Generative KI-Workload-Bewertung

AWS Präskriptive Leitlinien

AWS Präskriptive Leitlinien: Generative KI-Workload-Bewertung

Table of Contents

Generative KI-Workload-Bewertung

Tabby Ward und Deepak Dixit, Amazon Web Services ()AWS

November 2024 (Geschichte der Dokumente)

Die generative KI-Workload-Assessment ist eine strategische Methode zur Bewertung und Verbesserung der Bereitschaft eines Unternehmens, seine generativen KI-Workloads zu erstellen oder zu aktualisieren. Diese Bewertung ist wichtig, da die Integration generativer KI in den Geschäftsbetrieb die Funktionsweise der Dinge stark verändern und neue Effizienzen und Funktionen bieten kann. Um generative KI erfolgreich einzuführen, ist es jedoch unerlässlich, die aktuellen Systeme gründlich zu verstehen und einen klaren Plan für die future zu haben.

Generative KI-Workloads beziehen sich auf Rechenaufgaben, bei denen Modelle künstlicher Intelligenz verwendet werden, mit denen neue Inhalte wie Text, Bilder, Code oder andere Datentypen erstellt werden können. Diese Workloads erfordern in der Regel eine beträchtliche Rechenleistung, spezielle Hardware und große Datensätze für Training und Inferenz. GPUs Die Integration generativer KI-Workloads in den Betrieb bringt mehrere Herausforderungen mit sich:

- Infrastrukturanforderungen: Bereitstellung der erheblichen Rechenressourcen und der speziellen Hardware, die für generative KI-Modelle erforderlich sind.
- Datenmanagement: Sicherstellung von Datenqualität, Datenschutz und Compliance bei der Verarbeitung großer Datenmengen.
- Qualifikationslücke: Mangelndes Fachwissen in KI-Technologien und Modellbereitstellung.
- Ethische Überlegungen: Bekämpfung von Vorurteilen, Fairness und Transparenz bei KI-generierten Inhalten.
- Komplexität der Integration: Nahtlose Integration generativer KI in bestehende Workflows und Altsysteme.
- Kostenmanagement: Abwägen der potenziellen Vorteile mit den hohen Implementierungs- und Betriebskosten.

Die Bewältigung dieser Herausforderungen erfordert sorgfältige Planung, Investitionen in Infrastruktur und Talente sowie einen strategischen Umsetzungsansatz.

Zweck dieses Leitfadens

Generative KI wird in vielen Branchen schnell zu einer wichtigen Komponente. Sie bietet transformative Möglichkeiten, bringt aber auch Herausforderungen in Bezug auf Integration, Compliance und Skalierbarkeit mit sich. Viele Unternehmen haben aufgrund schwacher technologischer Grundlagen, Widerstand gegen Veränderungen und Datenqualitätsprobleme Schwierigkeiten, KI voll auszuschöpfen. Die generative KI-Workload-Assessment adressiert diese Herausforderungen, indem sie die Anforderungen für eine Modernisierung identifiziert, den Umfang der Implementierung definiert und bestehende Systeme und Denkweisen in Frage stellt. Es hilft Ihnen auch bei der Bestimmung von Produkten, die am wenigsten praktikabel sind (MVPs) und hilft Ihnen bei der Entwicklung einer Ziellösungsarchitektur, wodurch ein strukturierter und strategischer Ansatz für die Einführung von KI gewährleistet wird.

Dieser Leitfaden dient als strukturierter Ansatz, der Unternehmen dabei unterstützt, die Komplexität der Einführung generativer KI-Technologien zu bewältigen. Anstatt die Anforderungen von Anfang an klar zu definieren, hilft der Leitfaden bei:

- Identifizierung potenzieller Anwendungsfälle für generative KI in Ihrem Unternehmen.
- Bewertung der Bereitschaft Ihres Unternehmens für die Einführung generativer KI.
- Definition und Verfeinerung von Zielen für Anwendungsfälle und ehrgeizige Ziele.
- Festlegung des Umfangs und der Anforderungen für die Implementierung generativer KI.
- Entwicklung einer Architektur für eine Ziellösung.

Zielgruppe und Vorteile

Diese Bewertung richtet sich speziell an Lösungsarchitekten, Unternehmensarchitekten und Anwendungsarchitekten, die die technischen Aspekte der generativen KI-Workload-Modernisierung bewerten möchten. Es ist auch nützlich für Programm- und Personalmanager, die die allgemeine Bereitschaft, die Ressourcenzuweisung und die Anforderungen an die Befähigung ihres Teams einschätzen möchten. Die bewährten Verfahren der Branche unterstreichen, wie wichtig eine umfassende Bewertung ist, um sicherzustellen, dass die Voraussetzungen für die Einführung von KI erfüllt sind. Dazu gehören die Bewertung von Architektur, Speicher, Compliance, Integration, Tests, Bereitstellung und Automatisierung.

Scope

Die folgenden Themen sind Gegenstand der generativen KI-Methode zur Bewertung der Arbeitslast:

- Aktuelle generative KI-Technologien und -Modelle (z. B. große Sprachmodelle, Modelle zur Bilderzeugung)
- Enge KI-Anwendungen, die generative Techniken verwenden
- Integration generativer KI in bestehende Systeme und Workflows
- Datenstrategien für das Training und die Feinabstimmung generativer KI-Modelle
- Ethische Überlegungen und verantwortungsvolle KI-Praktiken für aktuelle generative KI-Anwendungen
- Test- und Einsatzstrategien für generative KI in Produktionsumgebungen
- Überlegungen zur Sicherheit und zum Datenschutz bei generativen KI-Implementierungen
- Leistungsoptimierung und Skalierbarkeit generativer KI-Workloads
- Anwendungsfälle und Anwendungen generativer KI in verschiedenen Branchen
- Bewertung generativer KI-Ergebnisse und Qualitätssicherungsprozesse

Die folgenden Themen fallen nicht in den Geltungsbereich:

- Szenarien mit künstlicher allgemeiner Intelligenz (AGI) und künstlicher Superintelligenz (ASI)
- Spekulative future Fortschritte in der KI, die über aktuelle generative Modelle hinausgehen
- Quantencomputer-Anwendungen in der KI
- Neuromorphes Computing und Gehirn-Computer-Schnittstellen
- Bewusstsein und Selbstbewusstsein in KI-Systemen
- Langfristige gesellschaftliche Auswirkungen fortschrittlicher KI, die über aktuelle generative KI-Anwendungen hinausgehen
- Regulatorische Rahmenbedingungen für hypothetische future KI-Technologien
- Philosophische Debatten über die Natur von Intelligenz und Bewusstsein in Maschinen
- Extreme Randfälle oder hochspekulative Anwendungsfälle von KI
- Detaillierte technische Spezifikationen proprietärer KI-Modelle oder -Architekturen

Gezielte Geschäftsergebnisse

Die generative KI-Workload-Assessment zielt darauf ab, mehrere gezielte Ergebnisse zu erzielen, die für die erfolgreiche Modernisierung generativer KI-Workloads entscheidend sind. Diese Ergebnisse stellen sicher, dass Unternehmen gut darauf vorbereitet sind, KI-Technologien effektiv und effizient zu integrieren.

Für jedes angestrebte Ergebnis konzentriert sich die generative KI-Workload-Bewertung auf:

- Interdependenzen: Identifizieren und klären Sie etwaige Interdependenzen zwischen dem Ergebnis und anderen Aspekten des Modernisierungsprozesses. Dazu gehört auch, zu verstehen, wie ein Ergebnis andere beeinflussen oder von ihnen beeinflusst werden könnte, um einen ganzheitlichen Modernisierungsansatz zu gewährleisten.
- Abstimmung der Interessengruppen: Skizzieren Sie Strategien, um verschiedene Interessengruppen mit den einzelnen Ergebnissen in Einklang zu bringen. Dies beinhaltet die Vermittlung des Werts und der Auswirkungen der einzelnen Ergebnisse an verschiedene Organisationsebenen und Abteilungen, um die Akzeptanz und Unterstützung zu fördern.
- Priorisierung: In Fällen, in denen mehrere Anwendungsfälle oder Ergebnisse identifiziert werden, sollten Sie einen Rahmen für deren Priorisierung auf der Grundlage von Faktoren wie Geschäftsauswirkungen, Ressourcenanforderungen und strategischer Ausrichtung bereitstellen.
- Kontinuierliche Verbesserung: Richten Sie für jedes Ergebnis Mechanismen zur kontinuierlichen Bewertung und Verbesserung ein. Dadurch wird sichergestellt, dass die Modernisierungsbemühungen anpassungsfähig bleiben und auf sich ändernde Technologielandschaften und Geschäftsanforderungen reagieren.

Im Folgenden finden Sie eine ausführliche Erläuterung der einzelnen angestrebten Ergebnisse:

Zielarchitektur

- Definition: Die Bewertung hilft bei der Definition einer klaren und skalierbaren Zielarchitektur für generative KI-Workloads.
- Komponenten: Dazu gehören die Auswahl geeigneter Cloud-Dienste, die Gestaltung von Datenpipelines und die Sicherstellung der Systeminteroperabilität.
- Vorteile: Eine klar definierte Architektur unterstützt Skalierbarkeit, Zuverlässigkeit und Leistungsoptimierung und bietet eine solide Grundlage für die Modernisierung.

Bereitschaft der Kunden

- Bewertung: Beurteilen Sie den aktuellen Stand der Infrastruktur, der Prozesse und der Unternehmenskultur, um festzustellen, ob Sie für die Einführung generativer KI-Modernisierung bereit sind.
- Kriterien: Dazu gehören die Bewertung der technischen Fähigkeiten, der Datenqualität und der Bereitschaft der Organisation, Veränderungen anzunehmen.
- Ergebnis: Durch die Identifizierung von Lücken und Verbesserungsmöglichkeiten wird sichergestellt, dass das Unternehmen auf einen reibungslosen Übergang zu modernen Lösungen und Technologien vorbereitet ist.

Ziele für Anwendungsfälle und weitreichende Ziele

- Mit den Zielen von Anwendungsfällen werden klare Ziele für die Implementierung der Ziellösung festgelegt, wobei der Schwerpunkt auf bestimmten Geschäftsproblemen oder Geschäftschancen liegt.

  Ein Anwendungsfallziel im Kontext der generativen KI-Modernisierung bezieht sich auf ein bestimmtes, messbares Ziel, das eine Organisation durch die Implementierung generativer KI-Lösungen erreichen möchte. Diese Ziele sind in der Regel auf umfassendere Geschäftsziele ausgerichtet und konzentrieren sich auf die Bewältigung bestimmter Herausforderungen oder Chancen innerhalb des Unternehmens. Zu den Zielen von Anwendungsfällen könnten beispielsweise gehören:

  - Reduzierung der Reaktionszeit des Kundendienstes um 50 Prozent durch den Einsatz generativer KI-gestützter Chatbots.
  - Verbesserung der Effizienz der Codeüberprüfung um 30 Prozent durch generative KI-gestützte Codeanalyse.
  - Verbesserung der Genauigkeit der Betrugserkennung um 25 Prozent durch den Einsatz generativer KI-Mustererkennung.
- Langfristige Ziele definieren ehrgeizige Ziele, die die Grenzen dessen, was die generative KI-Modernisierung innerhalb des Unternehmens erreichen kann, erweitern.
- Wirkung: Die Festlegung sowohl erreichbarer als auch ehrgeiziger Ziele trägt dazu bei, Initiativen zur generativen KI-Modernisierung mit strategischen Geschäftszielen in Einklang zu bringen und Innovationen zu fördern.

Schätzung des Aufwands

- Zweck: Eine genaue Aufwandsschätzung hilft bei der Ressourcenplanung und stellt sicher, dass Projekte pünktlich und innerhalb des Budgets abgeschlossen werden.
- Umfang: Schätzen Sie die Ressourcen, die Zeit und das Budget ab, die für die Umsetzung des generativen KI-Modernisierungsplans erforderlich sind.
- Faktoren: Berücksichtigen Sie die technische Komplexität, Integrationsherausforderungen und potenzielle Risiken.

Unterstützungsanforderungen

- Schulung und Entwicklung: Identifizieren Sie die Fähigkeiten und Kenntnisse, die für eine erfolgreiche Einführung der generativen KI-Modernisierung erforderlich sind.
- Ressourcen: Ermitteln Sie den Bedarf an Schulungsprogrammen, Workshops und anderen unterstützenden Aktivitäten.
- Ergebnis: Die Sicherstellung, dass die Mitarbeiter mit den erforderlichen Fähigkeiten ausgestattet sind, erhöht die Effektivität generativer KI-Modernisierungsinitiativen und unterstützt den langfristigen Erfolg.

Umsetzungsplan

- Roadmap: Entwickeln Sie einen detaillierten Plan, der die Schritte beschreibt, die zur generativen KI-Modernisierung erforderlich sind.
- Meilensteine: Definieren Sie wichtige Meilensteine und Ergebnisse, um den Fortschritt zu verfolgen.
- Vorteile: Ein klarer Implementierungsplan gibt Orientierung und Rechenschaftspflicht vor und ermöglicht einen strukturierten Ansatz für die generative KI-Modernisierung.

Überlegungen und Voraussetzungen für die Bewertung

Beginnen Sie mit klaren Anwendungsfällen

Identifizieren Sie spezifische Geschäftsprobleme oder Chancen, die mit generativer KI angegangen werden können. Konzentrieren Sie sich auf Anwendungsfälle, die auf strategische Geschäftsziele abgestimmt sind und messbare Vorteile bieten. Priorisieren Sie Anwendungsfälle, die auf häufig auftretende Herausforderungen innerhalb des Unternehmens abzielen, um sicherzustellen, dass die Lösungsarchitektur als Muster für mehrere Szenarien dienen kann.

Die Einleitung des Bewertungsprozesses mit einem allgemeinen Verständnis potenzieller generativer KI-Anwendungen ist von Vorteil, aber nicht zwingend erforderlich. Der Fragebogen, der diesem Leitfaden beiliegt, berücksichtigt verschiedene Vorbereitungsstufen, von Unternehmen mit klar definierten Anwendungsfällen bis hin zu Unternehmen, die nur allgemeine Ideen haben. Das Bewertungsverfahren dient folgenden Zwecken:

- Verfeinern und verdeutlichen Sie diese ersten Ideen für Anwendungsfälle.
- Identifizieren Sie neue potenzielle Anwendungsfälle.
- Entwickeln Sie spezifische, messbare Ziele für jeden Anwendungsfall.
- Beurteilen Sie die Machbarkeit und die potenziellen Auswirkungen jedes Anwendungsfalls.

Betrachten wir ein hypothetisches Beispiel: Ein Finanzdienstleistungsunternehmen beschließt, die generative KI-Modernisierung in Betracht zu ziehen. Sie beginnen mit einer umfassenden Idee zur Verbesserung ihres Kundendienstes und ihrer Prozesse zur Betrugserkennung.

- Erste Bewertung: Der Fragebogen hilft ihnen dabei, ihre aktuellen Systeme, die Datenqualität und die organisatorische Eignung für die Einführung generativer KI zu bewerten.
- Verfeinerung der Anwendungsfälle: Im Rahmen des Bewertungsprozesses verfeinern sie ihre ursprünglichen Ideen in zwei spezifische Anwendungsfälle:
  - Implementierung eines generativen KI-gestützten Chatbots für Kundenanfragen
  - Einsatz generativer KI zur Erkennung von Transaktionsbetrug in Echtzeit
- Zielsetzung: Für jeden Anwendungsfall definieren sie spezifische Ziele:
  - Reduzieren Sie die Reaktionszeit des Kundendienstes innerhalb von 6 Monaten um 40 Prozent
  - Verbessern Sie die Genauigkeit der Betrugserkennung um 20 Prozent und reduzieren Sie Fehlalarme um 15 Prozent

- Langfristige Ziele: Sie haben sich auch diese ehrgeizigen Ziele gesetzt:
  - Erreichen Sie mit KI-gestützten Antworten eine Kundenzufriedenheit von 80 Prozent
  - Entwickeln Sie ein Modell zur prädiktiven Betrugserkennung, das neue Betrugsmuster identifiziert
- MVP-Definition: Der Fragebogen hilft ihnen dabei, für jeden Anwendungsfall ein MVP zu ermitteln, wobei der Schwerpunkt auf wesentlichen Funktionen liegt, die einen unmittelbaren Nutzen bieten.
- Zielarchitektur: Schließlich entwickeln sie eine Zielarchitektur, die einen oder beide Anwendungsfälle unterstützt und Skalierbarkeit und Integration in bestehende Systeme gewährleistet.

Stellen Sie die Geschäftsausrichtung sicher

Stimmen Sie generative KI-Initiativen auf die allgemeine Geschäftsstrategie und die allgemeinen Unternehmensziele ab. Entwickeln Sie für jeden Anwendungsfall ein klares Wertversprechen, das zeigt, wie generative KI zu Unternehmenswachstum, Effizienz oder Innovation beiträgt. Legen Sie Kennzahlen fest, um die Auswirkungen generativer KI-Implementierungen auf wichtige Leistungsindikatoren zu messen ()KPIs.

Implementieren Sie Steuerung und Aufsicht

Richten Sie einen funktionsübergreifenden Lenkungsausschuss ein, der generative KI-Initiativen überwacht. Entwickeln Sie Richtlinien und Richtlinien für einen verantwortungsvollen Einsatz von KI und berücksichtigen Sie dabei ethische Überlegungen und potenzielle Vorurteile. Richten Sie einen Überprüfungsprozess für generative KI-Projekte ein, um die Einhaltung organisatorischer Standards und regulatorischer Anforderungen sicherzustellen.

Adressdaten und technische Voraussetzungen

Beurteilen und verbessern Sie die Datenqualität und implementieren Sie Datenverwaltungspraktiken, um zuverlässige Inputs für generative KI-Modelle zu gewährleisten. Entwickeln Sie eine Datenstrategie, die sich mit der Erfassung, Speicherung und Verwaltung von Daten befasst, die speziell auf die Bedürfnisse generativer KI zugeschnitten sind. Evaluieren und verbessern Sie die Dateninfrastruktur, um das Volumen und die Geschwindigkeit der Daten zu unterstützen, die für generative KI-Workloads erforderlich sind.

Berücksichtigen Sie die Anforderungen an die Rechenressourcen

Beurteilen Sie die aktuelle IT-Infrastruktur und identifizieren Sie Lücken in der Rechenkapazität für generative KI-Workloads. Planen Sie skalierbare Rechenressourcen ein und ziehen Sie Optionen wie Cloud-Dienste oder lokale Hochleistungs-Computing-Cluster in Betracht. Optimieren Sie die Ressourcenzuweisung, um ein ausgewogenes Verhältnis zwischen Leistung und Kosteneffektivität sowohl für Schulungs- als auch für Inferenz-Workloads zu erreichen.

Gehen Sie auf die Auswirkungen auf Datenschutz und Sicherheit ein

Implementieren Sie robuste Sicherheitsmaßnahmen zum Schutz sensibler Daten, die in generativen KI-Trainings und -Operationen verwendet werden. Achten Sie beim Umgang mit personenbezogenen Daten auf die Einhaltung von Datenschutzbestimmungen wie der Allgemeinen Datenschutzverordnung (GDPR) oder dem California Consumer Privacy Act (CCPA). Entwickeln Sie Protokolle für die sichere Implementierung und Überwachung von Modellen, um unbefugten Zugriff oder Missbrauch generativer KI-Funktionen zu verhindern.

Binden Sie Interessengruppen frühzeitig ein

Binden Sie wichtige Interessengruppen von Anfang an ein, um die Zustimmung und Unterstützung der Führung zu gewinnen. Kommunizieren Sie klar und deutlich die Vorteile und potenziellen Auswirkungen von Modernisierungsinitiativen, insbesondere für generative KI-Workloads. Bieten Sie Schulungen und Ressourcen an, um Interessengruppen dabei zu helfen, generative KI-Technologien und ihre Auswirkungen zu verstehen.

Iteriere und lerne

Verfolgen Sie einen schrittweisen Ansatz, mit dem Sie Ihre Ziellösungen verfeinern können. Verwenden Sie Feedback-Schleifen, um die Workload-Architektur und die Prozesse kontinuierlich zu verbessern. Beurteilen Sie regelmäßig die Leistung und die Auswirkungen generativer KI-Implementierungen und passen Sie die Strategien nach Bedarf an, basierend auf realen Ergebnissen und sich ändernden Geschäftsanforderungen.

Fragebogen zur generativen KI-Workload-Bewertung

In den folgenden Abschnitten finden Sie Fragen, anhand derer Sie verschiedene Aspekte der generativen KI-Workload-Modernisierung für Ihr Unternehmen bewerten können. In diesem umfassenden Fragebogen wird die Bereitschaft Ihres Unternehmens zur Einführung und Implementierung generativer KI-Workloads anhand von Fragen zu Schlüsselbereichen bewertet, darunter Anwendungsfälle, Architektur, Speicher, Compliance, Integration, Tests, Bereitstellung und Datenstrategie. Dieser Fragebogen befasst sich mit wichtigen Aspekten der generativen KI-Implementierung, von der technischen Infrastruktur bis hin zu regulatorischen Überlegungen, und hilft Ihnen dabei, Stärken, Lücken und Chancen auf Ihrem Weg zur KI-Modernisierung zu identifizieren.

Abschnitte:

Sie können den Fragebogen auch im Microsoft Excel-Format herunterladen und zur Erfassung Ihrer Informationen verwenden.

Frageb herunterladen

Bereitschaft

Frage	Beispielantwort
Haben Sie AWS Konten, die für diese Workloads genutzt werden können?	Ja oder nein.

Frage	Beispielantwort
Haben Sie eine bestehende Unternehm ensvereinbarung mit AWS?	Ja oder nein.
Wie skalierbar ist Ihre aktuelle Cloud-Inf rastruktur, um generative KI-Workloads zu bewältigen?	Unsere Cloud-Infrastruktur ist hochgradig skalierbar und bietet automatische Skalierun gsfunktionen für Rechenressourcen und verteilte Speichersysteme, die darauf ausgelegt sind, umfangreiche generative KI-Workloads effizient zu bewältigen.
Verfügen Sie über Daten-Pipeline-Funktionen für Vorverarbeitung und Feature-Engineering in großem Maßstab?	Unsere Daten-Pipelines verwenden verteilte Verarbeitungs-Frameworks wie Apache Spark für umfangreiche Datenvorverarbeitung und Feature-Engineering, wobei sowohl die Batch- als auch die Streaming-Datenverarbeitung unterstützt wird.
Verfügen Sie über Funktionen zur Kontobere itstellung und -verwaltung?	Ja oder nein.
Wie würden Sie die KI-Kompetenz und die Bereitschaft Ihres Unternehmens zur Einführun g generativer KI-Technologien beschreiben?	Unsere Organisation hat stark in KI-Bildun gsprogramme investiert, und die meisten technischen Mitarbeiter haben eine KI/ML-Gru ndausbildung abgeschlossen. Die Organisation hat eine Innovationskultur, die neue Technolog ien, einschließlich generativer KI, umfasst.
Welches KI/ML-Fachwissen gibt es in Ihrem Unternehmen und wie wird es verteilt?	Wir haben ein eigenes KI-Exzellenzzentrum mit erfahrenen Datenwissenschaftlern und ML-Ingenieuren. Wir bilden Fachexperten in verschiedenen Geschäftsbereichen weiter, damit sie sich mit KI auskennen und generative KI-Anwendungsfälle identifizieren können.

Frage	Beispielantwort
Haben Sie ein übergeordnetes Geschäfts szenario, in dem die Ziele, Vorteile und Kosten des Cloud-Programms dargelegt werden?	Ja oder nein.
Was ist Ihr Zeitplan, um die Lösung zur Produktion zu bringen?	Wochen, Monate und so weiter.
Wurden von Ihren wichtigsten Stakeholdern (z. B. CFO, CIT/CTO, COO) finanzielle Mittel zugesagt?	Ja oder nein.
Wie stellen Sie die Einhaltung der Datenschu tzbestimmungen in Ihren generativen KI-Initia tiven sicher?	Wir haben ein engagiertes Compliance-Team, das eng mit unseren KI-Teams zusammena rbeitet. Wir führen regelmäßig Folgenabs chätzungen zum Datenschutz durch, setzen die Grundsätze des Datenschutzes durch Technikgestaltung um und führen detaillierte Datenverarbeitungsaufzeichnungen für alle generativen KI-Projekte.
Wie ausgereift sind Ihre bestehenden Systeme, die sich in neue generative KI-Technologien integrieren lassen?	Unsere IT-Architektur basiert auf Microserv ices und ermöglicht APIs die flexible Integrati on neuer generativer KI-Technologien. Diese Systeme sind auf gängige Datenformate und Protokolle standardisiert, um die Interoper abilität zu gewährleisten.
Welche Erfahrung haben Sie mit der Operation alisierung von ML-Modellen und wie könnte dies auf generative KI-Systeme zutreffen?	Wir verfügen über etablierte MLOps Verfahren, darunter automatisierte Pipelines zur Modellber eitstellung, Überwachungssysteme und Frameworks für A/B-Tests. Diese Praktiken werden an die besonderen Anforderungen groß angelegter generativer KI-Modelle angepasst.

Anwendungsfälle

Frage	Beispielantwort
Was ist das primäre Ziel oder die Erfolgskriterien des Anwendungsfalls?	Um die Reaktionszeit des Kundensupports zu verbessern, die Konversionsraten zu steigern und die Produktempfehlungen zu verbessern. Außerdem: Um die Benutzerzufriedenheit, die Abschlussquote von Aufgaben, die Antwortqualität usw. zu verbessern.
Wie passt dieser Anwendungsfall zu den strategischen Zielen Ihres Unternehmens?	Dies steht im Einklang mit unserem strategischen Ziel, die Kundenzufriedenheit durch Verkürzung der Reaktionszeiten im Kundenservice zu erhöhen.
Wie hoch ist die erwartete Menge an Daten oder Anfragen für den Anwendungsfall?	500 Transaktionen pro Sekunde (TPS).
Welche Arten von Datenquellen sind erforderlich, um Ihre generativen KI-Workloads zu unterstützen?	Interne strukturierte Datenbanken (Kundendaten, Verkaufsdaten usw.); unstrukturierte Textdaten aus Dokumenten, E-Mails und sozialen Medien; Audio- und Videodateien für Sprach- und Bilderkennungsaufgaben; Echtzeit-Streaming-Daten von IoT-Geräten und -Sensoren; öffentliche Datensätze und APIs zur Anreicherung.
Wie oft müssen Sie Daten aus diesen Quellen aktualisieren oder aktualisieren?	Transaktionsdatenbanken: Updates nahezu in Echtzeit; Dokumentenspeicher: tägliche Batch-Updates; Social-Media-Feeds: stündliche Updates; IoT-Sensordaten: kontinuierliches Echtzeit-Streaming; öffentliche Datensätze: monatliche oder vierteljährliche Updates.
Welche Datenformate benötigen Ihre generativen KI-Modelle als Eingabe?	Strukturierte Daten: CSV-, JSON- und SQL-Datenbanktabellen; Textdaten: Klartext, PDF und HTML; Bilddaten: JPEG, PNG und TIFF;

Frage	Beispielantwort
	Audiodaten: WAV und MP3; Videodaten: MP4 und AVI.
Was sind Ihre wichtigsten Bedenken hinsichtlich der Datenqualität bei generativen KI-Workloads?	Vollständigkeit: Sicherstellung, dass keine kritischen Felder fehlen; Genauigkeit: Überprüfung der Datenrichtigkeit und Beseitigung von Fehlern; Konsistenz: Beibehaltung einheitlicher Formate und Werte in allen Quellen; Aktualität: Sicherstellung, dass die Daten aktuell sind, sodass Rückschlüsse in Echtzeit möglich sind; Relevanz: Bestätigung, dass die Daten mit der spezifischen generativen KI-Aufgabe übereinstimmen.
Was sind die wichtigsten Leistungsanforderungen (z. B. Reaktionszeit, Durchsatz, Genauigkeit)?	Genauigkeit von 95%; Reaktionszeit < 500 ms; Fähigkeit, 1000 Anfragen pro Sekunde zu verarbeiten. Hohe Genauigkeit (95% +), mäßige Genauigkeit (80-90%), beste Leistung usw.
Haben Sie weitere Möglichkeiten, den Erfolg dieses Anwendungsfalls KPIs zu messen?	KPIs Zu den wichtigsten gehören die Reduzierung der Fehlerquote, die Zeitersparnis pro Transaktion und die Kundenzufriedenheit.
Wie viel Modellgenauigkeit wird gewünscht, und wie steht sie im Einklang mit den Kosten?	Hohe Genauigkeit (> 90%) bei moderaten Kosten, mäßige Genauigkeit (70-80%) bei niedrigen Kosten usw.
Was sind die wichtigsten Anwendungsfälle oder Szenarien für die generative KI-Lösung?	Kundenservice-Chatbot, Inhaltsgenerierung, Produktempfehlung und so weiter.
Was sind die Zielbenutzer oder Personas für das generative KI-System?	Kundendienstmitarbeiter, Marketingteam, Mitarbeiter, Endbenutzer usw.
Wie hoch ist das erwartete Volumen an Anfragen oder Benutzern?	1.000 Anfragen pro Tag; 10.000 aktive Benutzer pro Monat.

Frage	Beispielantwort
Gibt es spezielle Einschränkungen oder Anforderungen für den Anwendungsfall?	Reaktion in Echtzeit, mehrsprachiger Support, Datenschutz usw.
Haben Sie ein zugewiesenes Budget für die Entwicklung und Wartung der generativen KI-Lösung?	Die anfänglichen Entwicklungskosten werden auf 200.000$ geschätzt, die jährlichen Wartungskosten belaufen sich auf 50.000$.
Wie hoch sind die voraussichtliche Investiti onsrendite (ROI) und die Amortisationszeit für diesen Anwendungsfall?	Erwarteter ROI von 150% über drei Jahre mit einer Amortisationszeit von 18 Monaten.
Gibt es versteckte Kosten oder potenziel le Einsparungen, die berücksichtigt werden sollten?	Zu den möglichen Einsparungen gehören geringere Kosten für Überstunden. Zu versteckt en Kosten könnten zusätzliche Schulungen für das Personal gehören.
Was sind die Skalierbarkeit und die future Erweiterungsmöglichkeiten dieser generativen KI-Lösung?	Die Lösung ist so konzipiert, dass sie mit unseren Abläufen skaliert und die Möglichke it bietet, sie in future auf andere Abteilungen auszudehnen.
Wie sorgen Sie für Fairness und reduzieren Verzerrungen in Ihren generativen KI-Modell en?	Wir planen, Verzerrungen durch vielfältige Datenerhebungen, regelmäßige Prüfungen und die Implementierung von Techniken zur Minderung von Verzerrungen zu verringern.
Welche Verfahren haben Sie eingeführt, um ethische Bedenken oder unbeabsichtigte Folgen auszuräumen?	Wir werden ethische Bedenken durch einen etablierten Plan zur Reaktion auf KI-Vorfäl le, regelmäßige ethische Risikobewertungen, ein anonymes Berichtssystem für Mitarbeiter, die Zusammenarbeit mit externen Ethikexpe rten und die kontinuierliche Überwachung und Anpassung der eingesetzten Modelle auf der Grundlage von Feedback ausräumen.

Frage	Beispielantwort
Wie gehen Sie an die Priorisierung und Sequenzierung generativer KI-Workload-Assessments für verschiedene Projekte und Abteilungen in Ihrem Unternehmen heran?	Indem wir eine hochrangige Umfrage in allen Abteilungen durchführen, um potenzielle Anwendungsfälle für generative KI zu identifiz ieren und diese anhand von drei Schlüssel kriterien zu bewerten: Auswirkungen auf das Geschäft, technische Machbarkeit und ethische Überlegungen. Projekten mit hohem Wirkungsp otenzial, geringeren technischen Barrieren und minimalen ethischen Bedenken wird Vorrang eingeräumt.

Architektur

Frage	Beispielantwort
Welche Art von generativem KI-Modell oder Architektur wird in Betracht gezogen?	Transformator, konvolutionelles neuronale s Netzwerk (CNN), rekurrentes neuronales Netzwerk (RNN), Entscheidungsbäume usw.
Was ist der erwartete Umfang oder das erwartete Volumen von Daten und Berechnun gen?	Millionen von Benutzern, Petabyte an Daten und so weiter.
Was sind die Hardwareanforderungen (zum Beispiel CPUs oder GPUs) für Training und Inferenz?	High-End GPUs, CPU-Cluster, Cloud-Instanzen usw.
Wie wird das generative KI-Modell im Laufe der Zeit aktualisiert oder neu trainiert?	Durch kontinuierliches Lernen, regelmäßige Umschulungen, manuelle Updates usw.
Was sind die Anforderungen an die Datenvorv erarbeitung und das Feature-Engineering?	Textreinigung, Bildvergrößerung, Funktions auswahl usw.

Frage	Beispielantwort
Wie wird das generative KI-System mit Grenzfällen, Ausreißern oder Eingaben mit geringer Zuverlässigkeit umgehen?	Durch Rückgriff auf menschliche Aufsicht, Anfragen zur Klärung usw.
Was sind die Latenzanforderungen für die generative KI-Anwendung?	Stapelverarbeitung in Echtzeit, nahezu in Echtzeit usw.

Speicher

Frage	Beispielantwort
Wo werden die Trainingsdaten gespeichert?	Im Cloud-Speicher (z. B. Amazon S3, Dateispeicher, Blockspeicher oder Objektspeicher), im lokalen Speicher usw.
Was sind die Speicheranforderungen für die Trainingsdaten und Modellartefakte (z. B. Kapazität, Haltbarkeit, Verfügbarkeit)?	Speicher im Petabyte-Bereich, hohe Haltbarkeit (99,999999999% Haltbarkeit), hohe Verfügbarkeit usw.
Was sind die Datenaufbewahrungs- und Backup-Anforderungen für die Trainingsdaten und Modellartefakte?	Datenspeicherung für x Jahre, tägliche Backups, externe Backups usw.
Welche Dateiformate werden hauptsächlich zum Speichern Ihrer KI-Trainingsdatensätze verwendet (z. B. CSV, JSON, HDF5 Parquet)?	Parquet-Dateien für strukturierte Daten und HDF5 für große multidimensionale Arrays und unstrukturierte Daten wie Bilder und Text. Wir verwenden spezielle Formate, um beispielsweise das Laden von Daten während des Trainings TFRecord zu optimieren.
Wie sind Ihre Trainingsdatensätze organisiert: als einzelne Dateien, in Datenbanken oder mithilfe spezieller KI-Datenformate?	Kleine bis mittlere Datensätze werden aus Gründen der Flexibilität als einzelne Parquet-Dateien im Objektspeicher gespeichert. Große Datensätze werden aus Skalierungsgründen

Frage	Beispielantwort
	in einer verteilten Datenbank (Cassandra) gespeichert.
Verwenden Sie Datenkomprimierungs- oder Kodierungstechniken speziell für generative KI-Trainingsdaten?	Für tabellarische Daten verwenden wir Wörterbuchkodierungs- und Bitpacking-Techniken, die in Parquet verfügbar sind. Für Bilder verwenden wir die verlustbehaftete JPEG-Komprimierung mit für unsere Modelle optimierten Qualitätseinstellungen.
Wie gehen Sie mit der Versionierung und Speicherung verschiedener Iterationen von Trainingsdatensätzen um? Welche Auswirkungen hat dies auf Ihren allgemeinen Speicherbedarf?	Wir verwenden ein Datenversionssystem (DVC), das in unsere ML-Plattform integriert ist.

Vorschriften und Einhaltung

Frage	Beispielantwort
Was sind die relevanten Vorschriften oder Compliance-Anforderungen für die generative KI-Lösung (z. B. GDPR, HIPAA, PCI-DSS)?	DSGVO für den Umgang mit personenbezogenen Daten, HIPAA für Gesundheitsdaten, PCI-DSS für Zahlungsdaten und so weiter.
Welche ethischen Leitlinien oder Frameworks für generative KI hat Ihre Organisation übernommen?	Wir haben unsere eigenen Richtlinien für verantwortungsvolle KI implementiert. Alle generativen KI-Projekte werden vor der Genehmigung und Einführung einer ethischen Überprüfung unterzogen.
Was sind die Sicherheitsanforderungen für das generative KI-System?	Datenverschlüsselung, sichere Netzwerkkommunikation, regelmäßige Sicherheitsaudits.
Was sind die Anforderungen an Datenschutz und Datenschutz?	Datenanonymisierung, Verschlüsselung, Zugriffskontrolle usw.

Frage	Beispielantwort
Was sind die Anforderungen an die Lösung für den Umgang mit sensiblen oder vertraulichen Daten?	Strenge Zugriffskontrollen, Datenmaskierung, Anforderungen an die Datenresidenz usw.
Wie werden Benutzerauthentifizierung und -autorisierung gehandhabt?	Mithilfe von API-Schlüsseln OAuth, Single Sign-On (SSO) und rollenbasierter Zugriffskontrolle (RBAC).
Wie wird die Lösung in der Produktion überwacht und verwaltet?	Durch die Verwendung von Überwachungstools wie Prometheus und Datadog, Protokoll ierungstools wie ELK Stack, Warnsystemen usw.

Integration

Frage	Beispielantwort
Was sind die Anforderungen für die Integrati on der generativen KI-Lösung in bestehende Systeme oder Datenquellen?	REST APIs, Nachrichtenwarteschlangen, Datenbankkonnektoren usw.
Wie werden Daten für die generative KI-Lösung aufgenommen und vorverarbeitet?	Mithilfe von Stapelverarbeitung, Streaming-Daten, Datentransformationen und Feature-E ngineering.
Wie wird der Output der generativen KI-Lösung genutzt oder in nachgelagerte Systeme integriert?	Über API-Endpunkte, Nachrichtenwartesc hlangen, Datenbankaktualisierungen usw.
Welche ereignisgesteuerten Integrationsmuster können für die generative KI-Lösung verwendet werden?	Nachrichtenwarteschlangen (wie Amazon SQS, Apache Kafka, RabbitMQ), Pub/Sub-Systeme, Webhooks, Event-Streaming-Plattformen.
Welche API-basierten Integrationsansätze können verwendet werden, um die generative	RESTful APIs, GraphQL APIs, SOAP APIs (für ältere Systeme).

Frage	Beispielantwort
KI-Lösung mit anderen Systemen zu verbinden ?	
Welche Komponenten der Microservices-Arch itektur können für die generative KI-Lösung sintegration verwendet werden?	Service Mesh für dienstübergreifende Kommunikation, API-Gateways, Container-Orchestrierung (zum Beispiel Kubernetes).
Wie kann die hybride Integration für die generative KI-Lösung implementiert werden?	Durch die Kombination ereignisgesteuerte r Muster für Aktualisierungen in Echtzeit, Stapelverarbeitung für historische Daten und APIs für die Integration externer Systeme.
Wie kann der Output der generativen KI-Lösung in nachgelagerte Systeme integriert werden?	Über API-Endpunkte, Nachrichtenwartesc hlangen, Datenbankaktualisierungen, Webhooks und Dateiexporte.
Welche Sicherheitsmaßnahmen sollten bei der Integration der generativen KI-Lösung in Betracht gezogen werden?	Authentifizierungsmechanismen (wie OAuth oder JWT), Verschlüsselung (bei der Übertragu ng und im Ruhezustand), API-Ratenbegrenzun g und Zugriffskontrolllisten (ACLs).
Wie planen Sie, Open-Source-Frameworks wie LlamaIndex oder LangChain in Ihre bestehende Datenpipeline und Ihren generativen KI-Workfl ow zu integrieren?	Wir planen, sie zur Entwicklung komplexer generativer KI-Anwendungen LangChain zu verwenden, insbesondere im Hinblick auf ihre Agenten- und Speicherverwaltungsfunktion en. Unser Ziel ist es, LangChain innerhalb der nächsten 6 Monate 60% unserer generativen KI-Projekte zu nutzen.

Frage	Beispielantwort
Wie stellen Sie die Kompatibilität zwischen den von Ihnen ausgewählten Open-Source-Frameworks und Ihrer bestehenden Dateninfrastruktur sicher?	Wir stellen ein spezielles Integrationsteam zusammen, um eine reibungslose Kompatibilität zu gewährleisten. Unser Ziel ist es, bis zum dritten Quartal über eine vollständig integrierte Pipeline zu verfügen, die eine effiziente Indexierung und den Abruf von Daten innerhalb unserer aktuellen Data-Lake-Struktur ermöglicht. LlamaIndex
Wie planen Sie, die modularen Komponenten von Frameworks zu nutzen, z. B. LangChain für Rapid Prototyping und Experimente?	Wir richten eine Sandbox-Umgebung ein, in der Entwickler mithilfe LangChain der Komponenten schnell Prototypen erstellen können.
Was ist Ihre Strategie, um mit Updates und neuen Funktionen in diesen sich schnell entwickelnden Open-Source-Frameworks Schritt zu halten?	Wir haben ein Team beauftragt, GitHub Repositorien und Community-Foren für LangChain und LlamaIndex zu überwachen. Wir planen, vierteljährlich wichtige Updates zu evaluieren und zu integrieren, wobei der Schwerpunkt auf Leistungsverbesserungen und neuen Funktionen liegt.

Testen

Frage	Beispielantwort
Was sind die Testanforderungen (z. B. Komponententests, Integrationstests, end-to-end Tests)?	Komponententests für einzelne Komponenten, end-to-end Integrationstests mit externen Systemen, Tests für kritische Szenarien usw.
Wie stellen Sie die Datenqualität und Konsistenz zwischen verschiedenen Quellen für generatives KI-Training sicher?	Wir gewährleisten die Datenqualität durch automatisierte Tools zur Datenprofilierung, regelmäßige Datenprüfungen und einen zentralisierten Datenkatalog. Wir haben Richtlinien zur Datenverwaltung eingeführt, um

Frage	Beispielantwort
	die Konsistenz zwischen den Quellen sicherzustellen und die Datenherkunft aufrechtzuerhalten.
Wie wird das generative KI-Modell evaluiert und validiert?	Mithilfe eines Holdout-Datensatzes, einer Bewertung durch Menschen, A/B-Tests usw.
Was sind die Kriterien für die Bewertung der Leistung und Genauigkeit des generativen KI-Modells?	Präzision, Erinnerungsvermögen, F1-Wert, Ratlosigkeit, menschliche Bewertung und so weiter.
Wie werden Randfälle und Eckfälle identifiziert und behandelt?	Durch den Einsatz einer umfassenden Testsuite, Evaluierung durch Menschen, kontradiktorische Tests usw.
Wie werden Sie das generative KI-Modell auf mögliche Verzerrungen testen?	Mithilfe demografischer Paritätsanalysen, Tests zur Chancengleichheit, Techniken zur Reduzierung von Vorurteilen, kontrafaktischen Tests usw.
Welche Kennzahlen werden verwendet, um die Fairness der Ergebnisse des Modells zu messen?	Uneinheitliches Wirkungsverhältnis, ausgeglichene Gewinnchancen, demografische Parität, individuelle Fairnesskennzahlen usw.
Wie stellen Sie sicher, dass Ihre Testdatensätze für die Erkennung von Verzerrungen eine vielfältige Repräsentation enthalten?	Durch die Verwendung stratifizierter Stichproben aus demografischen Gruppen, die Zusammenarbeit mit Experten für Diversität, die Verwendung synthetischer Daten zum Füllen von Lücken usw.
Welches Verfahren wird für die kontinuierliche Überwachung der Fairness der Modelle nach der Einführung eingeführt?	Regelmäßige Fairness-Audits, automatische Systeme zur Erkennung von Verzerrungen, Analyse von Benutzerfeedback, regelmäßige Fortbildung mit aktualisierten Datensätzen usw.

Frage	Beispielantwort
Wie werden Sie mit intersektionalen Vorurteilen im generativen KI-Modell umgehen?	Mithilfe intersektionaler Fairnessanalysen, Subgruppentests, Zusammenarbeit mit Fachexperten für Intersektionalität usw.
Wie werden Sie die Leistungsfähigkeit des Modells in verschiedenen Sprachen und kulturellen Kontexten testen?	Durch die Verwendung mehrsprachiger Testsets, die Zusammenarbeit mit Kulturexperten, lokalisierte Fairness-Metriken, interkulturelle Vergleichsstudien usw.

Einsatz und Automatisierung

Frage	Beispielantwort
Was sind die Anforderungen für Skalierung und Lastenausgleich?	Intelligentes Routing von Anfragen; automatisches Skalierungssystem; Optimierung für schnelle Kaltstarts durch den Einsatz von Techniken wie Modell-Caching, verzögertem Laden und verteilter Speichersysteme; Entwicklung des Systems zur Bewältigung von unvorhersehbaren Datenverkehrsmustern mit hohem Datenvolumen.
Was sind die Anforderungen für die Aktualisierung und Einführung neuer Versionen?	Blaue/grüne Bereitstellungen, Canary-Releases, fortlaufende Updates und so weiter.
Was sind die Anforderungen für Disaster Recovery und Geschäftskontinuität?	Backup- und Wiederherstellungsverfahren, Failover-Mechanismen, Hochverfügbarkeitskonfigurationen usw.
Was sind die Anforderungen für die Automatisierung der Schulung, Bereitstellung und Verwaltung des generativen KI-Modells?	Automatisierte Trainingspipeline, kontinuierliche Bereitstellung, automatische Skalierung usw.

Frage	Beispielantwort
Wie wird das generative KI-Modell aktualisiert und neu trainiert, sobald neue Daten verfügbar werden?	Durch regelmäßige Umschulungen, inkrement elles Lernen, Transferlernen usw.
Was sind die Anforderungen für die Automatis ierung von Überwachung und Verwaltung?	Automatisierte Benachrichtigungen, automatis che Skalierung, Selbstheilung usw.
Was ist Ihre bevorzugte Bereitstellungsumg ebung für generative KI-Workloads?	Ein hybrider Ansatz, der AWS für Modellsch ulungen und unsere lokale Infrastruktur für Inferenz verwendet, um die Anforderungen an die Datenresidenz zu erfüllen.
Gibt es spezielle Cloud-Plattformen, die Sie für generative KI-Bereitstellungen bevorzugen?	AWS-Services, insbesondere Amazon SageMaker AI für Modellentwicklung und -bereitstellung und Amazon Bedrock für Basismodelle.
Welche Containerisierungstechnologien ziehen Sie für generative KI-Workloads in Betracht?	Wir möchten Docker-Container standardi sieren, die mit Kubernetes orchestriert sind, um die Portabilität und Skalierbarkeit in unserer Hybridumgebung sicherzustellen.
Haben Sie bevorzugte Tools für CI/CD in Ihrer generativen KI-Pipeline?	GitLab für Versionskontrolle und CI/CD-Pip elines, integriert in Jenkins für automatisiertes Testen und Bereitstellen.
Welche Orchestrierungstools ziehen Sie für die Verwaltung generativer KI-Workflows in Betracht?	Apache Airflow für die Workflow-Orchestri erung, insbesondere für Datenvorverarbeitung und Modelltrainingspipelines.
Haben Sie spezielle Anforderungen an die lokale Infrastruktur zur Unterstützung generativ er KI-Workloads?	Wir investieren in GPU-beschleunigte Server und Hochgeschwindigkeitsnetzwerke, um Inferenz-Workloads vor Ort zu unterstützen.

Frage	Beispielantwort
Wie planen Sie, die Versionierung und Bereitstellung von Modellen in verschiedenen Umgebungen zu verwalten?	Wir planen, es MLflow für die Modellverfolgung und Versionierung zu verwenden und es in unsere Kubernetes-Infrastruktur zu integrier en, um eine nahtlose Bereitstellung in allen Umgebungen zu gewährleisten.
Welche Tools zur Überwachung und Beobachtb arkeit ziehen Sie für generative KI-Implem entierungen in Betracht?	Prometheus für die Erfassung von Metriken und Grafana für die Visualisierung mit zusätzlichen benutzerdefinierten Protokollierungslösungen für die modellspezifische Überwachung.
Wie gehen Sie mit Datenbewegung und Synchronisation in einem hybriden Bereitste llungsmodell um?	Wir werden AWS DataSync für eine effiziente Datenübertragung zwischen lokalen Speichern und automatische Synchronisierungsaufgaben verwenden AWS, die auf der Grundlage unserer Trainingszyklen geplant werden.
Welche Sicherheitsmaßnahmen implement ieren Sie für generative KI-Implementierungen in verschiedenen Umgebungen?	Wir werden IAM für Cloud-Ressourcen verwenden, das in unser lokales Active Directory integriert ist, um end-to-end Verschlüs selung und Netzwerksegmentierung zur Sicherung von Datenströmen zu implement ieren.

Datenstrategie

Frage	Beispielantwort
Welche spezifischen Datentypen sind für Ihre generativen KI-Workloads von entscheidender Bedeutung, und auf wie viel Prozent davon kann derzeit zugegriffen werden?	Kundenanrufprotokolle und Daten zu Produktre zensionen sind von entscheidender Bedeutung. Derzeit sind 85% dieser Datentypen für unsere generativen KI-Projekte zugänglich.
Wie stellen Sie die Qualität Ihrer Daten sicher und messen sie?	Wir haben Kennzahlen zur Datenqualität eingeführt, darunter Vollständigkeit, Genauigke

Frage	Beispielantwort
	it, Konsistenz und Aktualität. Wir verwenden automatisierte Tools, um diese Kennzahlen regelmäßig zu bewerten, und verfügen über ein engagiertes Team für die Datenbereinigung und -anreicherung.
Wie viel Prozent Ihrer Daten entsprechen Ihren Qualitätsstandards für den Einsatz generativer KI?	Derzeit entsprechen 78% unserer Daten unseren Qualitätsstandards. Wir streben durch verbesserte Datenbereinigungsprozesse einen Wert von 95% innerhalb der nächsten 12 Monate an.
Wie planen Sie, bei Ihren Stakeholdern Vertrauen in die Datennutzung im Rahmen generativer KI aufzubauen?	Wir führen ein KI-Ethikgremium ein, das KI-Entscheidungen klar erklärt und vierteljährliche KI-Audits durchführt, um Transparenz und Fairness zu gewährleisten.
Wie umfassend ist Ihre Dokumentation in Bezug auf Datenquellen und Herkunft?	Wir führen einen detaillierten Datenkatalog, der Metadaten für alle unsere Datenquellen enthält, einschließlich Herkunft, Aktualisierungshäu figkeit und Nutzung. Wir verwenden Data Lineage-Tools, um zu verfolgen, wie Daten in unseren Systemen fließen und sich transform ieren.
Wie stellen Sie die Vielfalt Ihrer Datensätze sicher, um Verzerrungen in KI-Modellen zu verhindern?	Wir beziehen aktiv Daten aus unterschiedlichen Bevölkerungsgruppen und überprüfen unsere Datensätze regelmäßig auf repräsentative Verzerrungen. Wir verwenden auch Techniken zur synthetischen Datengenerierung, um unterrepräsentierte Kategorien auszugleichen.

Frage	Beispielantwort
Wie hoch ist Ihre Datenaktualisierungsrate für kritische generative KI-Modelle, und wie bestimmen Sie diese Häufigkeit?	Kritische Modelle werden wöchentlich aktualisiert. Diese Häufigkeit wird anhand von Leistungskennzahlen für A/B-Tests bestimmt, und wir streben einen Rückgang von höchstens 2% zwischen den Aktualisierungen an.
Wie viele Versionen kritischer Datensätze verwalten Sie und für wie lange?	Wir verwalten die letzten fünf Versionen jedes kritischen Datensatzes mit einer Aufbewahrungsfrist von 18 Monaten für jede Version.
Wie viele funktionsübergreifende Teams sind an Ihren generativen KI-Initiativen beteiligt und haben Zugriff auf Ihre Daten?	Wir haben drei funktionsübergreifende Teams. Jedes Team besteht aus Datenwissenschaftlern, Fachexperten, Ethikern und Geschäftsanalysten.
Welche Richtlinien und Praktiken zur Datenverwaltung haben Sie eingeführt?	Wir haben einen funktionsübergreifenden Ausschuss für Datenverwaltung, der unsere Datenrichtlinien überwacht. Wir haben rollenbasierte Zugriffskontrollen, Datenklassifizierungssysteme und regelmäßige Audits eingeführt, um die Einhaltung unseres Governance-Frameworks sicherzustellen.
Welche Maßnahmen haben Sie getroffen, um den Datenschutz zu gewährleisten, die erforderlichen Einwilligungen einzuholen und die Vertraulichkeit zu wahren?	Wir haben einen umfassenden Datenschutzrahmen eingeführt, der auf die DSGVO und den CCPA abgestimmt ist. Dazu gehören die Einholung der ausdrücklichen Zustimmung zur Datennutzung, die Implementierung von Techniken zur Datenanonymisierung und regelmäßige Folgenabschätzungen für den Datenschutz.

Frage	Beispielantwort
Wie viel Prozent Ihrer KI-Schulungsdatensätze wurden im letzten Quartal auf Verzerrungen geprüft?	70% unserer KI-Trainingsdatensätze wurden im letzten Quartal auf Verzerrungen geprüft. Wir implementieren automatisierte Tools zur Erkennung von Verzerrungen, um vierteljä hrliche Audits zu 100% zu erreichen.
Wie hoch ist Ihre aktuelle Datenverarbeitungs kapazität und wie viel benötigen Sie voraussic htlich für future generative KI-Workloads?	Unsere aktuelle Kapazität liegt TB/day. We project needing 30 TB/day innerhalb eines Jahres bei 10 und wir skalieren unsere Infrastru ktur, um diesem Bedarf gerecht zu werden.
Was ist Ihre Strategie, um den Datenschutz mit den Datenanforderungen generativer KI-Modell e in Einklang zu bringen?	Wir implementieren fortschrittliche Anonymisi erungstechniken und die Generierung synthetis cher Daten. Unser Ziel ist es, unsere nutzbaren Daten für KI im nächsten Jahr um 40% zu erhöhen und gleichzeitig die Datenschutzrisiken um 60% zu reduzieren.
Wie viel Prozent Ihrer maschinellen Lerndaten sätze (ML) sind korrekt gekennzeichnet, und wie hoch ist Ihre Zielgenauigkeitsrate?	Derzeit sind 85% unserer ML-Datensätze korrekt gekennzeichnet. Wir streben innerhalb des nächsten Quartals eine Genauigkeitsrate von 95% an, indem wir sowohl menschliche als auch automatisierte Kennzeichnungstechniken einsetzen.

Umsetzung von Erkenntnissen aus der Bewertung in umsetzbare Ergebnisse

Dieser Abschnitt bietet einen Rahmen für die Analyse der Antworten auf den Fragebogen und die Nutzung dieser Erkenntnisse zur Gestaltung der Zielarchitektur und anderer wichtiger Ergebnisse der Initiative zur generativen KI-Modernisierung. Dieses Framework überbrückt die Lücke zwischen Datenerfassung und Implementierung und stellt sicher, dass die Bewertung direkt in Ihre Modernisierungsstrategie einfließen und diese vorantreibt.

Definition der Zielarchitektur:

- Verwenden Sie die Antworten auf den Fragebogen als Grundlage für die Auswahl von Cloud-Diensten und die Gestaltung von Daten-Pipelines.
- Stellen Sie sicher, dass das Architekturdesign Skalierbarkeit und Interoperabilität unterstützt, wie im Leitfaden beschrieben.

Bewertung der Kundenbereitschaft:

- Analysieren Sie die Antworten auf den Fragebogen im Zusammenhang mit der aktuellen Infrastruktur, den Prozessen und der Unternehmenskultur.
- Identifizieren Sie Lücken und erstellen Sie einen Plan, um diese zu beheben. Priorisieren Sie Lücken, die für den Erfolg von MVP entscheidend sind.

Fallstudie und ehrgeizige Ziele:

- Extrahieren Sie spezifische Geschäftsprobleme aus den Antworten auf den Fragebogen, um klare Ziele für den Anwendungsfall zu definieren.
- Setzen Sie sich ehrgeizige Ziele, die der langfristigen Vision Ihres Unternehmens für die generative KI-Modernisierung entsprechen.

Schätzung des Aufwands:

- Verwenden Sie die Fragebogendaten, um Ressourcen, Zeit und Budget sowohl für das MVP als auch für die vollständige Implementierung abzuschätzen.

- Erstellen Sie einen schrittweisen Ansatz, der mit dem MVP beginnt, und skizzieren Sie die nachfolgenden Phasen.

Für die Aktivierung ist Folgendes erforderlich:

- Identifizieren Sie anhand der Antworten auf den Fragebogen Qualifikationslücken und Schulungsbedarf.
- Entwickeln Sie einen Schulungsplan, der sowohl den unmittelbaren MVP-Bedarf als auch die langfristige Einführung generativer KI unterstützt.

Umsetzungsplan:

- Erstellen Sie eine umfassende Roadmap, die mit dem MVP beginnt und die Schritte zur vollständigen generativen KI-Modernisierung skizziert.
- Definieren Sie klare Meilensteine und Ergebnisse für jede Phase der Implementierung.

Praktische Schritte:

- Priorisierungsmatrix: Erstellen Sie eine Matrix, die die Antworten auf den Fragebogen den sechs Ergebnissen zuordnet, um die Priorisierung von Funktionen und Bemühungen zu erleichtern.
- Iterativer Ansatz: Entwerfen Sie das MVP so, dass es die erste Iteration in einer Reihe von geplanten Releases ist, wobei jede Version auf der vollständigen Zielarchitektur aufbaut.
- Abstimmung zwischen den Stakeholdern: Nutzen Sie die Ergebnisse des Fragebogens, um die Stakeholder über den Umfang des MVP und den schrittweisen Ansatz zur Erreichung aller Ergebnisse zu informieren.
- Kontinuierliche Feedback-Schleife: Implementieren Sie Mechanismen, um Feedback nach der MVP-Implementierung zu sammeln, und nutzen Sie die Erkenntnisse, um Pläne für nachfolgende Phasen zu verfeinern.
- Agile Implementierung: Verwenden Sie eine agile Methode, die Flexibilität bei der Bewältigung aller Ergebnisse im Laufe der Zeit ermöglicht, angefangen bei den kritischsten Ergebnissen im MVP.

Nächste Schritte

Gehen Sie nach Abschluss der generativen KI-Workload-Bewertung wie folgt vor:

1. Stellen Sie eine detaillierte Zielarchitektur bereit
   - Ziel: Der Lösungsarchitekt erstellt eine umfassende Zielarchitektur, die auf die Unternehmensziele und die Ergebnisse der Bewertung abgestimmt ist.
   - Komponenten: Diese Architektur umfasst das Design der Datenaufnahme, der Integrationspunkte und der Systeminteroperabilität, um Skalierbarkeit, Zuverlässigkeit und Leistungsoptimierung sicherzustellen.
2. Erläutern Sie, wie spezifisch der AWS-Services Anwendungsfall ist
   - Servicemapping: Identifizieren und ordnen Sie spezifische Dienste zu AWS-Services , die am besten zu den identifizierten Anwendungsfällen passen.
   - Vorteile: Erläutern Sie, wie diese Services spezifische Geschäftsanforderungen erfüllen, die Effizienz steigern und Skalierbarkeit bieten.
3. Bieten Sie optionale alternative Lösungen mit Vor- und Nachteilen
   - Alternativen: Präsentieren Sie alternative Lösungen, die auch die Anforderungen des Unternehmens erfüllen könnten.
   - Analyse: Bieten Sie eine detaillierte Analyse der Vor- und Nachteile der einzelnen Alternativen an, indem Sie Faktoren wie Kosten, Komplexität und Ausrichtung auf die Geschäftsziele berücksichtigen.
4. Geben Sie eine detaillierte Preisschätzung von AWS-Services
   - Kostenanalyse: Erstellen Sie eine detaillierte Kostenschätzung für das Angebot AWS-Services, einschließlich potenzieller Nutzungsszenarien und Preismodelle.
   - Budgetausrichtung: Stellen Sie sicher, dass die Kosten den Budgetbeschränkungen der Organisation entsprechen und dass Sie sich über die finanziellen Auswirkungen im Klaren sind.
5. Holen Sie sich Feedback zur vorgeschlagenen Architektur
   - Einbindung der Stakeholder: Nehmen Sie Kontakt zu den Stakeholdern auf, um die vorgeschlagene Architektur vorzustellen und Feedback einzuholen.
   - Iterative Verbesserung: Nutzen Sie das Feedback, um die Lösung zu verfeinern und zu verbessern, und stellen Sie sicher, dass sie den Bedürfnissen und Erwartungen aller Beteiligten entspricht.

Häufig gestellte Fragen

Was ist das Hauptziel der generativen KI-Workload-Assessment?

Das Hauptziel der Bewertung besteht darin, die Bereitschaft eines Unternehmens zur Modernisierung seiner generativen KI-Workloads zu bewerten, Anwendungsfälle zu identifizieren und eine Ziellösungsarchitektur zu entwickeln. Ziel ist es, Modernisierungsanforderungen zu definieren, den Implementierungsumfang festzulegen und sich auf eine erfolgreiche generative KI-Modernisierung vorzubereiten.

Wer sollte diese Bewertung verwenden?

Diese Bewertung richtet sich an Lösungsarchitekten, Unternehmensarchitekten und Anwendungsarchitekten, die die technischen Aspekte der generativen KI-Modernisierung bewerten möchten. Es ist auch für Programm- und Personalmanager nützlich, um die allgemeine Bereitschaft, die Ressourcenzuweisung und den Bedarf an Ressourcen einzuschätzen.

Was sind die wichtigsten Komponenten, die bei der Bewertung bewertet wurden?

Die Bewertung umfasst die allgemeine Eignung, den Anwendungsfall, die Architektur, den Speicher, die Vorschriften und die Einhaltung von Vorschriften, Integration, Tests, Bereitstellungsautomatisierung und Datenstrategie. Diese Komponenten sind entscheidend für die Beurteilung der technischen und organisatorischen Eignung für die Einführung generativer KI-Modernisierung.

Wie hilft die Bewertung bei der Definition der Zielarchitektur?

Die Bewertung bietet einen strukturierten Ansatz zur Bewertung der aktuellen Systeme und zur Identifizierung von Verbesserungen. Es hilft Ihnen bei der Auswahl geeigneter Technologien und beim Entwurf skalierbarer Architekturen, die auf die Geschäftsziele und die Anforderungen der Anwendungsfälle abgestimmt sind.

Was sind die Vorteile einer generativen KI-Workload-Assessment?

Zu den Vorteilen gehören eine höhere Effizienz, eine bessere Entscheidungsfindung, die Sicherstellung der Einhaltung von Vorschriften, die Förderung von Innovationen und die Vorbereitung auf Skalierbarkeit. Die Bewertung legt einen strategischen Ansatz für die generative KI-Modernisierung fest und maximiert den potenziellen Nutzen bei gleichzeitiger Minimierung der Risiken.

Wie können Unternehmen im Anschluss an die Bewertung eine erfolgreiche Implementierung sicherstellen?

Organizations sollten einen klaren Umsetzungsplan entwickeln, der definierte Meilensteine enthält, Interessengruppen frühzeitig einbeziehen und einen iterativen Ansatz verfolgen. Die Einrichtung eines Center of Excellence (CoE) und die Fokussierung auf die Talententwicklung sind ebenfalls empfohlene bewährte Verfahren.

Vor welchen Herausforderungen könnten Unternehmen bei der Bewertung stehen?

Zu den Herausforderungen könnten der Widerstand gegen Veränderungen, Probleme mit der Datenqualität und die Komplexität der Einhaltung von Vorschriften gehören. Die Bewältigung dieser Herausforderungen erfordert die Förderung einer Innovationskultur, die Sicherstellung der Datenverfügbarkeit und die Umsetzung robuster Sicherheitsmaßnahmen.

Wie berücksichtigt die Bewertung die regulatorischen und Compliance-Anforderungen?

Bei der Bewertung werden die aktuellen Compliance-Maßnahmen bewertet und Lücken aufgedeckt. Es stellt sicher, dass die Target-Lösungen den geltenden Vorschriften und Datenschutzgesetzen entsprechen und bewährte Sicherheitsverfahren zum Schutz vertraulicher Informationen beinhalten.

Welche Rolle spielt die Einbindung der Interessengruppen im Bewertungsprozess?

Die Einbindung der Interessengruppen ist entscheidend, um Zustimmung zu gewinnen, Modernisierungsinitiativen an Geschäftszielen auszurichten und eine erfolgreiche Umsetzung sicherzustellen. Eine frühzeitige Einbindung und klare Kommunikation der Vorteile sind entscheidend, um Widerstände zu überwinden und Unterstützung zu fördern.

Wie können Unternehmen den Erfolg ihrer Initiativen zur generativen KI-Modernisierung nach der Bewertung messen?

Der Erfolg kann anhand von Leistungskennzahlen (KPIs) gemessen werden, die auf die Geschäftsziele abgestimmt sind. Die regelmäßige Überwachung und Bewertung dieser Kennzahlen hilft bei der Entscheidungsfindung und zeigt den Stakeholdern den Wert der generativen KI-Modernisierung auf.

Wie unterscheidet sich der Bewertungsansatz für Unternehmen unterschiedlicher Größe (kleine, mittlere oder große Unternehmen) oder Branchen?

Kleine Organisationen:

- Möglicherweise stehen nur begrenzte Ressourcen und Fachkenntnisse für umfassende Bewertungen zur Verfügung
- Wird sich wahrscheinlich auf spezifische Anwendungsfälle mit großer Wirkung konzentrieren, anstatt sich auf eine unternehmensweite Einführung zu konzentrieren
- Könnte sich bei der Bewertung stärker auf Tools und Dienste von Drittanbietern verlassen
- Der Bewertungsprozess ist möglicherweise weniger formell und agiler

Mittelgroße Unternehmen:

- Sie verfügen häufig über eigene IT- oder Datenteams, verfügen jedoch möglicherweise nicht über spezielles KI-Fachwissen

- Möglicherweise wird ein schrittweiser Ansatz gewählt, der mit Pilotprojekten in wichtigen Abteilungen beginnt
- Innovation muss mit bestehenden Systemen und Prozessen in Einklang gebracht werden
- An der Bewertung sind wahrscheinlich funktionsübergreifende Teams beteiligt

Unternehmensorganisationen:

- Sie verfügen in der Regel über spezielle AI/ML Teams und mehr Ressourcen für eine umfassende Bewertung
- Müssen komplexe Integrationen mit bestehenden Unternehmenssystemen in Betracht ziehen
- Möglicherweise müssen branchenspezifische regulatorische Anforderungen berücksichtigt werden
- Die Bewertung umfasst häufig formelle Verwaltungsprozesse

Ressourcen

- Generative KI aktiviert AWS
- AWS bietet neue Leitfäden zu künstlicher Intelligenz, maschinellem Lernen und generativer KI zur Planung Ihrer KI-Strategie (AWS Blogbeitrag)
- Bewährte Methoden zur Erstellung generativer KI-Anwendungen AWS (AWS Blogbeitrag)
- Generativer KI-Anwendungsgenerator aktiviert AWS (AWS Lösungsbibliothek)
- Generative KI-Funktionen (AWS Sicherheitsreferenzarchitektur)
- AWS Framework für bewährte Verfahren im Bereich generativer KI (AWS Audit Manager Benutzerhandbuch)
- Auswahl eines generativen KI-Dienstes (AWS Entscheidungsleitfaden)
- Was ist Amazon Bedrock? (Amazon Bedrock-Benutzerhandbuch)
- Was ist Amazon SageMaker AI?(Amazon SageMaker AI-Entwicklerhandbuch)

Dokumentverlauf

In der folgenden Tabelle werden wichtige Änderungen in diesem Leitfaden beschrieben. Um Benachrichtigungen über zukünftige Aktualisierungen zu erhalten, können Sie einen RSS-Feed abonnieren.

Änderung	Beschreibung	Datum
Erste Veröffentlichung	—	6. 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

Siehe 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 verarbeitet Transaktionen von verbindenden Anwendungen, 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 Einführung der Cloud (AWS CAF)

Ein Framework mit Richtlinien und bewährten Verfahren, das Unternehmen bei der Entwicklung eines effizienten und effektiven Plans für die erfolgreiche Umstellung 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 normalerweise 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 AWS für die 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 wird allgemein als Pipeline beschrieben. CI/CD kann 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 Abweichung 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.

Einsatz

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, wie z. B. unbeabsichtigte Fehlkonfigurationen oder ein Malware-Angriff.

Notfallwiederherstellung (DR)

Die Strategie und der Prozess, mit denen Sie Ausfallzeiten und Datenverluste aufgrund einer Katastrophe 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 Sicherheitsepen 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 CSPM, 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.

Daten zurückhalten

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 Infrastruktur als Code.

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 des Modells 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

Siehe 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

Siehe Labelbasierte 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 Verfolgung, 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 Mitglied einer Organisation sein.

MES

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.

EICHE

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.

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. Diese Entität ist in der Regel ein 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 der Kontrolle entspricht, 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).

RAG

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 für die Erholungszeit.

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-Managementkonsole 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 Benutzeranmeldeinformationen, 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 Amazon EC2 EC2-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.

Eingabeaufforderung ohne Zwischenfälle

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.