Abrufen von Optionen und Architekturen für Augmented Generation auf AWS

AWS Präskriptive Leitlinien

AWS Präskriptive Leitlinien: Abrufen von Optionen und Architekturen für Augmented Generation auf AWS

Table of Contents

Optionen und Architekturen für Augmented Generation abrufen auf AWS

Mithil Shah, Rajeev Muralidhar und Natacha Fort, Amazon Web Services

Oktober 2024 (Geschichte der Dokumente)

Generative KI bezieht sich auf eine Untergruppe von KI-Modellen, die mit einer einfachen Texteingabe neue Inhalte und Artefakte wie Bilder, Videos, Text und Audio erstellen können. Generative KI-Modelle werden anhand riesiger Datenmengen trainiert, die eine Vielzahl von Themen und Aufgaben umfassen. Dadurch können sie eine bemerkenswerte Vielseitigkeit bei der Ausführung verschiedener Aufgaben unter Beweis stellen, auch bei solchen, für die sie nicht explizit geschult wurden. Aufgrund der Fähigkeit eines einzelnen Modells, mehrere Aufgaben auszuführen, werden diese Modelle oft als Basismodelle (FMs) bezeichnet.

Eine der bemerkenswerten Anwendungen generativer KI-Modelle ist ihre Fähigkeit, Fragen zu beantworten. Es gibt jedoch spezifische Herausforderungen, die sich ergeben, wenn diese Modelle zur Beantwortung von Fragen auf der Grundlage benutzerdefinierter Dokumente verwendet werden. Benutzerdefinierte Dokumente können firmeneigene Informationen, interne Websites, interne Dokumentation, Confluence Seiten, SharePoint Seiten und andere enthalten. Eine Option ist die Verwendung von Retrieval Augmented Generation (RAG). Bei RAG verweist das Foundation-Modell vor der Generierung einer Antwort auf eine autoritative Datenquelle, die sich außerhalb der Trainingsdatenquellen befindet (z. B. Ihre benutzerdefinierten Dokumente).

In diesem Leitfaden werden die verschiedenen generativen KI-Optionen beschrieben, die für die Beantwortung von Fragen aus benutzerdefinierter Dokumentation zur Verfügung stehen, einschließlich Retrieval Augmented Generation (RAG) -Systemen. Es bietet auch einen Überblick über die Erstellung von RAG-Systemen auf Amazon Web Services (AWS). Wenn Sie sich die RAG-Optionen und -Architekturen ansehen, können Sie zwischen vollständig verwalteten Services auf AWS und benutzerdefinierten RAG-Architekturen wählen.

Zielgruppe

Die Zielgruppe dieses Leitfadens sind Architekten und Manager generativer KI, die eine RAG-Lösung entwickeln, die verfügbaren Architekturen überprüfen und die Vor- und Nachteile der einzelnen Optionen verstehen möchten.

Ziele

Dieser Leitfaden hilft Ihnen bei folgenden Aufgaben:

- Machen Sie sich mit den generativen KI-Optionen vertraut, die zur Beantwortung von Fragen aus benutzerdefinierten Dokumenten zur Verfügung stehen
- Sehen Sie sich die Architekturoptionen für RAG-Systeme an unter AWS
- Machen Sie sich mit den Vor- und Nachteilen der einzelnen RAG-Optionen vertraut
- Wählen Sie eine RAG-Architektur für Ihre AWS Umgebung

Generative KI-Optionen für die Abfrage benutzerdefinierter Dokumente

Organizations verfügen häufig über verschiedene Quellen für strukturierte und unstrukturierte Daten. Dieser Leitfaden konzentriert sich darauf, wie Sie generative KI verwenden können, um Fragen aus unstrukturierten Daten zu beantworten.

Unstrukturierte Daten in Ihrem Unternehmen können aus verschiedenen Quellen stammen. Dies können Textdateien PDFs, interne Wikis, technische Dokumente, öffentlich zugängliche Websites, Wissensdatenbanken oder andere sein. Wenn Sie ein Basismodell benötigen, das Fragen zu unstrukturierten Daten beantworten kann, sind die folgenden Optionen verfügbar:

- Trainieren Sie ein neues Basismodell, indem Sie Ihre benutzerdefinierten Dokumente und andere Trainingsdaten verwenden
- Optimieren Sie ein vorhandenes Basismodell, indem Sie Daten aus Ihren benutzerdefinierten Dokumenten verwenden
- Verwenden Sie kontextbezogenes Lernen, um ein Dokument an das Foundation-Modell weiterzugeben, wenn Sie eine Frage stellen
- Verwenden Sie einen RAG-Ansatz (Retrieval Augmented Generation)

Es ist ein ehrgeiziges Unterfangen, ein neues Basismodell von Grund auf neu zu entwickeln, das Ihre benutzerdefinierten Daten enthält. Einige Unternehmen haben dies erfolgreich getan, beispielsweise Bloomberg mit ihrem BloombergGPTModell. Ein anderes Beispiel ist das multimodale EXAONEModell vonLG AI Research, das anhand von 600 Milliarden Kunstwerken und 250 Millionen hochauflösenden Bildern mit Text trainiert wurde. Laut The Cost of AI: Should You Build or Buy Your Foundation Model (LinkedIn) Meta Llama 2 kostet die Schulung eines ähnlichen Modells rund 4,8 Millionen US-Dollar. Es gibt zwei Hauptvoraussetzungen, um ein Modell von Grund auf zu trainieren: Zugang zu Ressourcen (finanzielle, technische, zeitliche) und eine klare Investitionsrendite. Wenn dies nicht die richtige Lösung zu sein scheint, besteht die nächste Option darin, ein vorhandenes Basismodell zu verfeinern.

Bei der Feinabstimmung eines vorhandenen Modells wird ein Modell, z. B. ein Amazon Titan-, Mistral- oder Lama-Modell, verwendet und das Modell anschließend an Ihre benutzerdefinierten Daten angepasst. Es gibt verschiedene Techniken für die Feinabstimmung, von denen die meisten nur die Änderung einiger weniger Parameter beinhalten, anstatt alle Parameter im Modell zu ändern.

Dies wird als parametereffiziente Feinabstimmung bezeichnet. Es gibt zwei Hauptmethoden für die Feinabstimmung:

- Die überwachte Feinabstimmung verwendet beschriftete Daten und hilft Ihnen, das Modell für eine neue Art von Aufgabe zu trainieren. Wenn Sie beispielsweise einen Bericht auf der Grundlage eines PDF-Formulars erstellen möchten, müssen Sie dem Modell möglicherweise anhand von ausreichend Beispielen beibringen, wie das geht.
- Die unbeaufsichtigte Feinabstimmung ist aufgabenunabhängig und passt das Basismodell an Ihre eigenen Daten an. Es trainiert das Modell, den Kontext Ihrer Dokumente zu verstehen. Das fein abgestimmte Modell erstellt dann Inhalte, z. B. einen Bericht, und verwendet dabei einen Stil, der besser an Ihre Organisation angepasst ist.

Eine Feinabstimmung ist jedoch möglicherweise nicht ideal für Anwendungsfälle mit Fragen und Antworten. Weitere Informationen finden Sie unter Vergleich von RAG und Feinabstimmung in diesem Handbuch.

Wenn Sie eine Frage stellen, können Sie einem Dokument das Grundlagenmodell übergeben und das kontextbezogene Lernen des Modells nutzen, um Antworten aus dem Dokument zurückzugeben. Diese Option eignet sich für die Ad-hoc-Abfrage eines einzelnen Dokuments. Diese Lösung eignet sich jedoch nicht gut für die Abfrage mehrerer Dokumente oder für die Abfrage von Systemen und Anwendungen wie Microsoft SharePoint oder Atlassian Confluence.

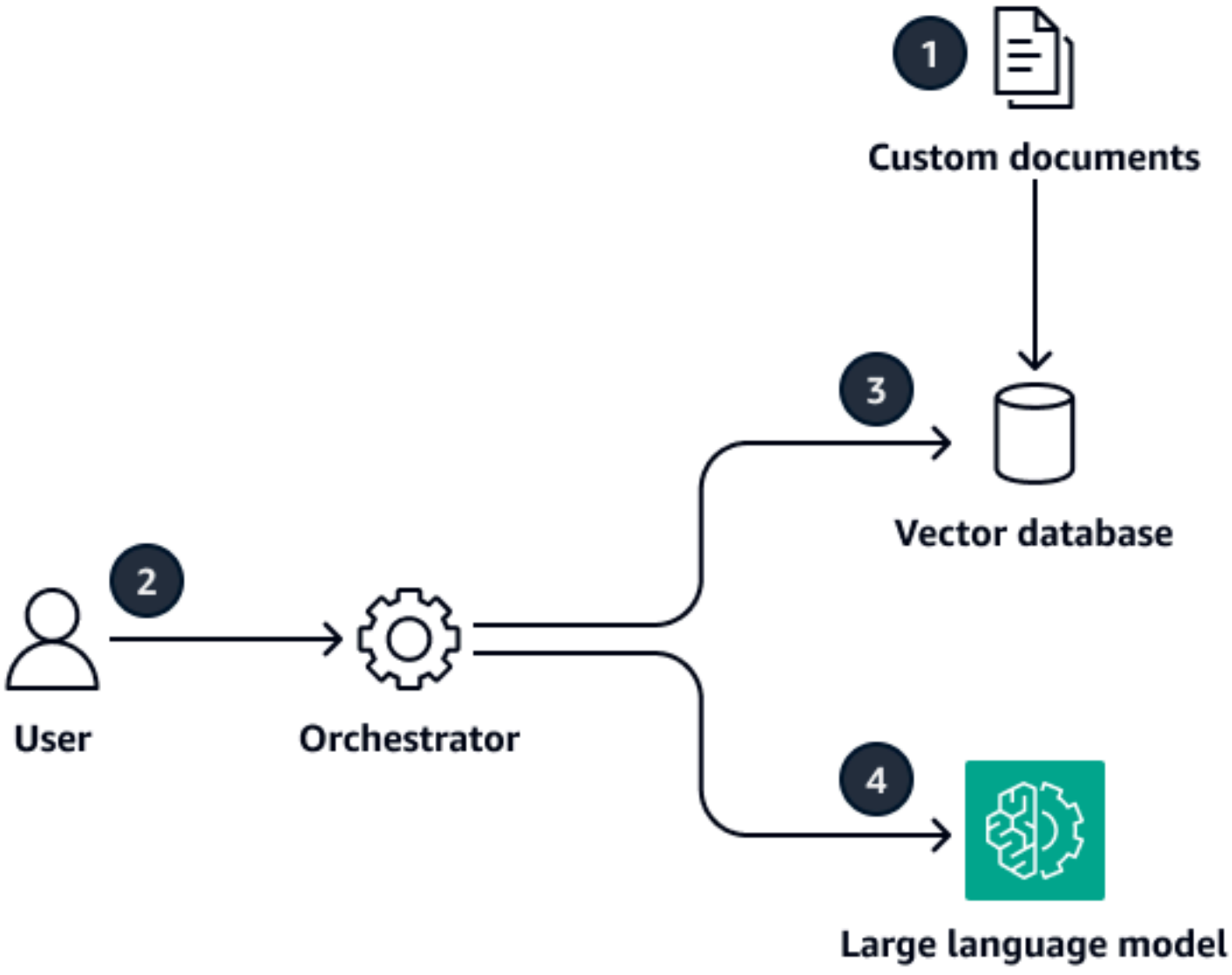
Die letzte Option ist die Verwendung von RAG. Bei RAG referenziert das Foundation-Modell Ihre benutzerdefinierten Dokumente, bevor eine Antwort generiert wird. RAG erweitert die Funktionen des Modells auf die interne Wissensdatenbank Ihres Unternehmens, ohne dass das Modell neu trainiert werden muss. Es ist ein kostengünstiger Ansatz zur Verbesserung der Modellergebnisse, sodass sie in verschiedenen Kontexten relevant, genau und nützlich bleiben.

Themen in diesem Abschnitt:

- Grundlegendes zu Retrieval Augmented Generation
- Vergleich von Retrieval, Augmented Generation und Feinabstimmung
- Anwendungsfälle für Retrieval Augmented Generation

Grundlegendes zu Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) ist eine Technik, mit der ein Large Language Model (LLM) um externe Daten, wie z. B. interne Dokumente eines Unternehmens, erweitert wird. Auf diese Weise erhält das Modell den Kontext, den es benötigt, um genaue und nützliche Ergebnisse für Ihren spezifischen Anwendungsfall zu erzeugen. RAG ist ein pragmatischer und effektiver Ansatz für den Einsatz LLMs in einem Unternehmen. Das folgende Diagramm zeigt einen allgemeinen Überblick über die Funktionsweise eines RAG-Ansatzes.



Im Großen und Ganzen besteht der RAG-Prozess aus vier Schritten. Der erste Schritt wird einmal ausgeführt, und die anderen drei Schritte werden so oft wie nötig ausgeführt:

1. Sie erstellen Einbettungen, um die internen Dokumente in eine Vektordatenbank aufzunehmen. Einbettungen sind numerische Darstellungen von Text in den Dokumenten, die die semantische oder kontextuelle Bedeutung der Daten erfassen. Eine Vektordatenbank ist im Wesentlichen eine Datenbank dieser Einbettungen und wird manchmal auch als Vektorspeicher oder Vektorindex bezeichnet. Dieser Schritt erfordert das Bereinigen, Formatieren und Aufteilen von Daten, aber dies ist eine einmalige Aktivität, die im Voraus erfolgt.
2. Ein Mensch reicht eine Anfrage in natürlicher Sprache ein.
3. Ein Orchestrator führt eine Ähnlichkeitssuche in der Vektordatenbank durch und ruft die relevanten Daten ab. Der Orchestrator fügt die abgerufenen Daten (auch als Kontext bezeichnet) zur Eingabeaufforderung hinzu, die die Abfrage enthält.
4. Der Orchestrator sendet die Abfrage und den Kontext an das LLM. Das LLM generiert eine Antwort auf die Abfrage, indem es den zusätzlichen Kontext verwendet.

Aus der Sicht eines Benutzers sieht es so aus, als würde RAG mit einem beliebigen LLM interagieren. Das System weiß jedoch viel mehr über die fraglichen Inhalte und bietet Antworten, die genau auf die Wissensbasis des Unternehmens abgestimmt sind.

Weitere Informationen zur Funktionsweise eines RAG-Ansatzes finden Sie auf der AWS Website unter Was ist RAG.

Komponenten von RAG-Systemen auf Produktionsebene

Der Aufbau eines RAG-Systems auf Produktionsebene erfordert das Durchdenken verschiedener Aspekte des RAG-Workflows. Konzeptionell erfordert ein RAG-Workflow auf Produktionsebene unabhängig von der spezifischen Implementierung die folgenden Funktionen und Komponenten:

- Konnektoren — Diese verbinden verschiedene Unternehmensdatenquellen mit der Vektordatenbank. Beispiele für strukturierte Datenquellen sind Transaktions- und Analysedatenbanken. Beispiele für unstrukturierte Datenquellen sind Objektspeicher, Codebasen und SaaS-Plattformen (Software as a Service). Für jede Datenquelle sind möglicherweise unterschiedliche Verbindungsmuster, Lizenzen und Konfigurationen erforderlich.
- Datenverarbeitung — Daten liegen in vielen Formen vor, z. B. als PDFs gescannte Bilder, Dokumente, Präsentationen und Microsoft SharePoint Dateien. Sie müssen Datenverarbeitungstechniken verwenden, um die Daten zu extrahieren, zu verarbeiten und für die Indizierung vorzubereiten.

- Einbettungen — Um eine Relevanzsuche durchzuführen, müssen Sie Ihre Dokumente und Benutzerabfragen in ein kompatibles Format konvertieren. Mithilfe von eingebetteten Sprachmodellen konvertieren Sie die Dokumente in eine numerische Darstellung. Dies sind im Wesentlichen Eingaben für das zugrunde liegende Fundamentmodell.
- Vektordatenbank — Die Vektordatenbank ist ein Index der Einbettungen, des zugehörigen Textes und der Metadaten. Der Index ist für die Suche und den Abruf optimiert.
- Retriever — Für die Benutzerabfrage ruft der Retriever den relevanten Kontext aus der Vektordatenbank ab und ordnet die Antworten auf der Grundlage der Geschäftsanforderungen.
- Basismodell — Das Basismodell für ein RAG-System ist in der Regel ein LLM. Durch die Verarbeitung des Kontextes und der Aufforderung generiert und formatiert das Foundation-Modell eine Antwort für den Benutzer.
- Leitplanken — Leitplanken sollen sicherstellen, dass die Anfrage, die Aufforderung, der abgerufene Kontext und die LLM-Antwort korrekt, verantwortungsbewusst, ethisch und frei von Halluzinationen und Vorurteilen sind.
- Orchestrator — Der Orchestrator ist für die Planung und Verwaltung des Workflows verantwortlich. end-to-end
- Benutzererfahrung — In der Regel interagiert der Benutzer mit einer Konversationsschnittstelle, die über umfangreiche Funktionen verfügt, darunter die Anzeige des Chat-Verlaufs und das Sammeln von Benutzerfeedback zu Antworten.
- Identitäts- und Benutzerverwaltung — Es ist wichtig, den Benutzerzugriff auf die Anwendung genau zu kontrollieren. In der werden Richtlinien AWS Cloud, Rollen und Berechtigungen in der Regel über AWS Identity and Access Management (IAM) verwaltet.

Es liegt auf der Hand, dass die Planung, Entwicklung, Veröffentlichung und Verwaltung eines RAG-Systems mit einem erheblichen Arbeitsaufwand verbunden ist. Vollständig verwaltete Services wie Amazon Bedrock oder Amazon Q Business können Ihnen helfen, einen Teil der undifferenzierten Schwerarbeit zu bewältigen. Benutzerdefinierte RAG-Architekturen können jedoch mehr Kontrolle über die Komponenten wie den Retriever oder die Vektordatenbank bieten.

Vergleich von Retrieval, Augmented Generation und Feinabstimmung

In der folgenden Tabelle werden die Vor- und Nachteile der Feinabstimmungs- und RAG-basierten Ansätze beschrieben.

Ansatz	Vorteile	Nachteile
Feinabstimmung	• Wenn ein fein abgestimmtes Modell mithilfe des unbeaufsichtigten Ansatzes trainiert wird, ist es in der Lage, Inhalte zu erstellen, die dem Stil Ihrer Organisation besser entsprechen. • Ein fein abgestimmtes Modell, das auf firmeneigenen oder regulatorischen Daten trainiert wurde, kann Ihrem Unternehmen helfen, interne oder branchenspezifische Daten- und Compliance-Standards einzuhalten.	• Die Feinabstimmung kann je nach Größe des Modells einige Stunden bis Tage dauern. Daher ist es keine gute Lösung, wenn sich Ihre benutzerdefinierten Dokumente häufig ändern. • Die Feinabstimmung erfordert ein Verständnis von Techniken wie Low-Rank Adaptation (LoRa) und parametereffizientes Fine-Tuning (PEFT). Für die Feinabstimmung ist möglicherweise ein Datenwissenschaftler erforderlich. • Die Feinabstimmung ist möglicherweise nicht für alle Modelle verfügbar. • Bei der Feinabstimmung von Modellen wird in ihren Antworten nicht auf die Quelle verwiesen. • Bei der Verwendung eines fein abgestimmten Modells zur Beantwortung von Fragen kann das Halluzinationsrisiko erhöht sein.
RAG	• Mit RAG können Sie ohne Feinabstimmung ein System zur Beantwortung von	• RAG funktioniert nicht gut, wenn es darum geht, Informationen aus ganzen

Ansatz	Vorteile	Nachteile
	Fragen für Ihre benutzerd efinierten Dokumente erstellen. • RAG kann die neuesten Dokumente in wenigen Minuten integrieren. • AWS bietet vollständig verwaltete RAG-Lösun gen. Daher sind weder Datenwissenschaftler noch Spezialkenntnisse im Bereich maschinelles Lernen erforderlich. • In seiner Antwort gibt ein RAG-Modell einen Verweis auf die Informationsquelle. • Da RAG den Kontext aus der Vektorsuche als Grundlage für die generierte Antwort verwendet, besteht ein geringeres Halluzina tionsrisiko.	Dokumenten zusammenz ufassen.

Wenn Sie eine Lösung zur Beantwortung von Fragen entwickeln müssen, die auf Ihre benutzerdefinierten Dokumente verweist, empfehlen wir Ihnen, von einem RAG-basierten Ansatz auszugehen. Verwenden Sie die Feinabstimmung, wenn Sie das Modell für zusätzliche Aufgaben, wie z. B. die Zusammenfassung, benötigen.

Sie können die Feinabstimmungs- und RAG-Ansätze in einem einzigen Modell kombinieren. In diesem Fall ändert sich die RAG-Architektur nicht, aber das LLM, das die Antwort generiert, wird ebenfalls an die benutzerdefinierten Dokumente angepasst. Dies kombiniert das Beste aus beiden Welten und könnte eine optimale Lösung für Ihren Anwendungsfall sein. Weitere Informationen zur

Kombination von überwachter Feinabstimmung mit RAG finden Sie in der Studie RAFT: Adapting Language Model to Domain Specific RAG von derUniversity of California, Berkeley.

Anwendungsfälle für Retrieval Augmented Generation

Im Folgenden sind gängige Anwendungsfälle für die Verwendung eines RAG-Ansatzes aufgeführt:

- Suchmaschinen — RAG-fähige Suchmaschinen können genauere und up-to-date aussagekräftigere Snippets in ihren Suchergebnissen bereitstellen.
- Systeme zur Beantwortung von Fragen — RAG kann die Qualität der Antworten in Systemen zur Beantwortung von Fragen verbessern. Das auf Abrufen basierende Modell verwendet die Ähnlichkeitssuche, um relevante Passagen oder Dokumente zu finden, die die Antwort enthalten. Anschließend generiert es auf der Grundlage dieser Informationen eine präzise und relevante Antwort.
- Einzelhandel oder E-Commerce — RAG kann das Benutzererlebnis im E-Commerce verbessern, indem es relevantere und personalisierte Produktempfehlungen gibt. Durch das Abrufen und Integrieren von Informationen über Benutzerpräferenzen und Produktdetails kann RAG genauere und hilfreichere Empfehlungen für Kunden erstellen.
- Industrie oder Fertigung — In der Fertigung hilft Ihnen RAG dabei, schnell auf wichtige Informationen zuzugreifen, z. B. über den Betrieb von Fabrikanlagen. Es kann auch bei Entscheidungsprozessen, bei der Fehlerbehebung und bei organisatorischen Innovationen helfen. Für Hersteller, die innerhalb strenger regulatorischer Rahmenbedingungen arbeiten, kann RAG schnell aktualisierte Vorschriften und Compliance-Standards aus internen und externen Quellen abrufen, z. B. aus Industriestandards oder Aufsichtsbehörden.
- Gesundheitswesen — RAG hat Potenzial in der Gesundheitsbranche, wo der Zugang zu genauen und aktuellen Informationen von entscheidender Bedeutung ist. Durch das Abrufen und Integrieren von relevantem medizinischem Wissen aus externen Quellen kann RAG genauere und kontextsensivere Antworten in Anwendungen im Gesundheitswesen bereitstellen. Solche Anwendungen erweitern die Informationen, auf die ein menschlicher Arzt zugreifen kann, der letztlich die Entscheidung trifft und nicht das Modell.
- Rechtliches — RAG kann hervorragend in rechtlichen Szenarien wie Fusionen und Übernahmen eingesetzt werden, bei denen komplexe Rechtsdokumente den Kontext für Anfragen bieten. Dies kann Juristen helfen, komplexe regulatorische Probleme schnell zu lösen.

Vollständig verwaltete Optionen für Retrieval Augmented Generation auf AWS

Um Workflows mit Retrieval Augmented Generation (RAG) zu verwalten AWS, können Sie benutzerdefinierte RAG-Pipelines verwenden oder einige der Funktionen der vollständig verwalteten Dienste nutzen, die die Lösung bietet. AWS Da sie viele der Kernkomponenten eines RAG-basierten Systems enthalten, können Fully Managed Services Ihnen helfen, einen Teil der undifferenzierten Schwerarbeit zu bewältigen. Diese Dienste bieten jedoch weniger Möglichkeiten zur Anpassung.

Die vollständig verwalteten AWS-Services Systeme verwenden Konnektoren, um Daten aus externen Datenquellen wie Websites, Atlassian Confluence oder Microsoft aufzunehmen. SharePoint Die unterstützten Datenquellen variieren je nach. AWS-Service

In diesem Abschnitt werden die folgenden vollständig verwalteten Optionen für die Erstellung von RAG-Workflows untersucht AWS:

- Wissensdatenbanken für Amazon Bedrock
- Amazon Q Business
- Amazon SageMaker AI-Leinwand

Weitere Informationen darüber, wie Sie zwischen diesen Optionen wählen können, finden Sie Wählen Sie eine Option zum Abrufen erweiterter Generierung auf AWS in diesem Handbuch.

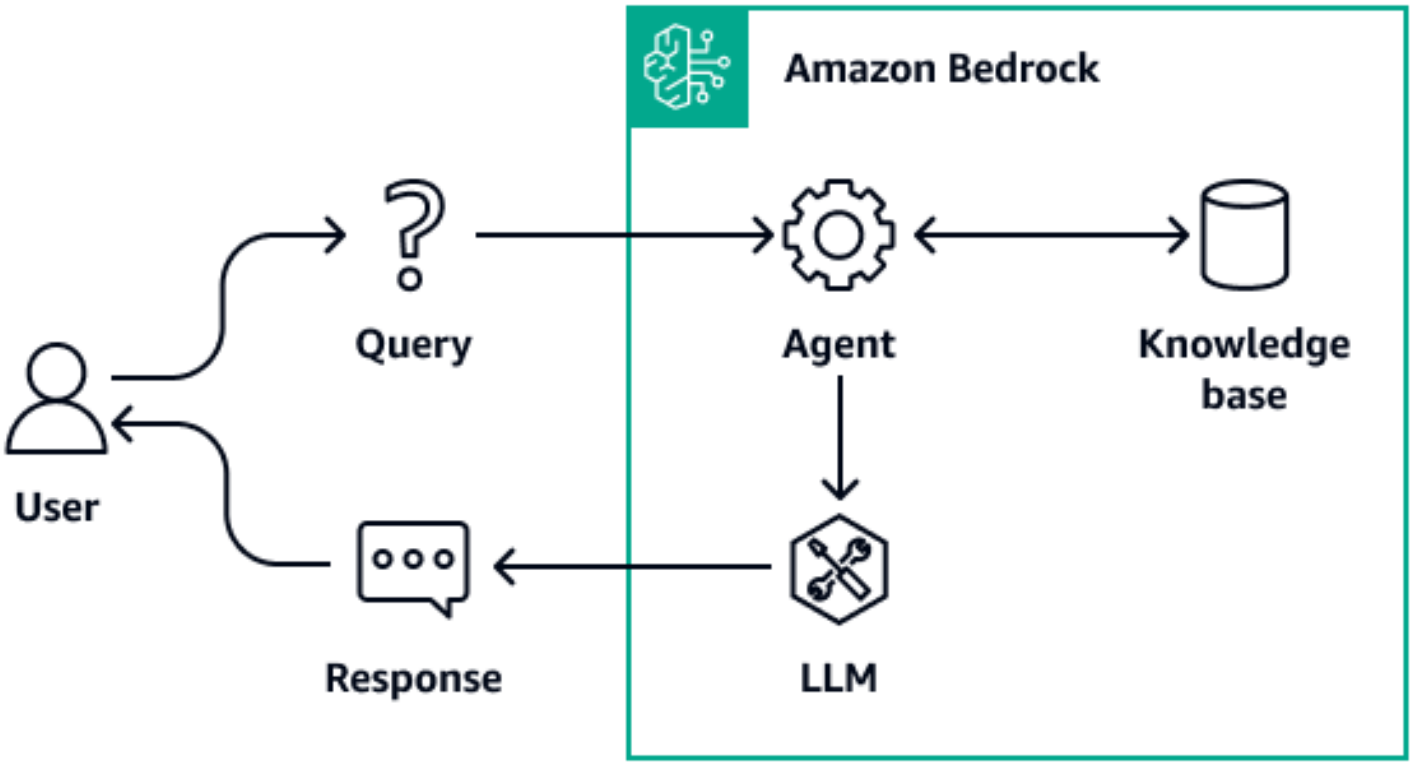
Wissensdatenbanken für Amazon Bedrock

Amazon Bedrock ist ein vollständig verwalteter Service, der Ihnen leistungsstarke Basismodelle (FMs) von führenden KI-Startups und Amazon über eine einheitliche API zur Verfügung stellt. Knowledge Bases ist eine Funktion von Amazon Bedrock, mit der Sie den gesamten RAG-Workflow implementieren können, von der Aufnahme über den Abruf bis hin zur sofortigen Erweiterung. Es ist nicht erforderlich, benutzerdefinierte Integrationen für Datenquellen zu erstellen oder Datenflüsse zu verwalten. Das Sitzungskontextmanagement ist integriert, sodass Ihre generative KI-Anwendung problemlos Multi-Turn-Konversationen unterstützen kann.

Nachdem Sie den Speicherort Ihrer Daten angegeben haben, ruft Knowledge Bases für Amazon Bedrock die Dokumente intern ab, teilt sie in Textblöcke auf, konvertiert den Text in Einbettungen

und speichert die Einbettungen dann in der Vektordatenbank Ihrer Wahl. Amazon Bedrock verwaltet und aktualisiert die Einbettungen und sorgt dafür, dass die Vektordatenbank mit den Daten synchron bleibt. Weitere Informationen zur Funktionsweise von Wissensdatenbanken finden Sie unter So funktionieren Amazon Bedrock-Wissensdatenbanken.

Wenn Sie einem Amazon Bedrock-Agenten Wissensdatenbanken hinzufügen, identifiziert der Agent die entsprechende Wissensdatenbank auf der Grundlage der Benutzereingabe. Der Agent ruft die relevanten Informationen ab und fügt sie der Eingabeaufforderung hinzu. Die aktualisierte Aufforderung stellt dem Modell mehr Kontextinformationen zur Verfügung, um eine Antwort zu generieren. Um die Transparenz zu erhöhen und Halluzinationen zu minimieren, können die aus der Wissensdatenbank abgerufenen Informationen bis zu ihrer Quelle zurückverfolgt werden.



Amazon Bedrock unterstützt die folgenden beiden APIs für RAG:

- RetrieveAndGenerate— Sie können diese API verwenden, um Ihre Wissensdatenbank abzufragen und anhand der abgerufenen Informationen Antworten zu generieren. Intern konvertiert Amazon Bedrock die Abfragen in Einbettungen, fragt die Wissensdatenbank ab, ergänzt die Eingabeaufforderung mit den Suchergebnissen als Kontextinformationen und gibt die vom LLM generierte Antwort zurück. Amazon Bedrock verwaltet auch das Kurzzeitgedächtnis der Konversation, um kontextbezogenere Ergebnisse zu erzielen.

- Abrufen — Sie können diese API verwenden, um Ihre Wissensdatenbank mit Informationen abzufragen, die direkt aus der Wissensdatenbank abgerufen wurden. Sie können die von dieser API zurückgegebenen Informationen verwenden, um den abgerufenen Text zu verarbeiten, seine Relevanz zu bewerten oder einen separaten Workflow für die Antwortgenerierung zu entwickeln. Intern wandelt Amazon Bedrock die Abfragen in Einbettungen um, durchsucht die Wissensdatenbank und gibt die entsprechenden Ergebnisse zurück. Sie können zusätzliche Workflows zusätzlich zu den Suchergebnissen erstellen. Sie können das LangChain`AmazonKnowledgeBasesRetriever`Plugin beispielsweise verwenden, um RAG-Workflows in generative KI-Anwendungen zu integrieren.

Architekturmuster und step-by-step Anleitungen zur Verwendung von finden Sie unter Knowledge Bases now provides fully managed RAG experience in Amazon Bedrock (AWS Blogbeitrag). APIs Weitere Informationen zur Verwendung der `RetrieveAndGenerate` API zum Erstellen eines RAG-Workflows für eine intelligente Chat-basierte Anwendung finden Sie unter Erstellen einer kontextbezogenen Chatbot-Anwendung mithilfe von Amazon Bedrock Knowledge Bases (Blogbeitrag).AWS

Datenquellen für Wissensdatenbanken

Sie können Ihre eigenen Daten mit einer Wissensdatenbank verbinden. Nachdem Sie einen Datenquellen-Connector konfiguriert haben, können Sie Ihre Daten mit Ihrer Wissensdatenbank synchronisieren oder auf dem neuesten Stand halten und Ihre Daten für Abfragen zur Verfügung stellen. Amazon Bedrock Knowledge Bases unterstützen Verbindungen zu den folgenden Datenquellen:

- Amazon Simple Storage Service (Amazon S3) — Sie können einen Amazon S3 S3-Bucket mit einer Amazon Bedrock-Wissensdatenbank verbinden, indem Sie entweder die Konsole oder die API verwenden. Die Wissensdatenbank nimmt die Dateien im Bucket auf und indexiert sie. Diese Art von Datenquelle unterstützt die folgenden Funktionen:
  - Metadatenfelder für Dokumente — Sie können eine separate Datei hinzufügen, um die Metadaten für die Dateien im Amazon S3 S3-Bucket anzugeben. Sie können diese Metadatenfelder dann verwenden, um die Relevanz von Antworten zu filtern und zu verbessern.
  - Inklusions- oder Ausschlussfilter — Sie können beim Crawlen bestimmte Inhalte ein- oder ausschließen.
  - Inkrementelle Synchronisierung — Die Inhaltsänderungen werden nachverfolgt, und es werden nur Inhalte gecrawlt, die sich seit der letzten Synchronisierung geändert haben.

- Confluence— Sie können eine Atlassian Confluence Instance mithilfe der Konsole oder der API mit einer Amazon Bedrock-Wissensdatenbank verbinden. Diese Art von Datenquelle unterstützt die folgenden Funktionen:
  - Automatische Erkennung der wichtigsten Dokumentfelder — Die Metadatenfelder werden automatisch erkannt und gecrawlt. Sie können diese Felder zum Filtern verwenden.
  - Inhaltsfilter zum Ein- oder Ausschließen — Sie können bestimmte Inhalte ein- oder ausschließen, indem Sie ein Präfix oder ein Muster mit regulären Ausdrücken für den Bereich, den Seitentitel, den Blogtitel, den Kommentar, den Namen des Anhangs oder die Erweiterung verwenden.
  - Inkrementelle Synchronisierung — Die Inhaltsänderungen werden nachverfolgt, und es werden nur Inhalte gecrawlt, die sich seit der letzten Synchronisierung geändert haben.
  - OAuth 2.0-Authentifizierung, Authentifizierung mit Confluence API-Token — Die Authentifizierungsdaten werden in gespeichert. AWS Secrets Manager
- Microsoft SharePoint— Sie können eine SharePoint Instanz mit einer Wissensdatenbank verbinden, indem Sie entweder die Konsole oder die API verwenden. Diese Art von Datenquelle unterstützt die folgenden Funktionen:
  - Automatische Erkennung der wichtigsten Dokumentfelder — Die Metadatenfelder werden automatisch erkannt und gecrawlt. Sie können diese Felder zum Filtern verwenden.
  - Inhaltsfilter zum Ein- oder Ausschließen — Sie können bestimmte Inhalte ein- oder ausschließen, indem Sie ein Präfix oder ein Muster mit regulären Ausdrücken für den Titel der Hauptseite, den Veranstaltungsnamen und den Dateinamen (einschließlich der Erweiterung) verwenden.
  - Inkrementelle Synchronisierung — Die Inhaltsänderungen werden nachverfolgt, und es werden nur Inhalte gecrawlt, die sich seit der letzten Synchronisierung geändert haben.
  - OAuth 2.0-Authentifizierung — Die Authentifizierungsdaten werden in gespeichert. AWS Secrets Manager
- Salesforce— Sie können eine Salesforce Instanz mit einer Wissensdatenbank verbinden, indem Sie entweder die Konsole oder die API verwenden. Diese Art von Datenquelle unterstützt die folgenden Funktionen:
  - Automatische Erkennung der wichtigsten Dokumentfelder — Die Metadatenfelder werden automatisch erkannt und gecrawlt. Sie können diese Felder zum Filtern verwenden.
  - Inhaltsfilter zum Einschließen oder Ausschließen — Sie können bestimmte Inhalte mit einem Präfix oder einem Muster für reguläre Ausdrücke ein- oder ausschließen. Eine Liste der Inhaltstypen, auf die Sie Filter anwenden können, finden Sie unter Einschluss-/Ausschlussfilter in der Amazon Bedrock-Dokumentation.

  - Inkrementelle Synchronisierung — Die Inhaltsänderungen werden nachverfolgt, und es werden nur Inhalte gecrawlt, die sich seit der letzten Synchronisierung geändert haben.
  - OAuth 2.0-Authentifizierung — Die Authentifizierungsdaten werden in gespeichert. AWS Secrets Manager
- Webcrawler — Ein Amazon Bedrock Web Crawler stellt eine Verbindung zu den von Ihnen bereitgestellten Daten her und crawlt diese. URLs Die folgenden Funktionen werden unterstützt:
  - Wählen Sie mehrere URLs zum Crawlen aus
  - Beachten Sie die Standardanweisungen von robots.txt wie und `Allow` `Disallow`
  - Schließt aus URLs , die einem Muster entsprechen
  - Beschränken Sie die Crawling-Rate
  - Sehen Sie CloudWatch sich in Amazon den Status jeder gecrawlten URL an

Weitere Informationen zu den Datenquellen, die Sie mit Ihrer Amazon Bedrock-Wissensdatenbank verbinden können, finden Sie unter Erstellen eines Datenquellen-Connectors für Ihre Wissensdatenbank.

Vektor-Datenbanken für Wissensdatenbanken

Wenn Sie eine Verbindung zwischen der Wissensdatenbank und der Datenquelle einrichten, müssen Sie eine Vektordatenbank konfigurieren, die auch als Vektorspeicher bezeichnet wird. In einer Vektordatenbank speichert, aktualisiert und verwaltet Amazon Bedrock die Einbettungen, die Ihre Daten repräsentieren. Jede Datenquelle unterstützt verschiedene Arten von Vektordatenbanken. Informationen darüber, welche Vektordatenbanken für Ihre Datenquelle verfügbar sind, finden Sie unter Datenquellentypen.

Wenn Sie es vorziehen, dass Amazon Bedrock automatisch eine Vektordatenbank in Amazon OpenSearch Serverless für Sie erstellt, können Sie diese Option bei der Erstellung der Wissensdatenbank wählen. Sie können sich jedoch auch dafür entscheiden, Ihre eigene Vektordatenbank einzurichten. Wenn Sie Ihre eigene Vektordatenbank einrichten, finden Sie eine Wissensdatenbank unter Voraussetzungen für Ihren eigenen Vektorspeicher. Jeder Vektordatenbanktyp hat seine eigenen Voraussetzungen.

Abhängig von Ihrem Datenquellentyp unterstützen die Amazon Bedrock-Wissensdatenbanken die folgenden Vektordatenbanken:

- Amazon OpenSearch Serverlos

- Amazon Aurora PostgreSQL-Compatible Edition
- Pinecone(PineconeDokumentation)
- Redis Enterprise Cloud(RedisDokumentation)
- MongoDB Atlas(MongoDBDokumentation)

Amazon Q Business

Amazon Q Business ist ein vollständig verwalteter, auf generativer KI basierender Assistent, den Sie so konfigurieren können, dass er Fragen beantwortet, Zusammenfassungen bereitstellt, Inhalte generiert und Aufgaben auf der Grundlage Ihrer Unternehmensdaten erledigt. Es ermöglicht Endbenutzern, sofortige, berechtigungsabhängige Antworten aus Unternehmensdatenquellen mit Quellenangaben zu erhalten.

Schlüssel-Features

Die folgenden Funktionen von Amazon Q Business können Ihnen helfen, eine produktionstaugliche RAG-basierte generative KI-Anwendung zu entwickeln:

- Integrierte Konnektoren — Amazon Q Business unterstützt mehr als 40 Arten von Konnektoren, z. B. Konnektoren für Adobe Experience Manager (AEM)Salesforce,Jira, undMicrosoft SharePoint. Eine vollständige Liste finden Sie unter Unterstützte Konnektoren. Wenn Sie einen Connector benötigen, der nicht unterstützt wird, können Sie Amazon verwenden, AppFlow um Daten aus Ihrer Datenquelle in Amazon Simple Storage Service (Amazon S3) abzurufen und dann Amazon Q Business mit dem Amazon S3-Bucket zu verbinden. Eine vollständige Liste der Datenquellen, die Amazon AppFlow unterstützt, finden Sie unter Unterstützte Anwendungen.
- Integrierte Indexierungs-Pipelines — Amazon Q Business bietet eine integrierte Pipeline für die Indizierung von Daten in einer Vektordatenbank. Sie können eine AWS Lambda Funktion verwenden, um Vorverarbeitungslogik für Ihre Indexierungspipeline hinzuzufügen.
- Indexoptionen — Sie können einen systemeigenen Index in Amazon Q Business erstellen und bereitstellen, und Sie verwenden einen Amazon Q Business Retriever, um Daten aus diesem Index abzurufen. Alternativ können Sie einen vorkonfigurierten Amazon Kendra Kendra-Index als Retriever verwenden. Weitere Informationen finden Sie unter Einen Retriever für eine Amazon Q Business-Anwendung erstellen.

- Foundation-Modelle — Amazon Q Business verwendet die Foundation-Modelle, die in Amazon Bedrock unterstützt werden. Eine vollständige Liste finden Sie unter Unterstützte Foundation-Modelle in Amazon Bedrock.
- Plugins — Amazon Q Business bietet die Möglichkeit, Plugins zur Integration in Zielsysteme zu verwenden, z. B. eine automatisierte Methode zur Zusammenfassung von Ticketinformationen und zur Ticketerstellung inJira. Nach der Konfiguration können Plugins Lese- und Schreibaktionen unterstützen, mit denen Sie die Produktivität der Endbenutzer steigern können. Amazon Q Business unterstützt zwei Arten von Plug-ins: integrierte Plug-ins und benutzerdefinierte Plug-ins.
- Guardrails — Amazon Q Business unterstützt globale Kontrollen und Kontrollen auf Themenebene. Diese Kontrollen können beispielsweise personenbezogene Daten (PII), Missbrauch oder vertrauliche Informationen in Eingabeaufforderungen erkennen. Weitere Informationen finden Sie unter Administratorkontrollen und Leitplanken in Amazon Q Business.
- Identitätsmanagement — Mit Amazon Q Business können Sie Benutzer und ihren Zugriff auf die RAG-basierte generative KI-Anwendung verwalten. Weitere Informationen finden Sie unter Identitäts- und Zugriffsverwaltung für Amazon Q Business. Außerdem indexieren Amazon Q Business Connectors Informationen zur Zugriffskontrollliste (ACL), die zusammen mit dem Dokument selbst an ein Dokument angehängt sind. Anschließend speichert Amazon Q Business die von ihm indizierten ACL-Informationen im Amazon Q Business User Store, um Benutzer- und Gruppenzuordnungen zu erstellen und Chat-Antworten basierend auf dem Zugriff des Endbenutzers auf Dokumente zu filtern. Weitere Informationen finden Sie unter Konzepte für Datenquellenkonnektoren.
- Anreicherung von Dokumenten — Mit der Funktion zur Anreicherung von Dokumenten können Sie kontrollieren, welche Dokumente und Dokumentattribute in Ihren Index aufgenommen werden und wie sie aufgenommen werden. Dies kann auf zwei Arten erreicht werden:
  - Grundoperationen konfigurieren — Verwenden Sie grundlegende Operationen, um Dokumentattribute zu Ihren Daten hinzuzufügen, zu aktualisieren oder zu löschen. Sie können beispielsweise personenbezogene Daten löschen, indem Sie alle Dokumentattribute löschen, die sich auf personenbezogene Daten beziehen.
  - Lambda-Funktionen konfigurieren — Verwenden Sie eine vorkonfigurierte Lambda-Funktion, um eine individuellere, erweiterte Logik zur Manipulation von Dokumentattributen auf Ihre Daten anzuwenden. Beispielsweise können Ihre Unternehmensdaten als gescannte Bilder gespeichert werden. In diesem Fall können Sie eine Lambda-Funktion verwenden, um die optische Zeichenerkennung (OCR) für die gescannten Dokumente auszuführen, um Text aus ihnen zu extrahieren. Anschließend wird jedes gescannte Dokument bei der Aufnahme als

Textdokument behandelt. Schließlich berücksichtigt Amazon Q während des Chats die aus den gescannten Dokumenten extrahierten Textdaten, wenn es Antworten generiert.

Bei der Implementierung Ihrer Lösung können Sie wählen, ob Sie beide Ansätze zur Anreicherung von Dokumenten kombinieren möchten. Sie können grundlegende Operationen verwenden, um eine erste Analyse Ihrer Daten durchzuführen, und dann eine Lambda-Funktion für komplexere Operationen verwenden. Weitere Informationen finden Sie unter Anreicherung von Dokumenten in Amazon Q Business.

- Integration — Nachdem Sie Ihre Amazon Q Business-Anwendung erstellt haben, können Sie sie in andere Anwendungen wie Slack oder integrierenMicrosoft Teams. Siehe beispielsweise Bereitstellen eines Slack Gateways für Amazon Q Business und Bereitstellen eines Microsoft Teams Gateways für Amazon Q Business (AWS Blogbeiträge).

Anpassung durch Endbenutzer

Amazon Q Business unterstützt das Hochladen von Dokumenten, die möglicherweise nicht in den Datenquellen und im Index Ihrer Organisation gespeichert sind. Hochgeladene Dokumente werden nicht gespeichert. Sie können nur für die Konversation verwendet werden, in der die Dokumente hochgeladen werden. Amazon Q Business unterstützt bestimmte Dokumenttypen für das Hochladen. Weitere Informationen finden Sie unter Dateien hochladen und chatten in Amazon Q Business.

Amazon Q Business beinhaltet eine Funktion zum Filtern nach Dokumentattributen. Sowohl Administratoren als auch Endbenutzer können diese Funktion verwenden. Administratoren können die Chat-Antworten für Endbenutzer mithilfe von Attributen anpassen und steuern. Wenn es sich bei dem Datenquellentyp beispielsweise um ein Attribut handelt, das an Ihre Dokumente angehängt ist, können Sie angeben, dass Chat-Antworten nur aus einer bestimmten Datenquelle generiert werden. Oder Sie können Endbenutzern ermöglichen, den Umfang der Chat-Antworten einzuschränken, indem Sie die von Ihnen ausgewählten Attributfilter verwenden.

Endbenutzer können schlanke, speziell entwickelte Amazon Q Apps in Ihrer breiteren Amazon Q Business-Anwendungsumgebung erstellen. Amazon Q-Apps ermöglichen die Automatisierung von Aufgaben für eine bestimmte Domain, z. B. eine speziell für das Marketingteam entwickelte App.

Amazon SageMaker AI-Leinwand

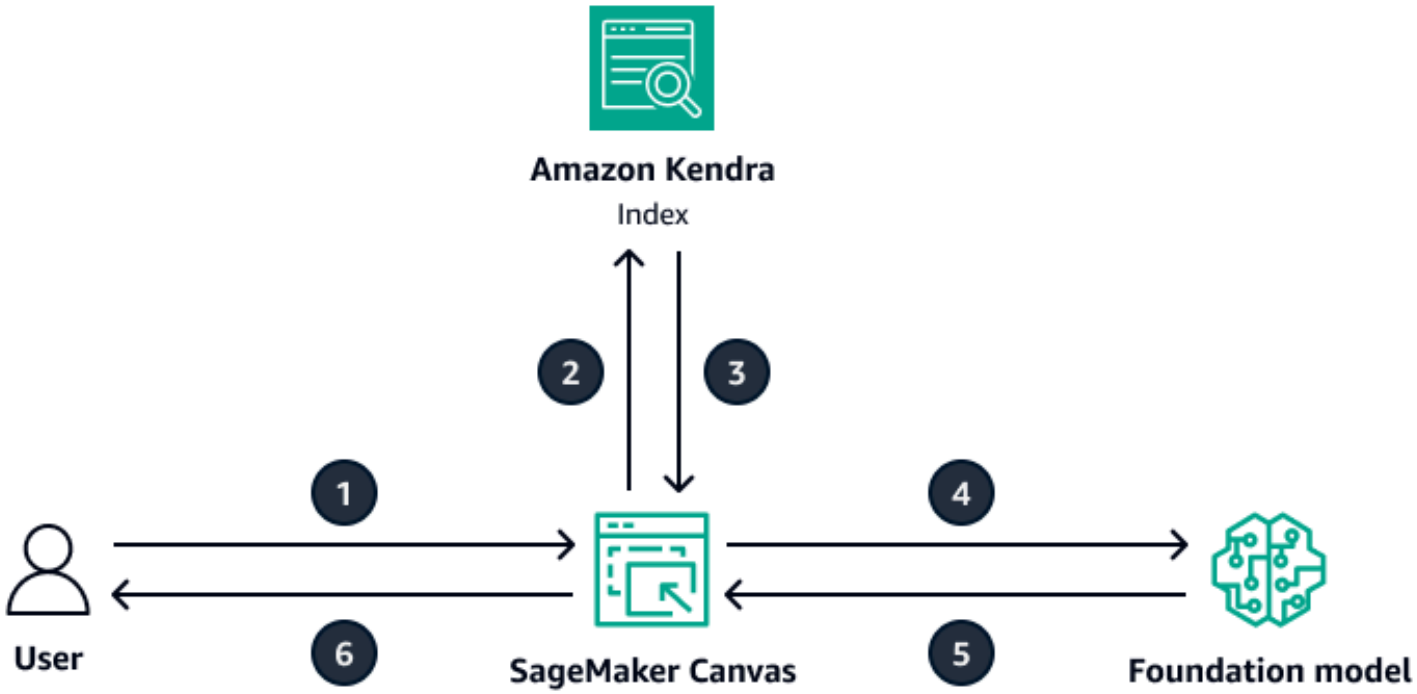
Amazon SageMaker AI Canvas hilft Ihnen, mithilfe von maschinellem Lernen Vorhersagen zu generieren, ohne Code schreiben zu müssen. Es bietet eine visuelle Oberfläche ohne Code,

mit der Sie Daten vorbereiten, ML-Modelle erstellen und bereitstellen und so den end-to-end ML-Lebenszyklus in einer einheitlichen Umgebung optimieren können. Die Komplexität der Datenaufbereitung, Modellentwicklung, Erkennung von Verzerrungen, Erklärbarkeit und Überwachung wird hinter einer intuitiven Oberfläche zusammengefasst. Benutzer müssen keine Experten für SageMaker KI oder maschinelles Lernen (MLOps) sein, um Modelle mit AI Canvas zu entwickeln, zu operationalisieren und zu überwachen. SageMaker

Bei SageMaker AI Canvas wird die RAG-Funktionalität über eine Funktion zur Dokumentenabfrage ohne Code bereitgestellt. Sie können das Chat-Erlebnis in SageMaker AI Canvas bereichern, indem Sie einen Amazon Kendra Kendra-Index als zugrunde liegende Unternehmenssuche verwenden. Weitere Informationen finden Sie unter Extrahieren von Informationen aus Dokumenten mit Dokumentenabfrage.

Die Verbindung von SageMaker AI Canvas mit dem Amazon Kendra Kendra-Index erfordert eine einmalige Einrichtung. Im Rahmen der Domain-Konfiguration kann ein Cloud-Administrator einen oder mehrere Kendra-Indizes auswählen, die der Benutzer bei der Interaktion mit SageMaker Canvas abfragen kann. Anweisungen zur Aktivierung der Dokumentabfragefunktion finden Sie unter Erste Schritte mit der Verwendung von Amazon SageMaker AI Canvas.

SageMaker AI Canvas verwaltet die zugrunde liegende Kommunikation zwischen Amazon Kendra und dem ausgewählten Foundation-Modell. Weitere Informationen zu den Basismodellen, die SageMaker AI Canvas unterstützt, finden Sie unter Generative KI-Grundmodelle in SageMaker AI Canvas. Das folgende Diagramm zeigt, wie die Funktion zur Dokumentenabfrage funktioniert, nachdem der Cloud-Administrator SageMaker AI Canvas mit einem Amazon Kendra Kendra-Index verbunden hat.



Das Diagramm zeigt den folgenden Workflow:

1. Der Benutzer startet einen neuen Chat in SageMaker AI Canvas, aktiviert Query documents, wählt den Zielindex aus und reicht dann eine Frage ein.
2. SageMaker AI Canvas verwendet die Abfrage, um den Amazon Kendra Kendra-Index nach relevanten Daten zu durchsuchen.
3. SageMaker AI Canvas ruft die Daten und ihre Quellen aus dem Amazon Kendra Kendra-Index ab.
4. SageMaker AI Canvas aktualisiert die Aufforderung, um den abgerufenen Kontext aus dem Amazon Kendra Kendra-Index aufzunehmen, und leitet die Aufforderung an das Foundation-Modell weiter.
5. Das Foundation-Modell verwendet die ursprüngliche Frage und den abgerufenen Kontext, um eine Antwort zu generieren.
6. SageMaker AI Canvas stellt dem Benutzer die generierte Antwort zur Verfügung. Es enthält Verweise auf die Datenquellen, z. B. Dokumente, die zur Generierung der Antwort verwendet wurden.

Custom Retrieval Augmented Generation-Architekturen auf AWS

Im vorherigen Abschnitt wird beschrieben, wie Sie eine vollständig verwaltete Augmented Generation (RAG) AWS-Service für Retrieval verwenden. In einigen Anwendungsfällen ist jedoch mehr Kontrolle über die Systemkomponenten wie den Retriever oder das LLM (auch Generator genannt) erforderlich. Beispielsweise benötigen Sie möglicherweise die Flexibilität, Ihre eigene Vektordatenbank auszuwählen oder auf eine nicht unterstützte Datenquelle zuzugreifen. Für diese Anwendungsfälle können Sie eine benutzerdefinierte RAG-Architektur erstellen.

In diesem Abschnitt werden folgende Themen behandelt:

- Retriever für RAG-Workflows
- Generatoren für RAG-Workflows

Weitere Informationen zur Auswahl zwischen den Optionen Retriever und Generator in diesem Abschnitt finden Sie Wählen Sie eine Option zum Abrufen erweiterter Generierung auf AWS in diesem Handbuch.

Retriever für RAG-Workflows

In diesem Abschnitt wird erklärt, wie Sie einen Retriever erstellen. Sie können eine vollständig verwaltete semantische Suchlösung wie Amazon Kendra verwenden oder mithilfe einer Vektordatenbank eine benutzerdefinierte semantische Suche erstellen. AWS

Bevor Sie sich mit den Retrieveroptionen befassen, stellen Sie sicher, dass Sie die drei Schritte des Vektorsuchprozesses verstanden haben:

1. Sie teilen die Dokumente, die indexiert werden müssen, in kleinere Teile auf. Dies wird als Chunking bezeichnet.
2. Sie verwenden einen Prozess namens Einbetten, um jeden Chunk in einen mathematischen Vektor umzuwandeln. Anschließend indizieren Sie jeden Vektor in einer Vektordatenbank. Der Ansatz, mit dem Sie die Dokumente indizieren, beeinflusst die Geschwindigkeit und Genauigkeit der Suche. Der Indizierungsansatz hängt von der Vektordatenbank und den von ihr bereitgestellten Konfigurationsoptionen ab.

3. Sie konvertieren die Benutzerabfrage mit demselben Verfahren in einen Vektor. Der Retriever durchsucht die Vektordatenbank nach Vektoren, die dem Abfragevektor des Benutzers ähnlich sind. Die Ähnlichkeit wird anhand von Metriken wie der euklidischen Distanz, der Kosinusdistanz oder dem Punktprodukt berechnet.

In diesem Handbuch wird beschrieben, wie Sie mit den folgenden Diensten AWS-Services oder Diensten von Drittanbietern eine benutzerdefinierte Abruf-Ebene erstellen können: AWS

- Amazon Kendra
- OpenSearch Amazon-Dienst
- Amazon Aurora PostgreSQL und pgvector
- Amazon Neptune Analytics
- Amazon MemoryDB
- Amazon DocumentDB
- Pinecone
- MongoDB Atlas
- Weaviate

Amazon Kendra

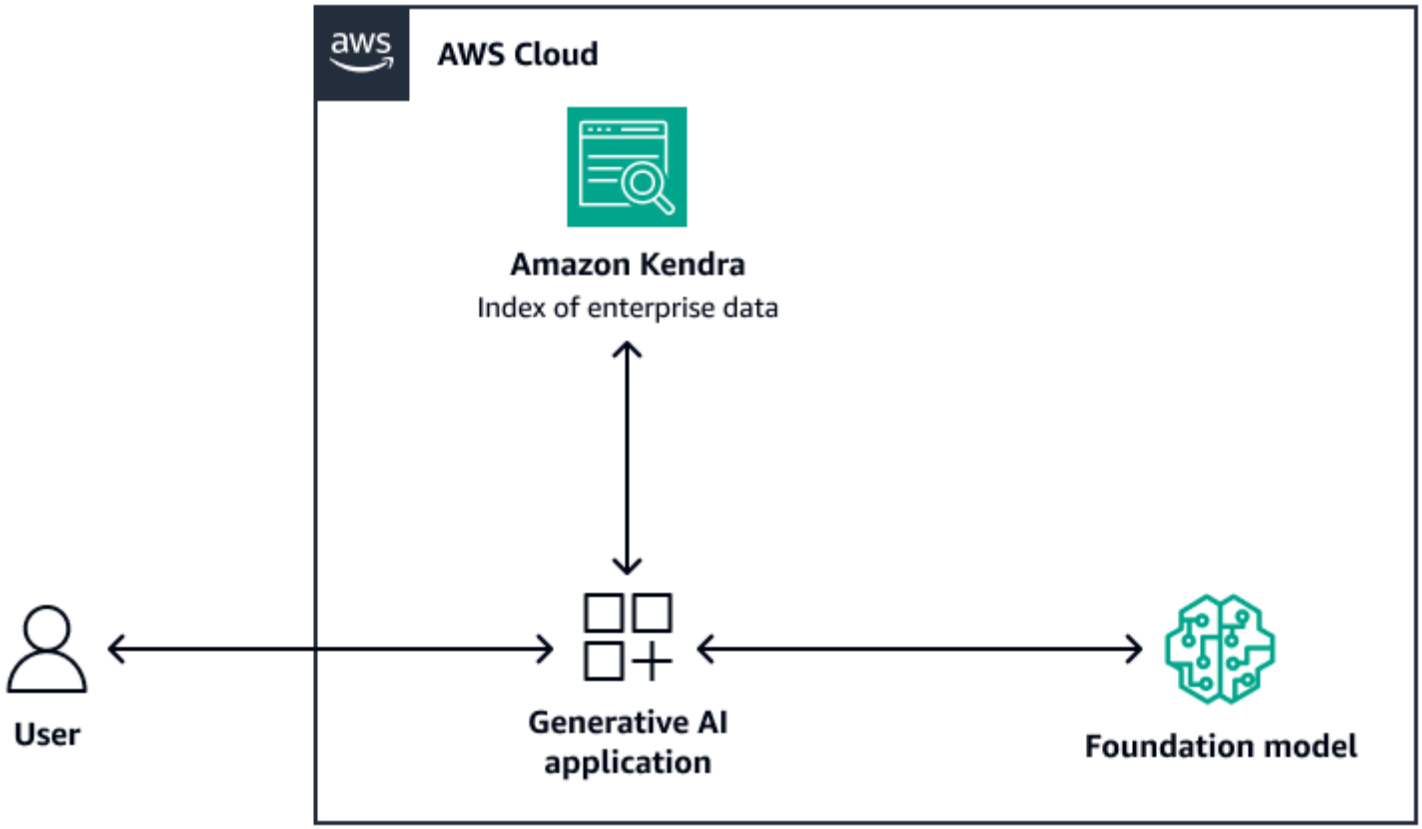
Amazon Kendra ist ein vollständig verwalteter, intelligenter Suchservice, der natürliche Sprachverarbeitung und fortschrittliche Algorithmen für maschinelles Lernen verwendet, um spezifische Antworten auf Suchfragen aus Ihren Daten zurückzugeben. Amazon Kendra hilft Ihnen dabei, Dokumente aus mehreren Quellen direkt aufzunehmen und die Dokumente abzufragen, nachdem sie erfolgreich synchronisiert wurden. Durch den Synchronisierungsprozess wird die erforderliche Infrastruktur geschaffen, um eine Vektorsuche im aufgenommenen Dokument zu erstellen. Daher benötigt Amazon Kendra nicht die traditionellen drei Schritte des Vektorsuchprozesses. Nach der ersten Synchronisierung können Sie einen definierten Zeitplan verwenden, um die laufende Datenaufnahme zu handhaben.

Im Folgenden sind die Vorteile der Verwendung von Amazon Kendra for RAG aufgeführt:

- Sie müssen keine Vektordatenbank verwalten, da Amazon Kendra den gesamten Vektorsuchprozess abwickelt.

- Amazon Kendra enthält vorgefertigte Konnektoren für beliebte Datenquellen wie Datenbanken, Website-Crawler, Amazon S3 S3-Buckets, Microsoft SharePoint Instances und Instances. Atlassian Confluence Von AWS Partnern entwickelte Konnektoren sind verfügbar, z. B. Konnektoren für und. Box GitLab
- Amazon Kendra bietet eine ACL-Filterung (Access Control List), die nur Dokumente zurückgibt, auf die der Endbenutzer Zugriff hat.
- Amazon Kendra kann Antworten auf der Grundlage von Metadaten wie Datum oder Quell-Repository beschleunigen.

Die folgende Abbildung zeigt eine Beispielarchitektur, die Amazon Kendra als Abruf-Ebene des RAG-Systems verwendet. Weitere Informationen finden Sie unter Schnelles Erstellen hochgenauer generativer KI-Anwendungen auf Unternehmensdaten mithilfe von Amazon Kendra und großen Sprachmodellen (AWS Blogbeitrag). LangChain



Für das Foundation-Modell können Sie Amazon Bedrock oder ein über Amazon SageMaker AI bereitgestelltes LLM verwenden. JumpStart Sie können AWS Lambda with verwenden LangChain, um den Fluss zwischen dem Benutzer, Amazon Kendra und dem LLM zu orchestrieren.

Informationen zum Erstellen eines RAG-Systems, das Amazon Kendra und andere verwendet LangChain LLMs, finden Sie im Amazon Kendra LangChain Extensions GitHub Repository.

OpenSearch Amazon-Dienst

Amazon OpenSearch Service bietet integrierte ML-Algorithmen für die Suche nach k-Nearest Neighbours (k-NN), um eine Vektorsuche durchzuführen. OpenSearch Der Service bietet auch eine Vektor-Engine für Amazon EMR Serverless. Mit dieser Vektor-Engine können Sie ein RAG-System erstellen, das über skalierbare und leistungsstarke Vektorspeicher- und Suchfunktionen verfügt. Weitere Informationen zum Erstellen eines RAG-Systems mithilfe von OpenSearch Serverless finden Sie unter Erstellen skalierbarer und serverloser RAG-Workflows mit einer Vektor-Engine für Amazon OpenSearch Serverless- und Amazon Bedrock Claude-Modelle (AWS Blogbeitrag).

Im Folgenden sind die Vorteile der Verwendung von OpenSearch Service für die Vektorsuche aufgeführt:

- Es bietet die vollständige Kontrolle über die Vektordatenbank, einschließlich der Erstellung einer skalierbaren Vektorsuche mithilfe von OpenSearch Serverless.
- Es bietet die Kontrolle über die Chunking-Strategie.
- Es verwendet ANN-Algorithmen (Approximate Nearest Neighbor) aus den Bibliotheken Non-Metric Space Library (NMSLIB), Faiss und Apache Lucene, um eine k-NN-Suche durchzuführen. Sie können den Algorithmus je nach Anwendungsfall ändern. Weitere Informationen zu den Optionen für die Anpassung der Vektorsuche über OpenSearch Service finden Sie unter Erläuterung der Funktionen der Amazon OpenSearch Service-Vektordatenbank (AWS Blogbeitrag).
- OpenSearch Serverless lässt sich als Vektorindex in die Wissensdatenbanken von Amazon Bedrock integrieren.

Amazon Aurora PostgreSQL und pgvector

Amazon Aurora PostgreSQL-Compatible Edition ist eine vollständig verwaltete relationale Datenbank-Engine, die Sie bei der Einrichtung, dem Betrieb und der Skalierung von PostgreSQL-Bereitstellungen unterstützt. pgvector ist eine Open-Source-PostgreSQL-Erweiterung, die Funktionen zur Suche nach Vektorähnlichkeit bietet. Diese Erweiterung ist sowohl für Aurora PostgreSQL-kompatibel als auch für Amazon Relational Database Service (Amazon RDS) für PostgreSQL verfügbar. Weitere Informationen zum Aufbau eines RAG-basierten Systems, das Aurora PostgreSQL-kompatibel und pgvector verwendet, finden Sie in den folgenden Blogbeiträgen: AWS

- Aufbau einer KI-gestützten Suche in PostgreSQL mit Amazon SageMaker AI und pgvector
- Nutzen Sie pgvector und Amazon Aurora PostgreSQL für die Verarbeitung natürlicher Sprache, Chatbots und Stimmungsanalyse

Im Folgenden sind die Vorteile der Verwendung von pgvector und Aurora PostgreSQL-kompatibel aufgeführt:

- Es unterstützt die exakte und ungefähre Suche nach dem nächsten Nachbarn. Es unterstützt auch die folgenden Ähnlichkeitsmetriken: L2-Entfernung, inneres Produkt und Kosinusdistanz.
- Es unterstützt die Indexierung Inverted File with Flat Compression (IVFFlat) und Hierarchical Navigable Small Worlds (HNSW).
- Sie können die Vektorsuche mit Abfragen über domänenspezifische Daten kombinieren, die in derselben PostgreSQL-Instanz verfügbar sind.
- Aurora PostgreSQL-kompatibel ist für mehrstufiges Caching optimiert I/O und bietet dieses. Bei Workloads, die den verfügbaren Instanzspeicher überschreiten, kann pgvector die Abfragen pro Sekunde für die Vektorsuche um das bis zu 8-fache erhöhen.

Amazon Neptune Analytics

Amazon Neptune Analytics ist eine speicheroptimierte Graphdatenbank-Engine für Analysen. Sie unterstützt eine Bibliothek mit optimierten Algorithmen für die Graphanalyse, Grafikabfragen mit niedriger Latenz und Vektorsuchfunktionen innerhalb von Graphendurchläufen. Es verfügt auch über eine integrierte Vektorähnlichkeitssuche. Es bietet einen Endpunkt, um ein Diagramm zu erstellen, Daten zu laden, Abfragen aufzurufen und eine Vektorähnlichkeitssuche durchzuführen. Weitere Informationen zum Erstellen eines RAG-basierten Systems, das Neptune Analytics verwendet, finden Sie unter Verwenden von Wissensgraphen zur Erstellung von GraphRag-Anwendungen mit Amazon Bedrock und Amazon Neptune (Blogbeitrag).AWS

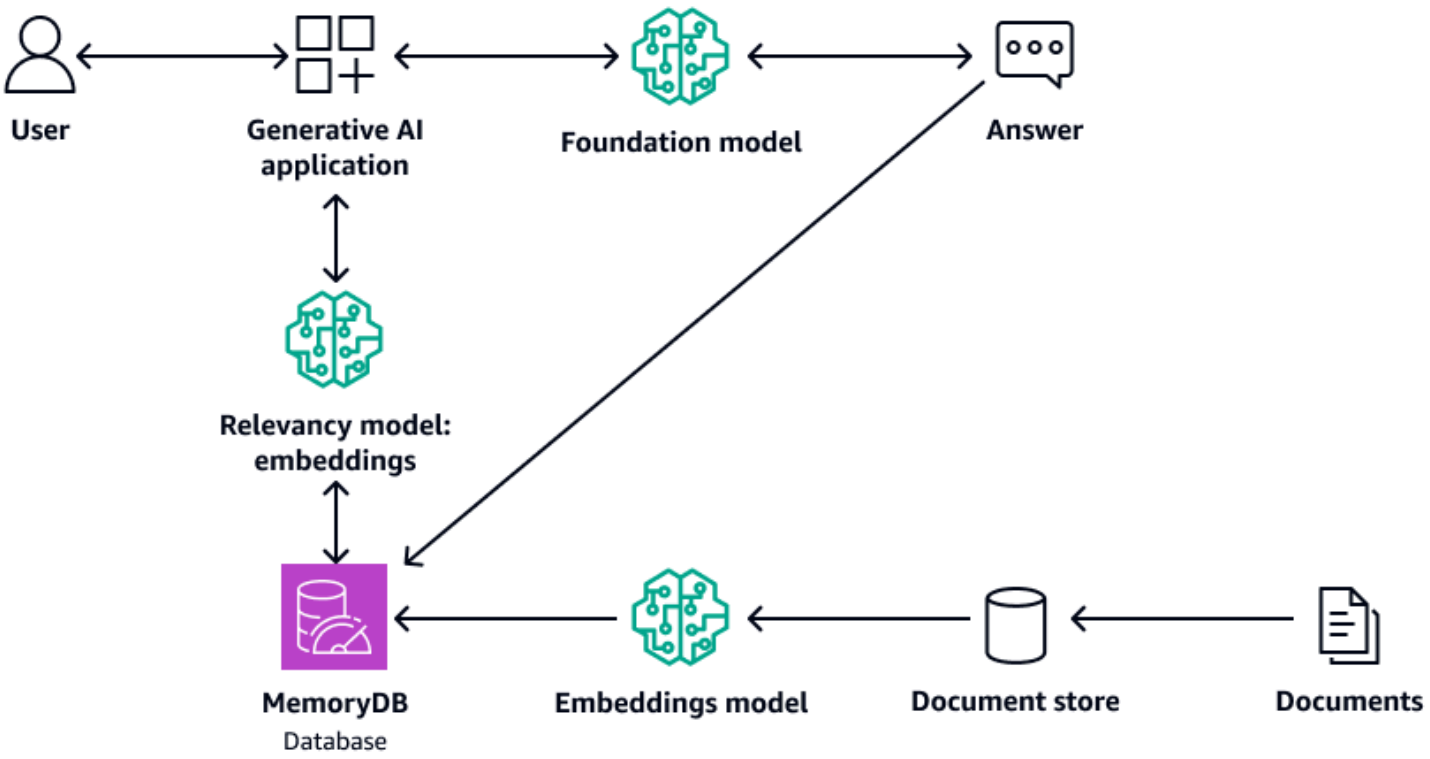
Im Folgenden sind die Vorteile der Verwendung von Neptune Analytics aufgeführt:

- Sie können Einbettungen in Grafikabfragen speichern und durchsuchen.
- Wenn Sie Neptune Analytics mit integrierenLangChain, unterstützt diese Architektur Graphabfragen in natürlicher Sprache.
- Diese Architektur speichert große Graphdatensätze im Speicher.

Amazon MemoryDB

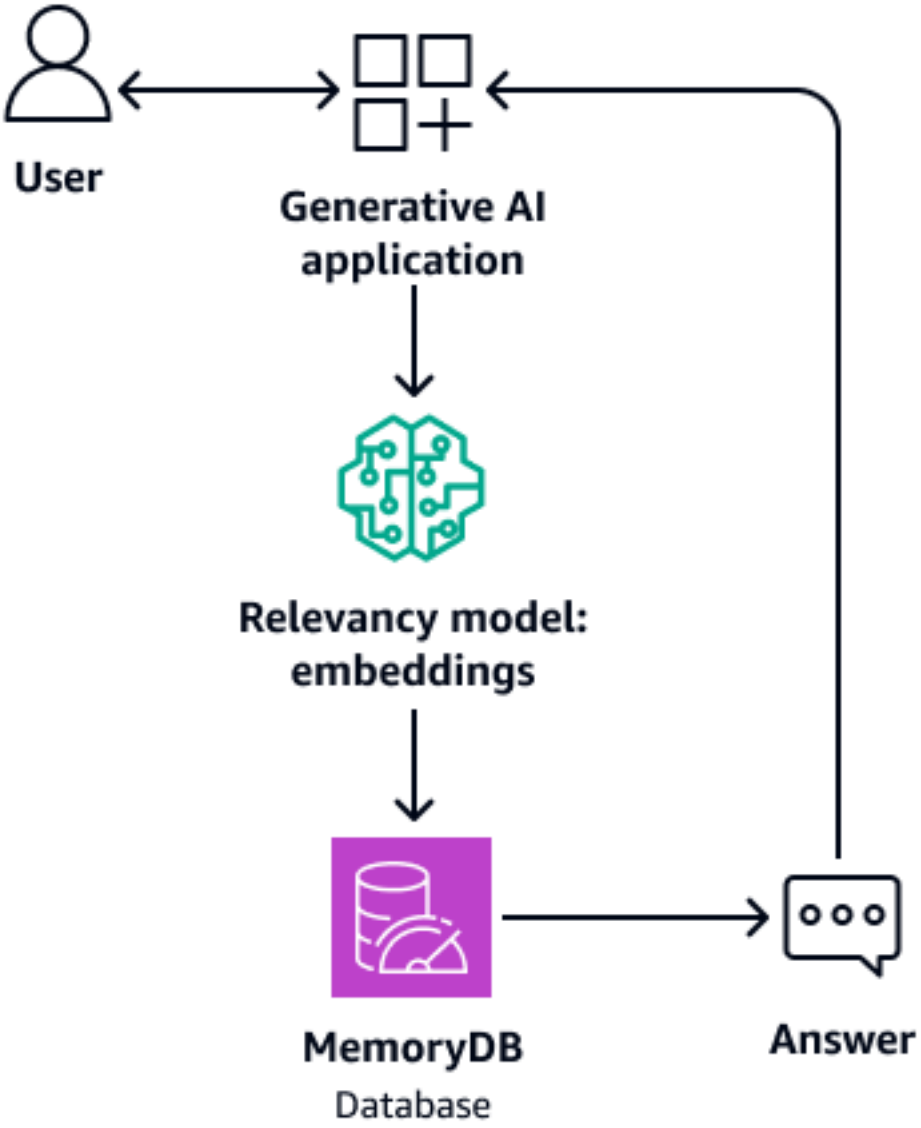
Amazon MemoryDB ist ein langlebiger In-Memory-Datenbankservice, der ultraschnelle Leistung bietet. Alle Ihre Daten werden im Speicher gespeichert, der Lesevorgänge im Mikrosekundenbereich, Schreiblatenz im einstelligen Millisekundenbereich und hohen Durchsatz unterstützt. Die Vektorsuche für MemoryDB erweitert die Funktionalität von MemoryDB und kann in Verbindung mit vorhandenen MemoryDB-Funktionen verwendet werden. Weitere Informationen finden Sie unter Fragen beantworten mit LLM und RAG-Repository auf. GitHub

Das folgende Diagramm zeigt eine Beispielarchitektur, die MemoryDB als Vektordatenbank verwendet.



Im Folgenden sind die Vorteile der Verwendung von MemoryDB aufgeführt:

- Es unterstützt sowohl Flat- als auch HNSW-Indizierungsalgorithmen. Weitere Informationen finden Sie unter Die Vektorsuche für Amazon MemoryDB ist jetzt allgemein im News-Blog verfügbar AWS
- Es kann auch als Pufferspeicher für das Foundation-Modell dienen. Dies bedeutet, dass zuvor beantwortete Fragen aus dem Puffer abgerufen werden, anstatt den Abruf- und Generierungsprozess erneut zu durchlaufen. Das folgende Diagramm zeigt diesen Prozess.



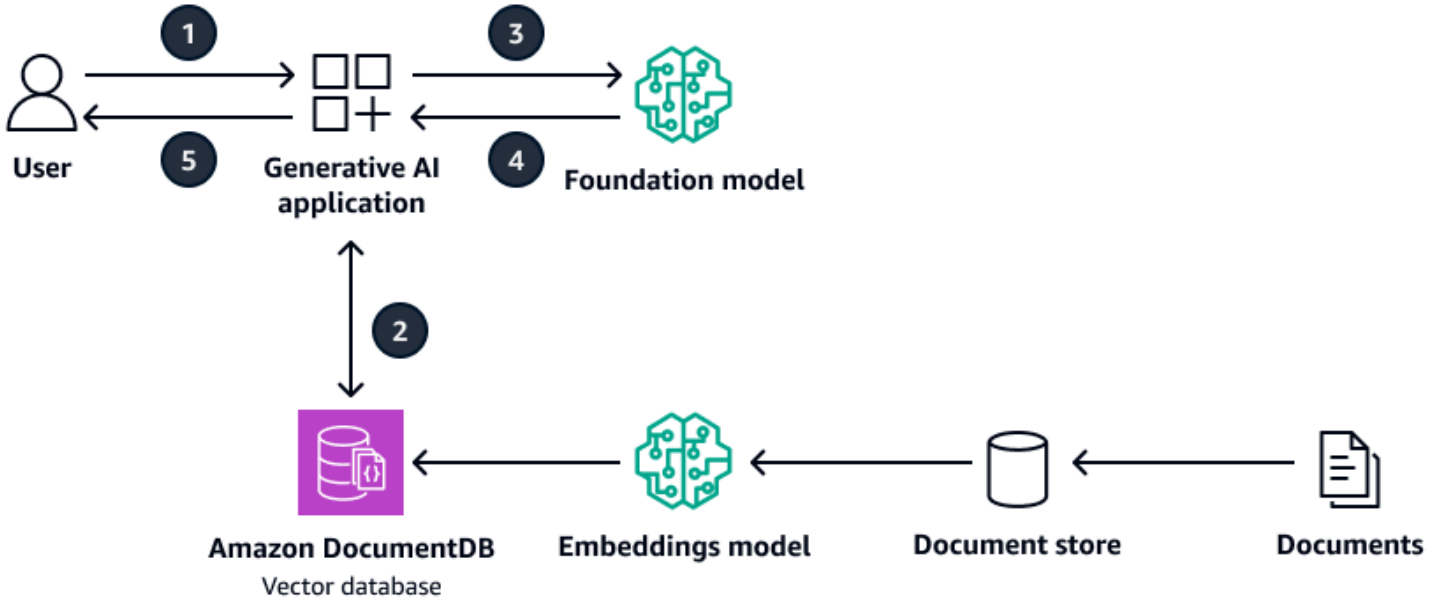
- Da sie eine In-Memory-Datenbank verwendet, bietet diese Architektur eine Abfragezeit im einstelligen Millisekundenbereich für die semantische Suche.
- Sie ermöglicht bis zu 33.000 Abfragen pro Sekunde bei einem Erinnerungsvermögen von 95—99% und 26.500 Abfragen pro Sekunde bei einem Wiedererkennungswert von mehr als 99% Weitere Informationen finden Sie im Video AWS re:Invent 2023 — Vektorsuche mit extrem niedriger Latenz für Amazon MemoryDB auf. YouTube

Amazon DocumentDB

Amazon DocumentDB (mit MongoDB-Kompatibilität) ist ein schneller, zuverlässiger und vollständig verwalteter Datenbankservice. Er macht es einfach, MongoDB kompatible Datenbanken in der Cloud

einzurichten, zu betreiben und zu skalieren. Die Vektorsuche für Amazon DocumentDB kombiniert die Flexibilität und die umfangreichen Abfragefunktionen einer JSON-basierten Dokumentendatenbank mit der Leistungsfähigkeit der Vektorsuche. Weitere Informationen finden Sie unter Fragen beantworten mit dem LLM - und RAG-Repository unter. GitHub

Das folgende Diagramm zeigt eine Beispielarchitektur, die Amazon DocumentDB als Vektordatenbank verwendet.



Das Diagramm zeigt den folgenden Workflow:

1. Der Benutzer sendet eine Anfrage an die generative KI-Anwendung.
2. Die generative KI-Anwendung führt eine Ähnlichkeitssuche in der Amazon DocumentDB DocumentDB-Vektordatenbank durch und ruft die entsprechenden Dokumentauszüge ab.
3. Die generative KI-Anwendung aktualisiert die Benutzerabfrage mit dem abgerufenen Kontext und leitet die Aufforderung an das Ziel-Foundation-Modell weiter.
4. Das Foundation-Modell verwendet den Kontext, um eine Antwort auf die Frage des Benutzers zu generieren, und gibt die Antwort zurück.
5. Die generative KI-Anwendung gibt die Antwort an den Benutzer zurück.

Im Folgenden sind die Vorteile der Verwendung von Amazon DocumentDB aufgeführt:

- Es unterstützt sowohl HNSW- als auch IVFFlat Indexierungsmethoden.

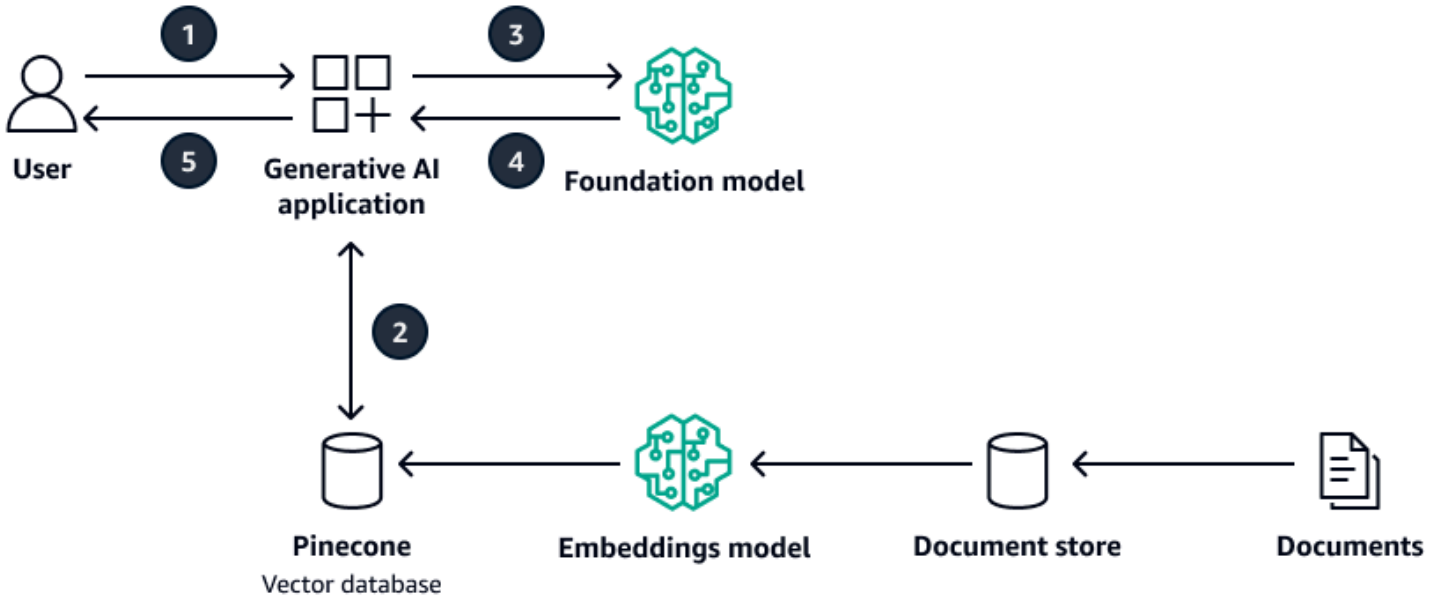
- Es unterstützt bis zu 2.000 Dimensionen in den Vektordaten und unterstützt die Entfernungsmetriken Euklid, Kosinus und Punktprodukt.
- Es bietet Reaktionszeiten im Millisekundenbereich.

Pinecone

Pineconeist eine vollständig verwaltete Vektordatenbank, mit der Sie Produktionsanwendungen um Vektorsuche erweitern können. Sie ist über die verfügbar AWS Marketplace. Die Abrechnung basiert auf der Nutzung. Die Gebühren werden berechnet, indem der Pod-Preis mit der Pod-Anzahl multipliziert wird. Weitere Informationen zum Aufbau eines RAG-basierten Systems, das Folgendes verwendetPinecone, finden Sie in den folgenden AWS Blogbeiträgen:

- Reduzieren Sie Halluzinationen mithilfe von RAG mithilfe der Pinecone Vektordatenbank und Llama-2 von Amazon AI SageMaker JumpStart
- Verwenden Sie Amazon SageMaker AI Studio, um mit Llama 2 eine RAG-Lösung zur Beantwortung von Fragen zu erstellen und schnell Pinecone zu experimentieren LangChain

Das folgende Diagramm zeigt eine Beispielarchitektur, die Pinecone als Vektordatenbank verwendet wird.



Das Diagramm zeigt den folgenden Workflow:

1. Der Benutzer sendet eine Anfrage an die generative KI-Anwendung.
2. Die generative KI-Anwendung führt eine Ähnlichkeitssuche in der Pinecone Vektordatenbank durch und ruft die entsprechenden Dokumentenauszüge ab.
3. Die generative KI-Anwendung aktualisiert die Benutzerabfrage mit dem abgerufenen Kontext und sendet die Aufforderung an das Ziel-Foundation-Modell.
4. Das Foundation-Modell verwendet den Kontext, um eine Antwort auf die Frage des Benutzers zu generieren, und gibt die Antwort zurück.
5. Die generative KI-Anwendung gibt die Antwort an den Benutzer zurück.

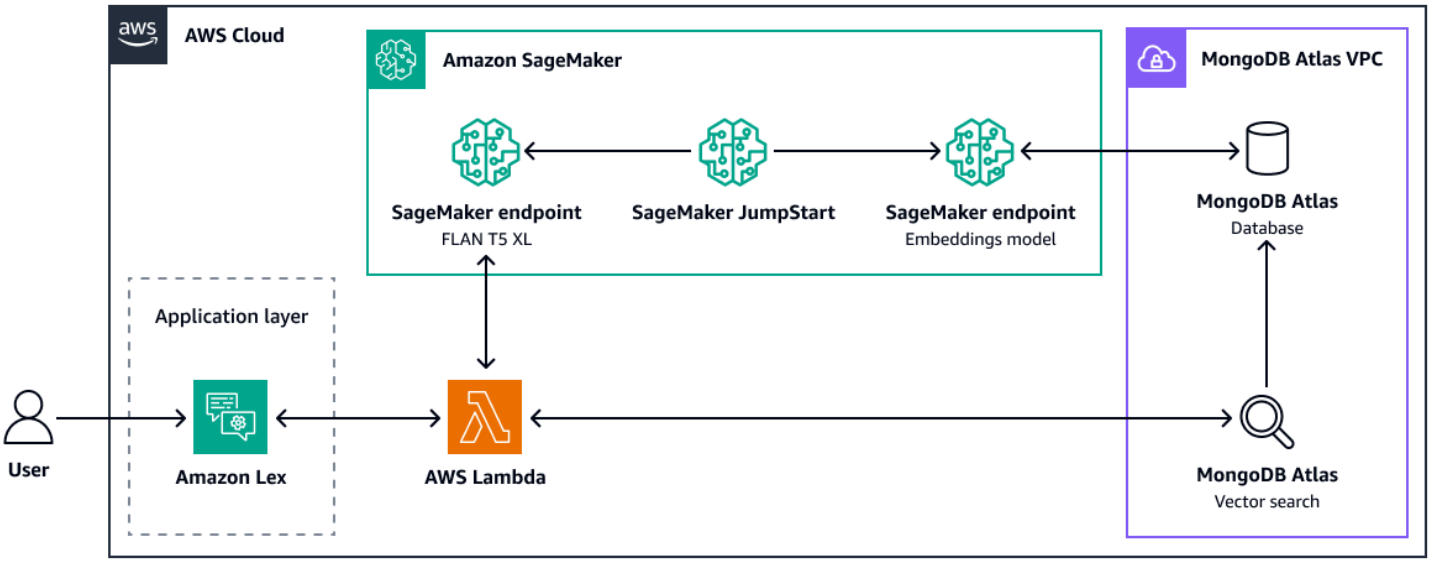
Im Folgenden sind die Vorteile der Verwendung von aufgeführtPinecone:

- Es handelt sich um eine vollständig verwaltete Vektordatenbank, die Ihnen den Aufwand für die Verwaltung Ihrer eigenen Infrastruktur nimmt.
- Sie bietet zusätzliche Funktionen wie Filterung, Live-Indexaktualisierungen und Keyword-Boosting (Hybridsuche).

MongoDB Atlas

MongoDB Atlasist eine vollständig verwaltete Cloud-Datenbank, die die gesamte Komplexität der Bereitstellung und Verwaltung Ihrer Bereitstellungen bewältigt. AWS Sie können Vector Search for verwendenMongoDB Atlas, um Vektor-Einbettungen in Ihrer Datenbank zu speichern. MongoDB Amazon Bedrock Knowledge Bases unterstützt MongoDB Atlas Vektorspeicher. Weitere Informationen finden Sie in der MongoDB Dokumentation unter Erste Schritte mit der Amazon Bedrock Knowledge Base-Integration.

Weitere Informationen zur Verwendung der MongoDB Atlas Vektorsuche für RAG finden Sie unter Retrieval-Augmented Generation withLangChain, Amazon SageMaker AI JumpStart und MongoDB Atlas Semantic Search (AWS Blogbeitrag). Das folgende Diagramm zeigt die Lösungsarchitektur, die in diesem Blogbeitrag detailliert beschrieben wird.



Im Folgenden sind die Vorteile der Verwendung der MongoDB Atlas Vektorsuche aufgeführt:

- Sie können Ihre bestehende Implementierung von verwendenMongoDB Atlas, um Vektoreinbettungen zu speichern und zu durchsuchen.
- Sie können die MongoDBAbfrage-API verwenden, um die Vektoreinbettungen abzufragen.
- Sie können die Vektorsuche und die Datenbank unabhängig voneinander skalieren.
- Vektoreinbettungen werden in der Nähe der Quelldaten (Dokumente) gespeichert, was die Indizierungsleistung verbessert.

Weaviate

Weaviateist eine beliebte Open-Source-Vektordatenbank mit niedriger Latenz, die multimodale Medientypen wie Text und Bilder unterstützt. In der Datenbank werden sowohl Objekte als auch Vektoren gespeichert, wodurch die Vektorsuche mit strukturierter Filterung kombiniert wird. Weitere Informationen zur Verwendung von Weaviate Amazon Bedrock zur Erstellung eines RAG-Workflows finden Sie unter Erstellen unternehmensfähiger generativer KI-Lösungen mit Cohere Foundation-Modellen in Amazon Bedrock und Weaviate Vector Database auf AWS Marketplace (Blogbeitrag).AWS

Im Folgenden sind die Vorteile der Verwendung von: Weaviate

- Es ist Open Source und wird von einer starken Community unterstützt.
- Es ist für die Hybridsuche (sowohl Vektoren als auch Schlüsselwörter) konzipiert.

- Sie können es AWS als verwaltetes Software-as-a-Service (SaaS) -Angebot oder als Kubernetes-Cluster bereitstellen.

Generatoren für RAG-Workflows

Große Sprachmodelle (LLMs) sind sehr große Deep-Learning-Modelle, die für riesige Datenmengen vorab trainiert wurden. Sie sind unglaublich flexibel. LLMs können vielfältige Aufgaben ausführen, z. B. Fragen beantworten, Dokumente zusammenfassen, Sprachen übersetzen und Sätze vervollständigen. Sie haben das Potenzial, die Erstellung von Inhalten und die Art und Weise, wie Menschen Suchmaschinen und virtuelle Assistenten verwenden, zu stören. Sie sind zwar nicht perfekt, LLMs weisen aber eine bemerkenswerte Fähigkeit auf, Vorhersagen auf der Grundlage einer relativ kleinen Aufforderung oder einer relativ geringen Anzahl von Eingaben zu treffen.

LLMs sind eine wichtige Komponente einer RAG-Lösung. Für benutzerdefinierte RAG-Architekturen gibt es zwei AWS-Services , die als Hauptoptionen dienen:

- Amazon Bedrock ist ein vollständig verwalteter Service, der Ihnen LLMs von führenden KI-Unternehmen und Amazon über eine einheitliche API zur Verfügung steht.
- Amazon SageMaker AI JumpStart ist ein ML-Hub, der Basismodelle, integrierte Algorithmen und vorgefertigte ML-Lösungen bietet. Mit SageMaker KI JumpStart können Sie auf vortrainierte Modelle zugreifen, einschließlich Basismodelle. Sie können auch Ihre eigenen Daten verwenden, um die vortrainierten Modelle zu optimieren.

Amazon Bedrock

Amazon Bedrock bietet branchenführende Modelle vonAnthropic,,Stability AI,Meta, Cohere AI21 LabsMistral AI, und Amazon. Eine vollständige Liste finden Sie unter Unterstützte Foundation-Modelle in Amazon Bedrock. Mit Amazon Bedrock können Sie Modelle auch mit Ihren eigenen Daten anpassen.

Sie können die Modellleistung bewerten, um festzustellen, welche Modelle für Ihren RAG-Anwendungsfall am besten geeignet sind. Sie können die neuesten Modelle testen und auch testen, welche Funktionen und Funktionen die besten Ergebnisse liefern und das zum besten Preis. Das Anthropic Claude Sonnet-Modell wird häufig für RAG-Anwendungen verwendet, da es sich bei einer Vielzahl von Aufgaben hervorragend eignet und ein hohes Maß an Zuverlässigkeit und Vorhersagbarkeit bietet.

SageMaker AI JumpStart

SageMaker KI JumpStart bietet vortrainierte Open-Source-Modelle für eine Vielzahl von Problemtypen. Sie können diese Modelle vor der Bereitstellung schrittweise trainieren und optimieren. Sie können auf die vortrainierten Modelle, Lösungsvorlagen und Beispiele über die SageMaker JumpStart KI-Landingpage in Amazon SageMaker AI Studio zugreifen oder das SageMaker AI Python SDK verwenden.

SageMaker KI JumpStart bietet state-of-the-art Basismodelle für Anwendungsfälle wie das Schreiben von Inhalten, Codegenerierung, Beantwortung von Fragen, Verfassen von Texten, Zusammenfassung, Klassifizierung, Informationsabruf und mehr. Verwenden Sie JumpStart Basismodelle, um Ihre eigenen generativen KI-Lösungen zu erstellen und benutzerdefinierte Lösungen mit zusätzlichen SageMaker KI-Funktionen zu integrieren. Weitere Informationen finden Sie unter Erste Schritte mit Amazon SageMaker AI JumpStart.

SageMaker KI integriert JumpStart und verwaltet öffentlich verfügbare Basismodelle, auf die Sie zugreifen, sie anpassen und in Ihre ML-Lebenszyklen integrieren können. Weitere Informationen finden Sie unter Öffentlich verfügbare Basismodelle. SageMaker KI umfasst JumpStart auch proprietäre Basismodelle von Drittanbietern. Weitere Informationen finden Sie unter Proprietäre Basismodelle.

Wählen Sie eine Option zum Abrufen erweiterter Generierung auf AWS

In den Abschnitten Vollständig verwaltete RAG-Optionen und Benutzerdefinierte RAG-Architekturen dieses Handbuchs werden verschiedene Ansätze zum Aufbau einer RAG-basierten Suchlösung beschrieben. AWS In diesem Abschnitt wird beschrieben, wie Sie je nach Anwendungsfall zwischen diesen Optionen wählen können. In einigen Situationen funktioniert möglicherweise mehr als eine Option. In diesem Szenario hängt die Wahl von der Einfachheit der Implementierung, den in Ihrer Organisation verfügbaren Fähigkeiten und den Richtlinien und Standards Ihres Unternehmens ab.

Wir empfehlen Ihnen, die vollständig verwalteten und benutzerdefinierten RAG-Optionen in der folgenden Reihenfolge zu betrachten und die erste Option auszuwählen, die zu Ihrem Anwendungsfall passt:

1. Verwenden Sie Amazon Q Business, es sei denn:
   - Dieser Service ist in Ihrem Land nicht verfügbar AWS-Region, und Ihre Daten können nicht in eine Region verschoben werden, in der er verfügbar ist
   - Sie haben einen bestimmten Grund, den RAG-Workflow anzupassen
   - Sie möchten eine bestehende Vektordatenbank oder ein bestimmtes LLM verwenden
2. Verwenden Sie Wissensdatenbanken für Amazon Bedrock, es sei denn:
   - Sie haben eine Vektordatenbank, die nicht unterstützt wird
   - Sie haben einen bestimmten Grund, den RAG-Workflow anzupassen
3. Kombinieren Sie Amazon Kendra mit einem Generator Ihrer Wahl, es sei denn:
   - Sie möchten Ihre eigene Vektordatenbank wählen
   - Sie möchten die Chunking-Strategie anpassen
4. Wenn Sie mehr Kontrolle über den Retriever haben und Ihre eigene Vektordatenbank auswählen möchten:
   - Wenn Sie nicht über eine bestehende Vektordatenbank verfügen und keine Abfragen mit niedriger Latenz oder Diagrammen benötigen, sollten Sie Amazon OpenSearch Service in Betracht ziehen.
   - Wenn Sie über eine bestehende PostgreSQL Vektordatenbank verfügen, sollten Sie die Amazon Aurora PostgreSQL and-Option in Betracht ziehen. pgvector

- Wenn Sie eine niedrige Latenz benötigen, sollten Sie eine In-Memory-Option wie Amazon MemoryDB oder Amazon DocumentDB in Betracht ziehen.
- Wenn Sie die Vektorsuche mit einer Grafikabfrage kombinieren möchten, sollten Sie Amazon Neptune Analytics in Betracht ziehen.
- Wenn Sie bereits eine Vektordatenbank eines Drittanbieters verwenden oder einen bestimmten Vorteil aus einer solchen Datenbank ziehen, sollten Sie PineconeMongoDB Atlas, und in Betracht ziehen. Weaviate

5. Wenn Sie sich für ein LLM entscheiden möchten:
   - Wenn Sie Amazon Q Business verwenden, können Sie den LLM nicht wählen.
   - Wenn Sie Amazon Bedrock verwenden, können Sie eines der unterstützten Foundation-Modelle wählen.
   - Wenn Sie Amazon Kendra oder eine benutzerdefinierte Vektordatenbank verwenden, können Sie einen der in diesem Handbuch beschriebenen Generatoren oder ein benutzerdefiniertes LLM verwenden.

   > **Note**
   >
   > Sie können Ihre benutzerdefinierten Dokumente auch verwenden, um ein vorhandenes LLM zu optimieren, um die Genauigkeit der Antworten zu erhöhen. Weitere Informationen finden Sie unter Vergleich von RAG und Feinabstimmung in diesem Handbuch.

6. Wenn Sie bereits über eine Implementierung von Amazon SageMaker AI Canvas verfügen, die Sie verwenden möchten, oder wenn Sie RAG-Antworten verschiedener Anbieter vergleichen möchten LLMs, sollten Sie Amazon SageMaker AI Canvas in Betracht ziehen.

Schlussfolgerung

In diesem Handbuch werden die verschiedenen Optionen beschrieben, auf denen ein Retrieval Augmented Generation (RAG) -System aufgebaut werden kann. AWS Sie können mit vollständig verwalteten Services wie Amazon Q Business und Amazon Bedrock Knowledge Bases beginnen. Wenn Sie mehr Kontrolle über den RAG-Workflow haben möchten, können Sie einen benutzerdefinierten Retriever wählen. Für einen Generator können Sie eine API verwenden, um ein unterstütztes LLM in Amazon Bedrock aufzurufen, oder Sie können Ihr eigenes LLM mithilfe von Amazon AI bereitstellen. SageMaker JumpStart Lesen Sie die Empfehlungen unter Auswahl einer RAG-Option, um herauszufinden, welche Option für Ihren Anwendungsfall am besten geeignet ist. Nachdem Sie die beste Option für Ihren Anwendungsfall ausgewählt haben, verwenden Sie die Referenzen in diesem Handbuch, um mit der Erstellung Ihrer RAG-basierten Anwendung zu beginnen.

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	—	28. Oktober 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.