



Generative KI-Workload-Bewertung

AWS Prescriptive Guidance



AWS Prescriptive Guidance: Generative KI-Workload-Bewertung

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

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

Table of Contents

Einführung	1
Zweck dieses Leitfadens	2
Zielgruppe und Vorteile	2
Scope	3
Gezielte Geschäftsergebnisse	4
Überlegungen und Voraussetzungen für die Bewertung	7
Beginnen Sie mit klaren Anwendungsfällen	7
Stellen Sie die Geschäftsausrichtung sicher	8
Implementieren Sie Steuerung und Aufsicht	8
Adressdaten und technische Voraussetzungen	8
Berücksichtigen Sie die Anforderungen an die Rechenressourcen	9
Gehen Sie auf die Auswirkungen auf Datenschutz und Sicherheit ein	9
Binden Sie Interessengruppen frühzeitig ein	9
Iteriere und lerne	9
Fragebogen zur generativen KI-Workload-Bewertung	10
Bereitschaft	10
Anwendungsfälle	13
Architektur	16
Speicher	17
Vorschriften und Einhaltung	18
Integration	19
Testen	21
Einsatz und Automatisierung	23
Datenstrategie	25
Umsetzung von Erkenntnissen aus der Bewertung in umsetzbare Ergebnisse	29
Nächste Schritte	31
Häufig gestellte Fragen	32
Was ist das Hauptziel?	32
Wer sollte diese Bewertung verwenden?	32
Was sind die wichtigsten Komponenten?	32
Wie hilft dies bei der Definition der Architektur?	32
Was sind die Vorteile?	33
Wie können wir dies erfolgreich umsetzen?	33
Was sind die Herausforderungen?	33

Was sind die regulatorischen und Compliance-Anforderungen?	33
Welche Rolle spielen die Interessengruppen?	34
Wie können wir den Erfolg messen?	34
Wie unterscheidet sich der Ansatz je nach Unternehmensgröße?	34
Ressourcen	36
Dokumentverlauf	37
.....	xxxviii

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:

- [Bereitschaft](#)
- [Anwendungsfälle](#)
- [Architektur](#)
- [Speicher](#)
- [Vorschriften und Einhaltung](#)
- [Integration](#)
- [Testen](#)
- [Einsatz und Automatisierung](#)
- [Datenstrategie](#)

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

 [herunterladen](#)

Frageb

Bereitschaft

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

Frage	Beispielantwort
Haben Sie eine bestehende Unternehmensvereinbarung mit AWS?	Ja oder nein.
Wie skalierbar ist Ihre aktuelle Cloud-Infrastruktur, um generative KI-Workloads zu bewältigen?	Unsere Cloud-Infrastruktur ist hochgradig skalierbar und bietet automatische Skalierungsfunktionen 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 Kontobereitstellung und -verwaltung?	Ja oder nein.
Wie würden Sie die KI-Kompetenz und die Bereitschaft Ihres Unternehmens zur Einführung generativer KI-Technologien beschreiben?	Unsere Organisation hat stark in KI-Bildungsprogramme investiert, und die meisten technischen Mitarbeiter haben eine KI/ML-Grundausbildung abgeschlossen. Die Organisation hat eine Innovationskultur, die neue Technologien, 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 Geschäftsszenario, 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 Datenschutzbestimmungen in Ihren generativen KI-Initiativen sicher?	Wir haben ein engagiertes Compliance-Team, das eng mit unseren KI-Teams zusammenarbeitet. Wir führen regelmäßig Folgenabschätzungen 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 Microservices und ermöglicht APIs die flexible Integration neuer generativer KI-Technologien. Diese Systeme sind auf gängige Datenformate und Protokolle standardisiert, um die Interoperabilität zu gewährleisten.
Welche Erfahrung haben Sie mit der Operationalisierung von ML-Modellen und wie könnte dies auf generative KI-Systeme zutreffen?	Wir verfügen über etablierte MLOps Verfahren, darunter automatisierte Pipelines zur Modellbereitstellung, Überwachungssysteme und Frameworks für A/B-Tests. Diese Praktiken werden an die besonderen Anforderungen großer 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 Investitionsrendite (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 potenzielle Einsparungen, die berücksichtigt werden sollten?	Zu den möglichen Einsparungen gehören geringere Kosten für Überstunden. Zu versteckten 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 Möglichkeit bietet, sie in future auf andere Abteilungen auszuweiten.
Wie sorgen Sie für Fairness und reduzieren Verzerrungen in Ihren generativen KI-Modellen?	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-Vorfälle, regelmäßige ethische Risikobewertungen, ein anonymes Berichtssystem für Mitarbeiter, die Zusammenarbeit mit externen Ethikexperten 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 identifizieren und diese anhand von drei Schlüsselkriterien zu bewerten: Auswirkungen auf das Geschäft, technische Machbarkeit und ethische Überlegungen. Projekten mit hohem Wirkungspotenzial, 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 neuronales Netzwerk (CNN), rekurrentes neuronales Netzwerk (RNN), Entscheidungsbäume usw.
Was ist der erwartete Umfang oder das erwartete Volumen von Daten und Berechnungen?	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 Datenvorverarbeitung und das Feature-Engineering?	Textreinigung, Bildvergrößerung, Funktionsauswahl 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 Datensicherheit?	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, Protokollierungstools wie ELK Stack, Warnsystemen usw.

Integration

Frage	Beispielantwort
Was sind die Anforderungen für die Integration 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-Engineering.
Wie wird der Output der generativen KI-Lösung genutzt oder in nachgelagerte Systeme integriert?	Über API-Endpunkte, Nachrichtenwarteschlangen, 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-Architektur können für die generative KI-Lösung integration 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, Nachrichtenwarteschlangen, Datenbankaktualisierungen, Webhooks und Dateieexporte.
Welche Sicherheitsmaßnahmen sollten bei der Integration der generativen KI-Lösung in Betracht gezogen werden?	Authentifizierungsmechanismen (wie OAuth oder JWT), Verschlüsselung (bei der Übertragung und im Ruhezustand), API-Ratenbegrenzung und Zugriffskontrolllisten (ACLs).
Wie planen Sie, Open-Source-Frameworks wie LlamaIndex oder LangChain in Ihre bestehende Datenpipeline und Ihren generativen KI-Workflow zu integrieren?	Wir planen, sie zur Entwicklung komplexer generativer KI-Anwendungen LangChain zu verwenden, insbesondere im Hinblick auf ihre Agenten- und Speicherverwaltungsfunktionen. 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, inkrementelles Lernen, Transferlernen usw.
Was sind die Anforderungen für die Automatisierung von Überwachung und Verwaltung?	Automatisierte Benachrichtigungen, automatische Skalierung, Selbstheilung usw.
Was ist Ihre bevorzugte Bereitstellungsumgebung für generative KI-Workloads?	Ein hybrider Ansatz, der AWS für Modellschulungen 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 standardisieren, 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-Pipelines, 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-Orchestrierung, insbesondere für Datenvorverarbeitung und Modelltrainingspipelines.
Haben Sie spezielle Anforderungen an die lokale Infrastruktur zur Unterstützung generativer 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 integrieren, um eine nahtlose Bereitstellung in allen Umgebungen zu gewährleisten.
Welche Tools zur Überwachung und Beobachtbarkeit ziehen Sie für generative KI-Implementierungen 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 Bereitstellungsmodell 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 implementieren 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 Verschlüsselung und Netzwerksegmentierung zur Sicherung von Datenströmen zu implementieren.

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 Produktrezensionen 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, Genauigkeit

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 Datenbereinigungsverfahren 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, Aktualisierungshäufigkeit und Nutzung. Wir verwenden Data Lineage-Tools, um zu verfolgen, wie Daten in unseren Systemen fließen und sich transformieren.
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 Datenschutzzahmen 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 vierteljährliche Audits zu 100% zu erreichen.
Wie hoch ist Ihre aktuelle Datenverarbeitungskapazität und wie viel benötigen Sie voraussichtlich 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 Infrastruktur, um diesem Bedarf gerecht zu werden.
Was ist Ihre Strategie, um den Datenschutz mit den Datenanforderungen generativer KI-Modelle in Einklang zu bringen?	Wir implementieren fortschrittliche Anonymisierungstechniken und die Generierung synthetischer 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 Lerndatensätze (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

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.