Entwicklung von Retrieval Augmented Generation-Lösungen für das Gesundheitswesen AWS

AWS Präskriptive Leitlinien

AWS Präskriptive Leitlinien: Entwicklung von Retrieval Augmented Generation-Lösungen für das Gesundheitswesen AWS

Table of Contents

Entwicklung von Retrieval Augmented Generation-Lösungen AWS für das Gesundheitswesen

Amazon Web Services, Accenture, und Cadiem (Mitwirkende)

März 2025 (Geschichte des Dokuments)

Vor großen Sprachmodellen (LLMs) und generativer KI war die Entwicklung automatisierter und hochpräziser Anwendungen im Gesundheitswesen eine Herausforderung. Herkömmliche Methoden beruhten stark auf manueller Dateneingabe und -analyse. Die Komplexität der Analyse medizinischer Bildgebung und Patientenakten erforderte umfangreiche menschliche Eingriffe, was häufig zu fragmentierten und ineffizienten Arbeitsabläufen führte. Die Weiterentwicklung der KI-Technologien hilft Ihnen dabei, hyperpersonalisierte Anwendungen in großem Maßstab zu entwickeln. Anwendungen im Gesundheitswesen können jetzt in medizinische Wissensdatenbanken integriert werden, diagnostische Bilder mit höherer Genauigkeit interpretieren und mithilfe von Vorhersagemodellen Behandlungsergebnisse prognostizieren.

In diesem Leitfaden erfahren Sie, wie LLMs Sie das Gesundheitswesen durch Retrieval Augmented Generation-Anwendungen, mit denen Sie bauen können, revolutionieren. AWS-ServicesRetrieval Augmented Generation (RAG) ist eine generative KI-Technologie, bei der ein LLM auf eine verlässliche Datenquelle verweist, die sich außerhalb seiner Trainingsdatenquellen befindet, bevor es eine Antwort generiert. RAG-Anwendungen stützen die Ergebnisse des Modells auf Wissen aus der realen Welt, wodurch Halluzinationen reduziert und die Relevanz der Antworten erhöht werden. Im Gesundheitswesen kann RAG eingesetzt werden, um genaue und up-to-date medizinische Informationen bereitzustellen und so sicherzustellen, dass Gesundheitsdienstleister Zugang zu den neuesten Forschungsergebnissen und klinischen Leitlinien haben. Durch die Umwandlung von Daten in verwertbare Erkenntnisse und die Automatisierung komplexer Prozesse tragen diese Technologien dazu bei, die Patientenversorgung zu verbessern, Abläufe zu rationalisieren und die Produktivität von medizinischem Fachpersonal zu steigern.

In Amazon Bedrock können Sie sie fein abstimmen LLMs und mit intelligenten Agenten integrieren, um fortschrittliche Gesundheitslösungen zu entwickeln. Der Leitfaden hebt die Synergie zwischen Amazon OpenSearch Service und Amazon Neptune hervor und zeigt, wie diese Dienste RAG-Lösungen durch verbesserte Suchrelevanz und erweiterten Datenabruf aus mehreren Quellen verbessern. Sie können umfassende Amazon Bedrock-Lösungen orchestrieren, die Amazon Bedrock-Agenten verwenden und LangChainum Interaktionen zwischen verschiedenen

Datenrepositorien nahtlos zu koordinieren. Diese Integration zeigt, wie leistungsfähig die Kombination spezialisierter Dienste zur Schaffung effektiverer und effizienterer KI-gestützter Systeme ist.

Patientenversorgung und Produktivität

In diesem Leitfaden werden zwei reale Anwendungsfälle für Patientenversorgung und Produktivität vorgestellt: die Erweiterung von Patientendaten und die Vorhersage von Risiken bei einer erneuten Aufnahme. Es enthält strategische Pläne für die Implementierung dieser Lösungen in großem Maßstab und bietet Organisationen im Gesundheitswesen einen klaren Weg zur Industrialisierung KI-gestützter Prozesse. Durch diese Erkenntnisse können Gesundheitseinrichtungen fortschrittliche KI-Technologien nutzen, um effizientere und intelligentere Arbeitsabläufe zu schaffen.

Talentmanagement

Dieser Leitfaden beschreibt auch Strategien zur Umschulung und Befähigung von Mitarbeitern im Gesundheitswesen, generative KI nahtlos in ihren Alltag zu integrieren. Dies kann sowohl die Produktivität als auch die Qualität der Patientenversorgung verbessern. Indem sie ihre Belegschaft mit den Fähigkeiten ausstatten, fortschrittliche KI-Tools effektiv zu nutzen, können Gesundheitsorganisationen ihre Kapitalrendite maximieren und Innovationen in der Patientenversorgung vorantreiben.

Diese KI-gestützte Talentmanagement-Lösung umfasst die folgenden Hauptfunktionen:

- Intelligenter Talent-Lebenslauf-Parser — Mithilfe der in Amazon Bedrock LLMs verfügbaren erweiterten Funktionen extrahiert und analysiert dieses Tool auf effiziente Weise wichtige Fähigkeiten und Eigenschaften von Talenten aus Lebensläufen. Dieses Tool kann den Rekrutierungsprozess rationalisieren.
- Wissensdatenbank für Talente — Diese dynamische Datenbank wird von Amazon Neptune unterstützt und bietet in Echtzeit Einblicke in Personalbestand, Qualifikationsverteilung und Branchentrends. Dies hilft Ihnen, datengestützte Entscheidungen zum Personalmanagement zu treffen.
- Engine für Lernempfehlungen — Dieses KI-gestützte Tool identifiziert Qualifikationslücken innerhalb des Unternehmens und empfiehlt personalisierte Schulungsprogramme für medizinisches Personal. Dieses Tool fördert die kontinuierliche berufliche Weiterentwicklung und hilft Ihrer Belegschaft, sich an die sich weiterentwickelnden Gesundheitstechnologien anzupassen.

Zusammen tragen diese KI-gestützten Funktionen dazu bei, die Leistung der Belegschaft zu optimieren und das Talentmanagement durch mehr Intelligenz und Effizienz zu revolutionieren.

Chancen und Herausforderungen

Amazon Bedrock bietet verbesserte Produktivität, Skalierbarkeit, Kosteneffektivität und datengestützte Einblicke. Amazon Bedrock ermöglicht es Organisationen im Gesundheitswesen, verschiedene Anwendungsfälle LLMs effektiv zu nutzen, von der Erstellung von Inhalten über die Datenanalyse bis hin zur automatisierten Entscheidungsfindung. Dieser Leitfaden bietet Ansätze zur Bewältigung gängiger Herausforderungen im Bereich der generativen KI, wie z. B. Probleme mit der Datenqualität, Skalierbarkeit der Infrastruktur, Aufrechterhaltung der Modellleistung und Anforderungen an die kontinuierliche Verbesserung beim Übergang vom Machbarkeitsnachweis zur Produktion.

Möglichkeiten für generative KI-Anwendungen im Gesundheitswesen

Die Gesundheitsbranche steht vor einem transformativen Wandel, der auf die Möglichkeiten zurückzuführen ist, die generative KI-Anwendungen bieten. Generative KI hat das Potenzial, die Patientenversorgung zu verbessern, Abläufe zu rationalisieren und die medizinische Forschung zu beschleunigen. Durch den Einsatz fortschrittlicher KI-Modelle können Gesundheitsdienstleister die Erweiterung von Patientenakten automatisieren. Umfassende up-to-date Patientenanamnesen ermöglichen genauere Diagnosen und Behandlungspläne. KI-gestützte Bildanalysen, wie z. B. die Interpretation von Sonogrammen und anderen medizinischen Bildgebungsverfahren, können schnelle und präzise Erkenntnisse liefern, wodurch der Arbeitsaufwand für medizinisches Fachpersonal reduziert und das Risiko menschlicher Fehler minimiert wird.

Neben Diagnose und Behandlung kann generative KI eine zentrale Rolle in der prädiktiven Analytik spielen. Prädiktive Analysen helfen Organisationen im Gesundheitswesen dabei, Patientenergebnisse zu antizipieren und Behandlungspläne entsprechend zu personalisieren. Diese Technologie kann auch Verwaltungsprozesse optimieren, von der Verwaltung von Patientendaten bis hin zur Rationalisierung der Kommunikation zwischen Anbietern und Patienten. Durch die Integration generativer KI-Lösungen in bestehende Gesundheitssysteme können medizinische Einrichtungen eine höhere Effizienz erzielen, die Kosten senken und letztendlich eine qualitativ hochwertigere Versorgung bieten. Die Integration von KI in das Gesundheitswesen ist nicht nur eine Verbesserung, sondern auch ein grundlegender Wandel hin zu einer intelligenteren, reaktionsfähigeren und patientenorientierteren Versorgung.

Fortschrittliche Bildanalyse

Die Kombination von Amazon Bedrock mit Datenspeichern wie Amazon Neptune und Amazon OpenSearch Service kann Ihnen helfen, die Komplexität der fortschrittlichen Bildanalyse im Gesundheitswesen zu bewältigen. Lösungen zum Abrufen von Informationen können den Prozess zur Entdeckung von Krankheiten verbessern und die Genauigkeit der Interpretation verbessern, indem sie diagnostische Bilder auswerten und Sonogramme interpretieren. Die Lösung kann die visuellen und textuellen Beurteilungsdaten in die manuelle Überprüfung der Patientenbeurteilung durch Ärzte integrieren.

Herausforderungen bei der Industrialisierung der Lösungen

Die Haupthindernisse, die es bei der Industrialisierung von KI-Lösungen im Gesundheitswesen zu bewältigen gilt, sind Datenqualität und Verfügbarkeit. Gesundheitsdaten liegen häufig in fragmentierten, inkonsistenten Formaten vor. Die Sicherstellung, dass KI-Modelle Zugriff auf saubere, strukturierte und repräsentative Daten haben, ist entscheidend für die Aufrechterhaltung der Leistung in realen Szenarien. Die Skalierbarkeit der Infrastruktur kann aufgrund von Produktionsumgebungen zu einer Herausforderung werden. Diese Umgebungen müssen große Mengen an Patientendaten in Echtzeit verarbeiten und gleichzeitig schnelle Reaktionszeiten bieten und die Einhaltung von Datenschutzbestimmungen wie dem Health Insurance Portability and Accountability Act (HIPAA) gewährleisten. Darüber hinaus müssen KI-Modelle angesichts neuer medizinischer Informationen und Patientendaten, die sich im Laufe der Zeit weiterentwickeln, neu trainiert und aktualisiert werden, damit sie relevant bleiben und genaue Empfehlungen geben. Schließlich kann die Integration dieser KI-Lösungen in bestehende Gesundheitssysteme aufgrund von Interoperabilitätsproblemen und der Notwendigkeit, sie an die aktuellen klinischen Arbeitsabläufe anzupassen, komplex sein. Diese Integration erfordert sowohl technische als auch betriebliche Änderungen.

Anwendungsfall: Entwicklung einer Anwendung für medizinische Intelligenz mit erweiterten Patientendaten

Generative KI kann dazu beitragen, die Patientenversorgung und die Produktivität des Personals zu verbessern, indem sowohl die klinischen als auch die administrativen Funktionen verbessert werden. KI-gestützte Bildanalysen, wie z. B. die Interpretation von Sonogrammen, beschleunigen die Diagnoseprozesse und verbessern die Genauigkeit. Sie kann wichtige Erkenntnisse liefern, die rechtzeitige medizinische Interventionen unterstützen.

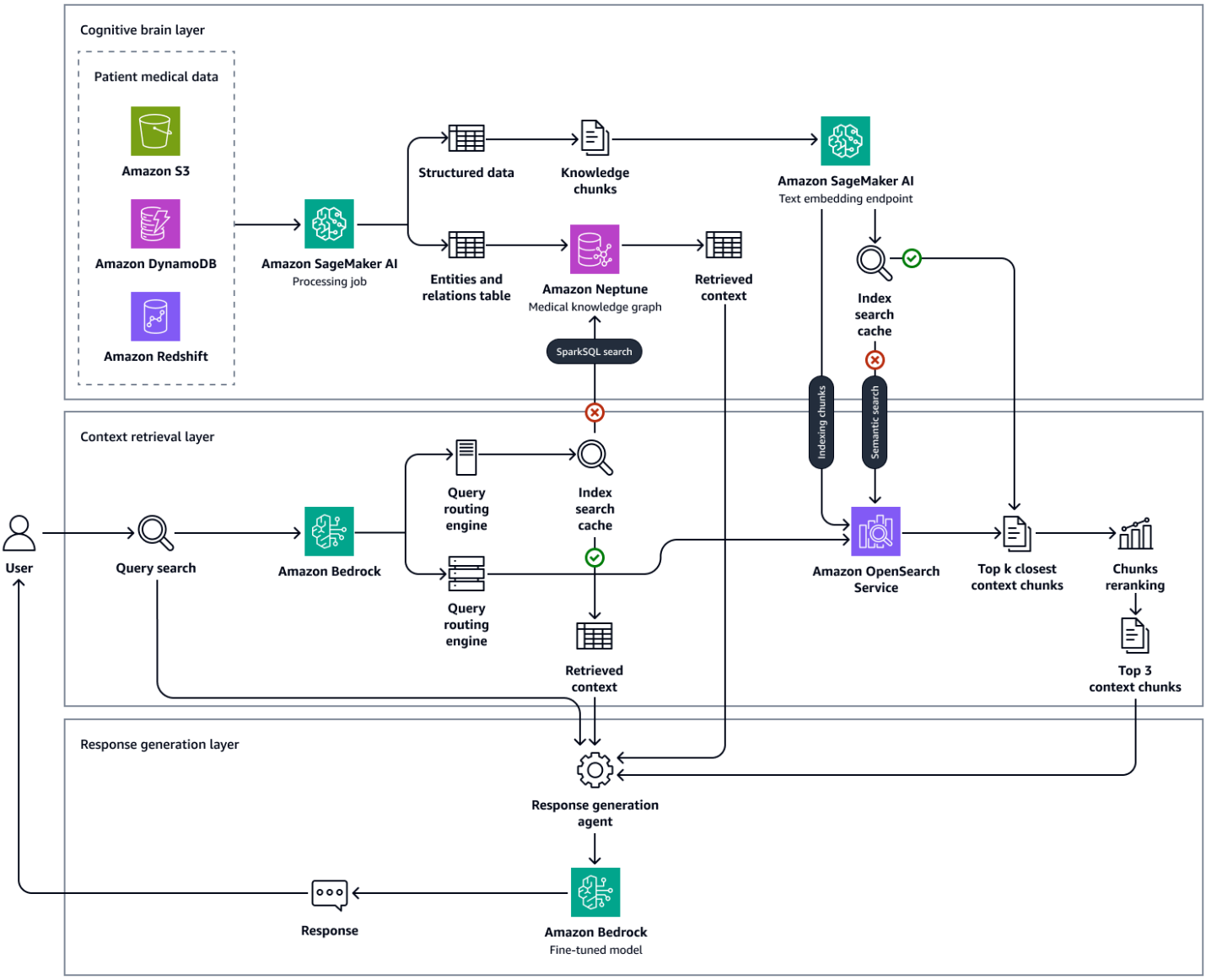
Wenn Sie generative KI-Modelle mit Wissensgraphen kombinieren, können Sie die chronologische Organisation elektronischer Patientenakten automatisieren. Auf diese Weise können Sie Echtzeitdaten aus Interaktionen, Symptomen, Diagnosen, Laborergebnissen und Bildanalysen zwischen Arzt und Patient integrieren. Dadurch erhält der Arzt umfassende Patientendaten. Diese Daten helfen dem Arzt, genauere und zeitnahere medizinische Entscheidungen zu treffen, was sowohl die Behandlungsergebnisse als auch die Produktivität der Gesundheitsdienstleister verbessert.

Übersicht über die Lösung

KI kann Ärzte und Kliniker unterstützen, indem sie Patientendaten und medizinisches Wissen zusammenführt, um wertvolle Erkenntnisse zu gewinnen. Bei dieser Retrieval Augmented Generation (RAG) -Lösung handelt es sich um eine Engine für medizinische Intelligenz, die umfassende Patientendaten und Erkenntnisse aus Millionen von klinischen Interaktionen nutzt. Sie nutzt die Leistungsfähigkeit der generativen KI, um evidenzbasierte Erkenntnisse für eine verbesserte Patientenversorgung zu gewinnen. Es wurde entwickelt, um die klinischen Arbeitsabläufe zu verbessern, Fehler zu reduzieren und die Behandlungsergebnisse zu verbessern.

Die Lösung umfasst eine automatisierte Bildverarbeitungsfunktion, die von unterstützt wird. LLMs Diese Funktion reduziert den Zeitaufwand, den das medizinische Personal für die manuelle Suche nach ähnlichen Diagnosebildern und die Analyse der Diagnoseergebnisse aufwenden muss.

Die folgende Abbildung zeigt die Lösung end-to-end-workflow für diese Lösung. Es verwendet Amazon Neptune, Amazon SageMaker AI, Amazon OpenSearch Service und ein Basismodell in Amazon Bedrock. Für den Context Retrieval Agent, der mit dem Medical Knowledge Graph in Neptune interagiert, können Sie zwischen einem Amazon Bedrock-Agenten und einem LangChain Agent.



In unseren Experimenten mit medizinischen Musterfragen stellten wir fest, dass die endgültigen Antworten, die durch unseren Ansatz mithilfe eines in Neptune verwalteten Wissensgraphen, einer OpenSearch Vektordatenbank mit klinischer Wissensdatenbank und Amazon Bedrock generiert LLMs wurden, auf Fakten beruhten und weitaus genauer sind, da die falsch positiven Ergebnisse reduziert und die wahren positiven Ergebnisse verstärkt werden. Diese Lösung kann evidenzbasierte Erkenntnisse über den Gesundheitszustand der Patienten liefern und zielt darauf ab, die klinischen Arbeitsabläufe zu verbessern, Fehler zu reduzieren und die Behandlungsergebnisse zu verbessern.

Der Aufbau dieser Lösung besteht aus den folgenden Schritten:

- Schritt 1: Daten ermitteln

- Schritt 2: Erstellung eines medizinischen Wissensgraphen
- Schritt 3: Erstellung von Context-Retrieval-Agenten zur Abfrage des medizinischen Wissensgraphen
- Schritt 4: Erstellen einer Wissensdatenbank mit beschreibenden Echtzeitdaten
- Schritt 5: Verwendung LLMs zur Beantwortung medizinischer Fragen

Schritt 1: Daten ermitteln

Es gibt viele medizinische Open-Source-Datensätze, die Sie verwenden können, um die Entwicklung einer KI-gestützten Lösung für das Gesundheitswesen zu unterstützen. Ein solcher Datensatz ist der MIMIC-IV-Datensatz, ein öffentlich zugänglicher Datensatz für elektronische Patientenakten (EHR), der in der Gesundheitsforschung weit verbreitet ist. MIMIC-IV enthält detaillierte klinische Informationen, einschließlich Freitext-Entlassungsnotizen aus Patientenakten. Sie können diese Aufzeichnungen verwenden, um mit Techniken zur Textsummierung und Extraktion von Entitäten zu experimentieren. Diese Techniken helfen Ihnen dabei, medizinische Informationen (wie Patientensymptome, verabreichte Medikamente und verschriebene Behandlungen) aus unstrukturiertem Text zu extrahieren.

Sie können auch einen Datensatz verwenden, der kommentierte, anonymisierte Zusammenfassungen von Patientenentlassungen enthält, die speziell für Forschungszwecke zusammengestellt wurden. Ein Datensatz mit einer Zusammenfassung der Entlassung kann Ihnen beim Experimentieren mit der Extraktion von Entitäten helfen, sodass Sie wichtige medizinische Entitäten (wie Erkrankungen, Verfahren und Medikamente) anhand des Textes identifizieren können. Schritt 2: Erstellung eines medizinischen WissensgraphenIn diesem Leitfaden wird beschrieben, wie Sie die strukturierten Daten aus den Datensätzen MIMIC-IV und Zusammenfassung der Entlassungsdaten verwenden können, um ein medizinisches Wissensdiagramm zu erstellen. Dieser Graph zum medizinischen Wissen dient als Grundlage für fortschrittliche Abfrage- und Entscheidungsunterstützungssysteme für medizinisches Fachpersonal.

Zusätzlich zu textbasierten Datensätzen können Sie Bilddatensätze verwenden. Zum Beispiel der Datensatz Musculoskeletal Radiographs (MURA), bei dem es sich um eine umfassende Datenbank mit Röntgenbildern von Knochen mit mehreren Ansichten handelt. Verwenden Sie solche Bilddatensätze, um mit der diagnostischen Beurteilung mithilfe medizinischer Bilddekodierungstechniken zu experimentieren. Diese Entschlüsselungstechniken sind entscheidend für die Früherkennung von Krankheiten wie Erkrankungen des Bewegungsapparates, Herz-Kreislauf-Erkrankungen und Osteoporose. Durch die Feinabstimmung von Seh- und Sprachmodellen anhand

des medizinischen Bilddatensatzes können Sie Auffälligkeiten in diagnostischen Bildern erkennen. Auf diese Weise kann das System Ärzten frühzeitige und genaue diagnostische Erkenntnisse liefern. Durch die Verwendung von Bild- und Textdatensätzen können Sie eine KI-gestützte Gesundheitsanwendung erstellen, die sowohl Text- als auch Bilddaten verarbeiten kann, um die Patientenversorgung zu verbessern.

Schritt 2: Erstellung eines medizinischen Wissensgraphen

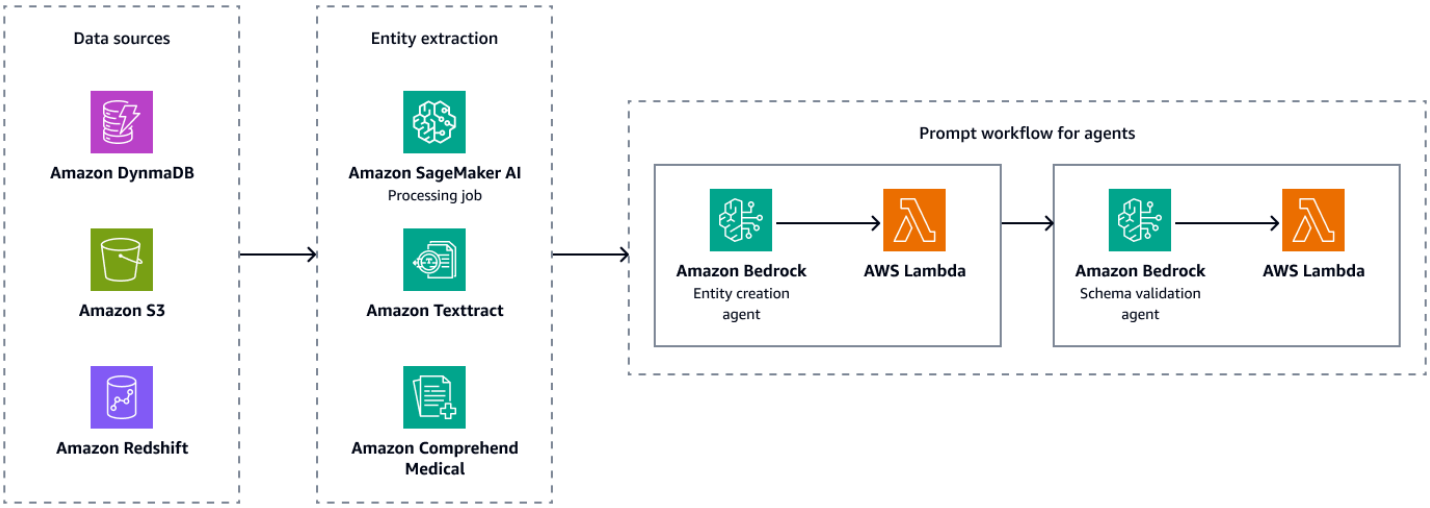
Für jede Gesundheitsorganisation, die ein System zur Entscheidungsunterstützung aufbauen möchte, das auf einer riesigen Wissensbasis basiert, besteht eine zentrale Herausforderung darin, die medizinischen Entitäten zu finden und zu extrahieren, die in den klinischen Aufzeichnungen, medizinischen Fachzeitschriften, Entlassungszusammenfassungen und anderen Datenquellen enthalten sind. Sie müssen auch die zeitlichen Zusammenhänge, Themen und Sicherheitseinschätzungen aus diesen Krankenakten erfassen, um die extrahierten Entitäten, Attribute und Beziehungen effektiv nutzen zu können.

Der erste Schritt besteht darin, medizinische Konzepte aus dem unstrukturierten medizinischen Text zu extrahieren, indem ein paar Eingabeaufforderungen für ein Basismodell verwendet werden, wie Llama 3 in Amazon Bedrock. Beim Few-Shot-Prompting stellen Sie einem LLM eine kleine Anzahl von Beispielen zur Verfügung, die die Aufgabe und das gewünschte Ergebnis demonstrieren, bevor Sie es bitten, eine ähnliche Aufgabe auszuführen. Mithilfe eines LLM-basierten Extraktors für medizinische Entitäten können Sie den unstrukturierten medizinischen Text analysieren und anschließend eine strukturierte Datendarstellung der medizinischen Wissenseinheiten generieren. Sie können die Patientenattribute auch für nachgelagerte Analysen und Automatisierung speichern. Der Vorgang zur Extraktion von Entitäten umfasst die folgenden Aktionen:

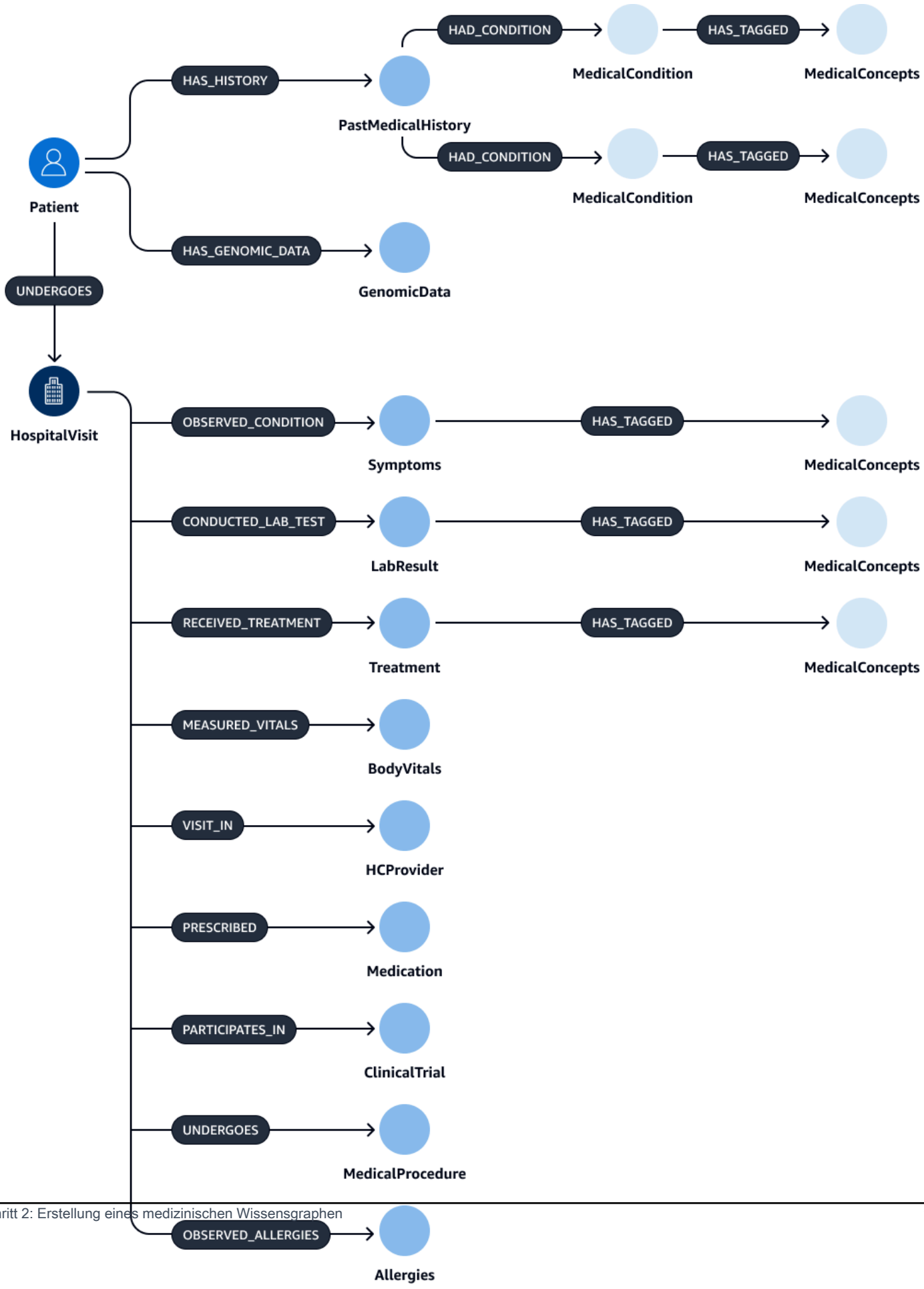
- Extrahieren Sie Informationen über medizinische Konzepte wie Krankheiten, Medikamente, Medizinprodukte, Dosierung, Häufigkeit und Dauer der Behandlung, Symptome, medizinische Eingriffe und deren klinisch relevante Eigenschaften.
- Erfassen Sie funktionale Merkmale, wie z. B. zeitliche Beziehungen zwischen extrahierten Entitäten, Subjekten und Sicherheitsbewertungen.
- Erweitern Sie medizinische Standardvokabeln, wie z. B. die folgenden:
  - Concept Identifiers (RxCUI) aus der Datenbank RxNorm
  - Codes aus der Internationalen Klassifikation der Krankheiten, 10. Revision, klinische Modifikation (ICD-10-CM)
  - Begriffe aus medizinischen Fachüberschriften (MeSH)

  - Konzepte aus der Systematisierten Nomenklatur der Medizin, klinische Begriffe (SNOMED CT)
  - Codes aus dem Unified Medical Language System (UMLS)
- Fassen Sie Entlassungsbescheinigungen zusammen und leiten Sie medizinische Erkenntnisse aus Zeugnissen ab.

Die folgende Abbildung zeigt die Schritte zur Entitätsextraktion und Schemavalidierung, um gültige Kombinationen von Entitäten, Attributen und Beziehungen zu erstellen. Sie können unstrukturierte Daten wie Entlassungszusammenfassungen oder Patientennotizen in Amazon Simple Storage Service (Amazon S3) speichern. Sie können strukturierte Daten wie ERP-Daten (Enterprise Resource Planning), elektronische Patientenakten und Laborinformationssysteme in Amazon Redshift und Amazon DynamoDB speichern. Sie können einen Amazon Bedrock Entity Creation Agent erstellen. Dieser Agent kann Dienste wie Amazon SageMaker AI-Datenextraktionspipelines, Amazon Textract und Amazon Comprehend Medical integrieren, um Entitäten, Beziehungen und Attribute aus den strukturierten und unstrukturierten Datenquellen zu extrahieren. Schließlich verwenden Sie einen Amazon Bedrock-Schemavalidierungsagenten, um sicherzustellen, dass die extrahierten Entitäten und Beziehungen dem vordefinierten Diagrammschema entsprechen, und um die Integrität der Knoten-Edge-Verbindungen und der zugehörigen Eigenschaften aufrechtzuerhalten.



Nach der Extraktion und Validierung der Entitäten, Beziehungen und Attribute können Sie sie verknüpfen, um ein subject-object-predicate Triplett zu erstellen. Sie nehmen diese Daten in eine Amazon Neptune Neptune-Graphdatenbank auf, wie in der folgenden Abbildung dargestellt. Graphdatenbanken sind für das Speichern und Abfragen der Beziehungen zwischen Datenelementen optimiert.


HAD_CONDITION
HAS_TAGGED
MedicalCondition
MedicalConcepts
HAS_HISTORY
PastMedicalHistory
HAD_CONDITION
HAS_TAGGED
MedicalCondition
MedicalConcepts
Patient
HAS_GENOMIC_DATA
GenomicData
UNDERGOES
HospitalVisit
OBSERVED_CONDITION
HAS_TAGGED
Symptoms
MedicalConcepts
CONDUCTED_LAB_TEST
HAS_TAGGED
LabResult
MedicalConcepts
RECEIVED_TREATMENT
HAS_TAGGED
Treatment
MedicalConcepts
MEASURED_VITALS
BodyVitals
VISIT_IN
HCProvider
PRESCRIBED
Medication
PARTICIPATES_IN
ClinicalTrial
UNDERGOES
MedicalProcedure
OBSERVED_ALLERGIES
Allergies

Sie könnten mit diesen Daten einen umfassenden Wissensgraphen erstellen. Ein Wissensgraph hilft Ihnen dabei, alle Arten von zusammenhängenden Informationen zu organisieren und abzufragen. Sie könnten beispielsweise einen Wissensgraphen mit den folgenden Hauptknoten erstellen:`HospitalVisit`,`PastMedicalHistory`, `SymptomsMedication`,`MedicalProcedures`, und`Treatment`.

In den folgenden Tabellen sind die Entitäten und ihre Attribute aufgeführt, die Sie aus Entlassungsnotizen extrahieren können.

Entität	Attribute
`Patient`	`PatientID` , `Name`, `Age`, `Gender`, `Address`, `ContactInformation`
`HospitalVisit`	`VisitDate` , `Reason`, `Notes`
`HealthcareProvider`	`ProviderID` , `Name`, `Specialty` , `ContactInformation` , `Address`, `AffiliatedInstitution`
`Symptoms`	`Description` , `RiskFactors`
`Allergies`	`AllergyType` , `Duration`
`Medication`	`MedicationID` , `Name`, `Description` , `Dosage`, `SideEffects` , `Manufacturer`
`PastMedicalHistory`	`ContinuingMedicines`
`MedicalCondition`	`ConditionName` , `Severity`, `Treatment Received` , `DoctorinCharge` , `HospitalName` , `MedicinesFollowed`
`BodyVitals`	`HeartRate` , `BloodPressure` , `RespiratoryRate` , `BodyTemperature` , `BMI`
`LabResult`	`LabResultID` , `PatientID` , `TestName`, `Result`, `Date`

Entität	Attribute
`ClinicalTrial`	`TrialID`, `Name`, `Description` , `Phase`, `Status`, `StartDate` , `EndDate`
`GenomicData`	`GenomicDataID` , `PatientID` , `SequenceData` , `VariantInformation`
`Treatment`	`TreatmentID` , `Name`, `Description` , `Type`, `SideEffects`
`MedicalProcedure`	`ProcedureID` , `Name`, `Description` , `Risks`, `Outcomes`
`MedicalConcepts`	`UMLSCodes` , `MedicalVocabularies`

In der folgenden Tabelle sind die Beziehungen aufgeführt, die möglicherweise zwischen Entitäten bestehen, und die entsprechenden Attribute. Beispielsweise könnte die `Patient` Entität eine Verbindung zu der `HospitalVisit` Entität mit der `[UNDERGOES]` Beziehung herstellen. Das Attribut für diese Beziehung ist`VisitDate`.

Betreff-Entität	Beziehung	Objekt-Entität	Attribute
`Patient`	`[UNDERGOES]`	`HospitalVisit`	`VisitDate`
`HospitalVisit`	`[VISIT_IN]`	`HealthcareProvider`	`ProviderName` , `Location`, `ProviderID` , `VisitDate`
`HospitalVisit`	`[OBSERVED_CONDITION]`	`Symptoms`	`Severity`, `CurrentStatus` , `VisitDate`
`HospitalVisit`	`[RECEIVED_TREATMENT]`	`Treatment`	`Duration`, `Dosage`, `VisitDate`

Betreff-Entität	Beziehung	Objekt-Entität	Attribute
HospitalVisit	[PRESCRIBED]	Medication	Duration, Dosage, Adherence , VisitDate
Patient	[HAS_HISTORY]	PastMedicalHistory	Keine
PastMedicalHistory	[HAD_CONDITION]	MedicalCondition	DiagnosisDate , CurrentStatus
HospitalVisit	[PARTICIPATES_IN]	ClinicalTrial	VisitDate , Status, Outcomes
Patient	[HAS_GENOMIC_DATA]	GenomicData	CollectionDate
HospitalVisit	[OBSERVED_ALLERGIES]	Allergies	VisitDate
HospitalVisit	[CONDUCTED_LAB_TEST]	LabResult	VisitDate , AnalysisDate , Interpretation
HospitalVisit	[UNDERGOES]	MedicalProcedure	VisitDate , Outcome
MedicalCondition	[HAS_TAGGED]	MedicalConcepts	Keine
LabResult	[HAS_TAGGED]	MedicalConcepts	Keine
Treatment	[HAS_TAGGED]	MedicalConcepts	Keine
Symptoms	[HAS_TAGGED]	MedicalConcepts	Keine

Schritt 3: Erstellung von Context-Retrieval-Agenten zur Abfrage des medizinischen Wissensgraphen

Nachdem Sie die medizinische Graphdatenbank erstellt haben, besteht der nächste Schritt darin, Agenten für die Graphinteraktion zu erstellen. Die Agenten rufen den richtigen und erforderlichen Kontext für die Abfrage ab, die ein Arzt oder Kliniker eingibt. Es gibt mehrere Optionen für die Konfiguration dieser Agenten, die den Kontext aus dem Knowledge Graph abrufen:

- Agenten von Amazon Bedrock
- LangChain Agenten

Amazon Bedrock-Agenten für die Interaktion mit Diagrammen

Amazon Bedrock-Agenten arbeiten nahtlos mit Amazon Neptune Neptune-Graphdatenbanken zusammen. Sie können erweiterte Interaktionen über Amazon Bedrock-Aktionsgruppen durchführen. Die Aktionsgruppe initiiert den Prozess, indem sie eine AWS Lambda Funktion aufruft, die Neptune OpenCypher-Abfragen ausführt.

Für die Abfrage eines Wissensgraphen können Sie zwei unterschiedliche Ansätze verwenden: direkte Abfrageausführung oder Abfragen mit Kontexteinbettung. Diese Ansätze können unabhängig voneinander oder kombiniert angewendet werden, abhängig von Ihrem spezifischen Anwendungsfall und Ihren Rangkriterien. Durch die Kombination beider Ansätze können Sie dem LLM einen umfassenderen Kontext bieten, was die Ergebnisse verbessern kann. Im Folgenden sind die beiden Ansätze zur Abfrageausführung aufgeführt:

- Direkte Ausführung von Cypher-Abfragen ohne Einbettungen — Die Lambda-Funktion führt Abfragen direkt gegen Neptune aus, ohne dass eine auf Einbettungen basierende Suche erforderlich ist. Im Folgenden finden Sie ein Beispiel für diesen Ansatz:

```
MATCH (p:Patient)-[u:UNDERGOES]->(h:HospitalVisit) WHERE h.Reason = 'Acute Diabetes'
 AND date(u.VisitDate) > date('2024-01-01')
RETURN p.PatientID, p.Name, p.Age, p.Gender, p.Address, p.ContactInformation
```

- Direkte Ausführung von Cypher-Abfragen mithilfe der eingebetteten Suche — Die Lambda-Funktion verwendet die eingebettete Suche, um die Abfrageergebnisse zu verbessern. Dieser Ansatz verbessert die Abfrageausführung durch die Integration von Einbettungen, bei denen es sich um dichte Vektordarstellungen von Daten handelt. Einbettungen sind besonders nützlich,

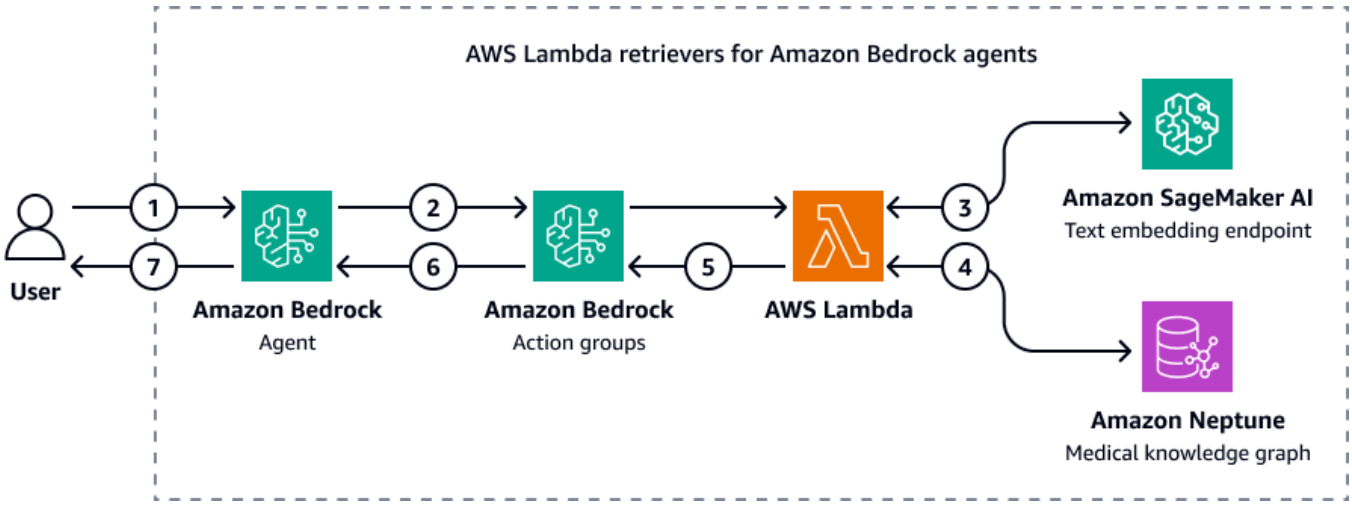
wenn für die Abfrage semantische Ähnlichkeit oder ein umfassenderes Verständnis erforderlich ist, das über exakte Übereinstimmungen hinausgeht. Sie können vortrainierte oder individuell trainierte Modelle verwenden, um Einbettungen für jede Erkrankung zu generieren. Im Folgenden finden Sie ein Beispiel für diesen Ansatz:

```
CALL { WITH "Acute Diabetes" AS query_term RETURN search_embedding(query_term) AS
 similar_reasons }

MATCH (p:Patient)-[u:UNDERGOES]->(h:HospitalVisit) WHERE h.Reason IN similar reasons
 AND date(u.VisitDate) > date('2024-01-01')
RETURN p.PatientID, p.Name, p.Age, p.Gender, p.Address, p.ContactInformation
```

In diesem Beispiel ruft die `search_embedding("Acute Diabetes")` Funktion Erkrankungen ab, die dem Begriff „Akuter Diabetes“ semantisch nahe kommen. Auf diese Weise kann die Abfrage auch nach Patienten mit Erkrankungen wie Prädiabetes oder metabolischem Syndrom suchen.

Die folgende Abbildung zeigt, wie Amazon Bedrock-Agenten mit Amazon Neptune interagieren, um eine Cypher-Abfrage eines medizinischen Wissensgraphen durchzuführen.



Das Diagramm zeigt den folgenden Workflow:

1. Der Benutzer sendet eine Frage an den Amazon Bedrock-Agenten.
2. Der Amazon Bedrock-Agent leitet die Frage und die Eingabefiltervariablen an die Amazon Bedrock-Aktionsgruppen weiter. Diese Aktionsgruppen enthalten eine AWS Lambda Funktion, die mit dem Amazon SageMaker AI-Endpunkt zur Texteinbettung und dem Amazon Neptune Medical Knowledge Graph interagiert.

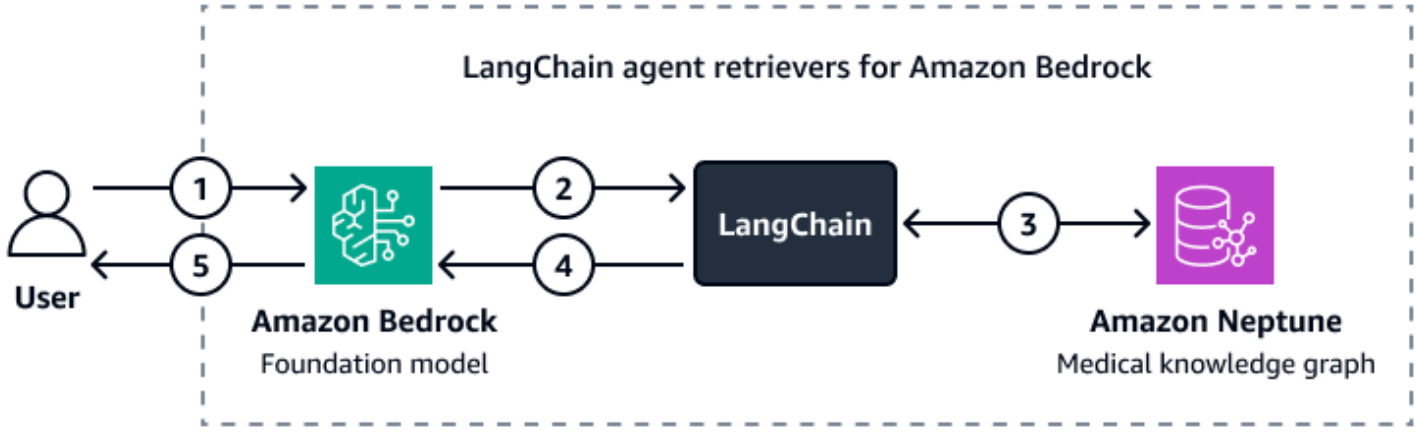
3. Die Lambda-Funktion ist in den SageMaker AI-Texteinbettungsendpunkt integriert, um eine semantische Suche innerhalb der OpenCypher-Abfrage durchzuführen. Sie konvertiert die Abfrage in natürlicher Sprache in eine OpenCypher-Abfrage, indem sie die zugrunde liegende LangChain Agenten.
4. Die Lambda-Funktion fragt den Neptune Medical Knowledge Graph nach dem richtigen Datensatz ab und empfängt die Ausgabe aus dem Neptune Medical Knowledge Graph.
5. Die Lambda-Funktion gibt die Ergebnisse von Neptune an die Amazon Bedrock-Aktionsgruppen zurück.
6. Die Amazon Bedrock-Aktionsgruppen senden den abgerufenen Kontext an den Amazon Bedrock-Agenten.
7. Der Amazon Bedrock-Agent generiert die Antwort anhand der ursprünglichen Benutzerabfrage und des abgerufenen Kontextes aus dem Knowledge Graph.

LangChain Agenten für die Interaktion mit Diagrammen

Sie können integrieren LangChain mit Neptune, um graphbasierte Abfragen und Abrufe zu ermöglichen. Dieser Ansatz kann KI-gesteuerte Workflows verbessern, indem er die Graphdatenbankfunktionen von Neptune nutzt. Der Brauch LangChain Der Retriever fungiert als Vermittler. Das grundlegende Modell in Amazon Bedrock kann mit Neptune interagieren, indem es sowohl direkte Cypher-Abfragen als auch komplexere Graphalgorithmen verwendet.

Sie können den benutzerdefinierten Retriever verwenden, um zu verfeinern, wie LangChain Der Agent interagiert mit den Neptun-Graph-Algorithmen. Sie können beispielsweise Few-Shot-Prompting verwenden, um die Antworten des Basismodells auf der Grundlage bestimmter Muster oder Beispiele anzupassen. Sie können auch LLM-identifizierte Filter anwenden, um den Kontext zu verfeinern und die Genauigkeit der Antworten zu verbessern. Dies kann die Effizienz und Genauigkeit des gesamten Abrufprozesses bei der Interaktion mit komplexen Grafikdaten verbessern.

Die folgende Abbildung zeigt, wie ein benutzerdefinierter LangChain Der Agent orchestriert die Interaktion zwischen einem Amazon Bedrock Foundation-Modell und einem Amazon Neptune Medical Knowledge Graph.



Das Diagramm zeigt den folgenden Workflow:

1. Ein Benutzer sendet eine Frage an Amazon Bedrock und LangChain Agent.
2. Das Amazon Bedrock Foundation-Modell verwendet das Neptun-Schema, das bereitgestellt wird von LangChain Agent, um eine Abfrage für die Frage des Benutzers zu generieren.
3. Das Tool LangChain Der Agent führt die Abfrage anhand des Amazon Neptune Medical Knowledge Graph aus.
4. Das Tool LangChain Der Agent sendet den abgerufenen Kontext an das Amazon Bedrock Foundation-Modell.
5. Das Amazon Bedrock Foundation-Modell verwendet den abgerufenen Kontext, um eine Antwort auf die Frage des Benutzers zu generieren.

Schritt 4: Erstellen einer Wissensdatenbank mit beschreibenden Echtzeitdaten

Als Nächstes erstellen Sie eine Wissensdatenbank mit beschreibenden Notizen zur Interaktion zwischen Arzt und Patient in Echtzeit, diagnostischen Bildbeurteilungen und Laboranalyseberichten. Bei dieser Wissensdatenbank handelt es sich um eine Vektordatenbank. Durch die Verwendung einer Vektordatenbank, in der beschreibendes medizinisches Wissen in indexierter, vektorisierter Form gespeichert werden kann, können Gesundheitsdienstleister relevante Informationen aus einem riesigen Datenbestand effizient abfragen und darauf zugreifen. Diese vektorisierten Darstellungen helfen Ihnen, semantisch ähnliche Daten abzurufen. Leistungserbringer können schnell durch klinische Notizen, medizinische Bilder und Laborergebnisse navigieren. Dies beschleunigt die fundierte Entscheidungsfindung, da der sofortige Zugriff auf kontextrelevante

Informationen ermöglicht wird, wodurch die Genauigkeit und Geschwindigkeit von Diagnosen und Behandlungsplänen verbessert wird.

Nutzung einer medizinischen Wissensdatenbank von OpenSearch Service

Amazon OpenSearch Service kann große Mengen hochdimensionaler medizinischer Daten verwalten. Es handelt sich um einen verwalteten Service, der eine leistungsstarke Suche und Echtzeitanalysen ermöglicht. Es eignet sich gut als Vektordatenbank für RAG-Anwendungen. OpenSearch Der Service dient als Backend-Tool zur Verwaltung großer Mengen unstrukturierter oder halbstrukturierter Daten wie Krankenakten, Forschungsartikeln und klinischen Notizen. Seine fortschrittlichen semantischen Suchfunktionen helfen Ihnen dabei, kontextrelevante Informationen abzurufen. Dies macht es besonders nützlich in Anwendungen wie Systemen zur Unterstützung klinischer Entscheidungen, Tools zur Lösung von Patientenanfragen und Wissensmanagementsystemen im Gesundheitswesen. So kann ein Arzt beispielsweise schnell relevante Patientendaten oder Forschungsstudien finden, die bestimmten Symptomen oder Behandlungsprotokollen entsprechen. Dies hilft Ärzten dabei, Entscheidungen zu treffen, die sich auf die relevantesten up-to-date und relevantesten Informationen stützen.

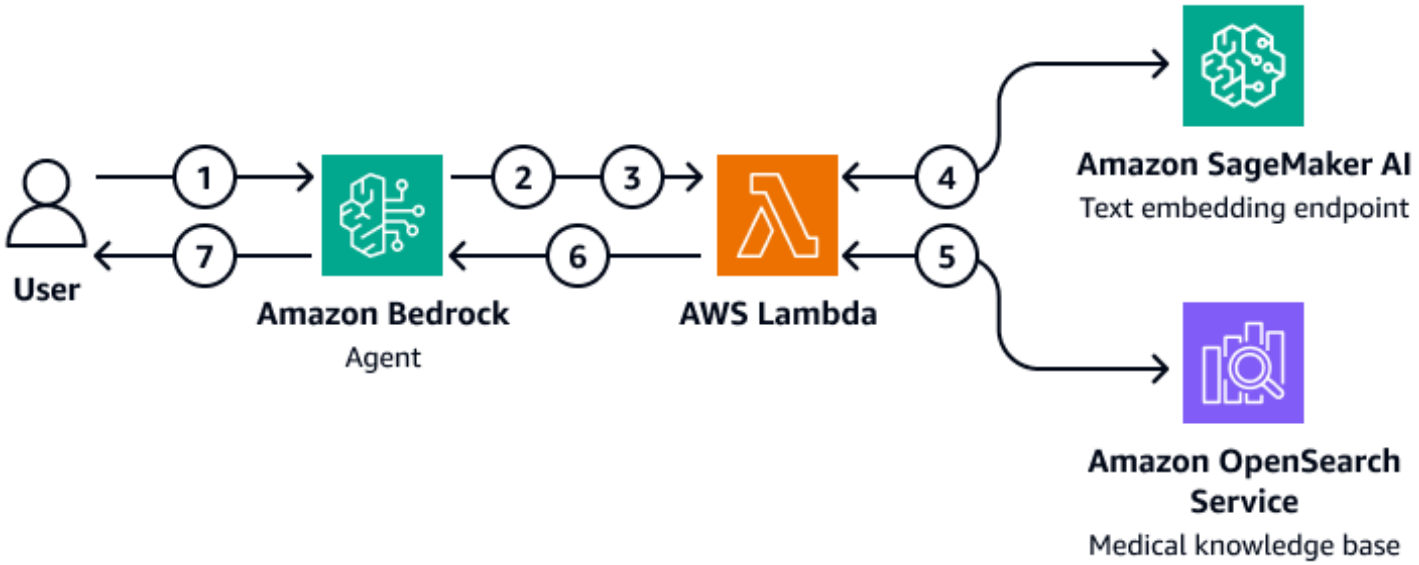
OpenSearch Der Service kann skaliert und die Indizierung und Abfrage von Daten in Echtzeit abgewickelt werden. Dies macht es ideal für dynamische Umgebungen im Gesundheitswesen, in denen der zeitnahe Zugriff auf genaue Informationen entscheidend ist. Darüber hinaus verfügt es über multimodale Suchfunktionen, die sich optimal für Suchen eignen, die mehrere Eingaben erfordern, z. B. medizinische Bilder und Arztnotizen. Bei der Implementierung von OpenSearch Service for Healthcare-Anwendungen ist es wichtig, dass Sie präzise Felder und Zuordnungen definieren, um die Indexierung und den Datenabruf zu optimieren. Felder stellen die einzelnen Daten dar, z. B. Patientenakten, Krankengeschichten und Diagnosecodes. Zuordnungen definieren, wie diese Felder gespeichert (in eingebetteter Form oder Originalform) und abgefragt werden. Für Anwendungen im Gesundheitswesen ist es wichtig, Zuordnungen zu erstellen, die verschiedene Datentypen berücksichtigen, darunter strukturierte Daten (wie numerische Testergebnisse), halbstrukturierte Daten (wie Patientennotizen) und unstrukturierte Daten (wie medizinische Bilder)

In OpenSearch Service können Sie mithilfe kuratierter Eingabeaufforderungen neuronale Volltext-Suchanfragen durchführen, um Krankenakten, klinische Notizen oder Forschungsarbeiten zu durchsuchen, um schnell relevante Informationen zu bestimmten Symptomen, Behandlungen oder Patientenanamnesen zu finden. Neuronale Suchanfragen übernehmen mithilfe integrierter neuronaler Netzwerkmodelle automatisch die Einbettung der Eingabeaufforderung und der Bilder. Dies hilft dabei, die tieferen semantischen Beziehungen in multimodalen Daten zu verstehen und zu erfassen,

und bietet im Vergleich zu anderen Suchabfragealgorithmen, wie der Suche nach k-Nearest Neighbor (k-NN), kontextsensitivere und präzisere Suchergebnisse.

Erstellung einer RAG-Architektur

Sie können eine maßgeschneiderte RAG-Lösung bereitstellen, die Amazon Bedrock-Agenten verwendet, um eine medizinische Wissensdatenbank in OpenSearch Service abzufragen. Um dies zu erreichen, erstellen Sie eine AWS Lambda Funktion, die mit OpenSearch Service interagieren und Anfragen abfragen kann. Die Lambda-Funktion bettet die Eingabefrage des Benutzers ein, indem sie auf einen SageMaker AI-Texteinbettungsendpunkt zugreift. Der Amazon Bedrock-Agent übergibt zusätzliche Abfrageparameter als Eingaben an die Lambda-Funktion. Die Funktion fragt die medizinische Wissensdatenbank in OpenSearch Service ab, die den relevanten medizinischen Inhalt zurückgibt. Nachdem Sie die Lambda-Funktion eingerichtet haben, fügen Sie sie als Aktionsgruppe innerhalb des Amazon Bedrock-Agenten hinzu. Der Amazon Bedrock-Agent nimmt die Eingaben des Benutzers entgegen, identifiziert die erforderlichen Variablen, übergibt die Variablen und die Frage an die Lambda-Funktion und initiiert dann die Funktion. Die Funktion gibt einen Kontext zurück, der dem Foundation-Modell hilft, eine genauere Antwort auf die Frage des Benutzers zu geben.

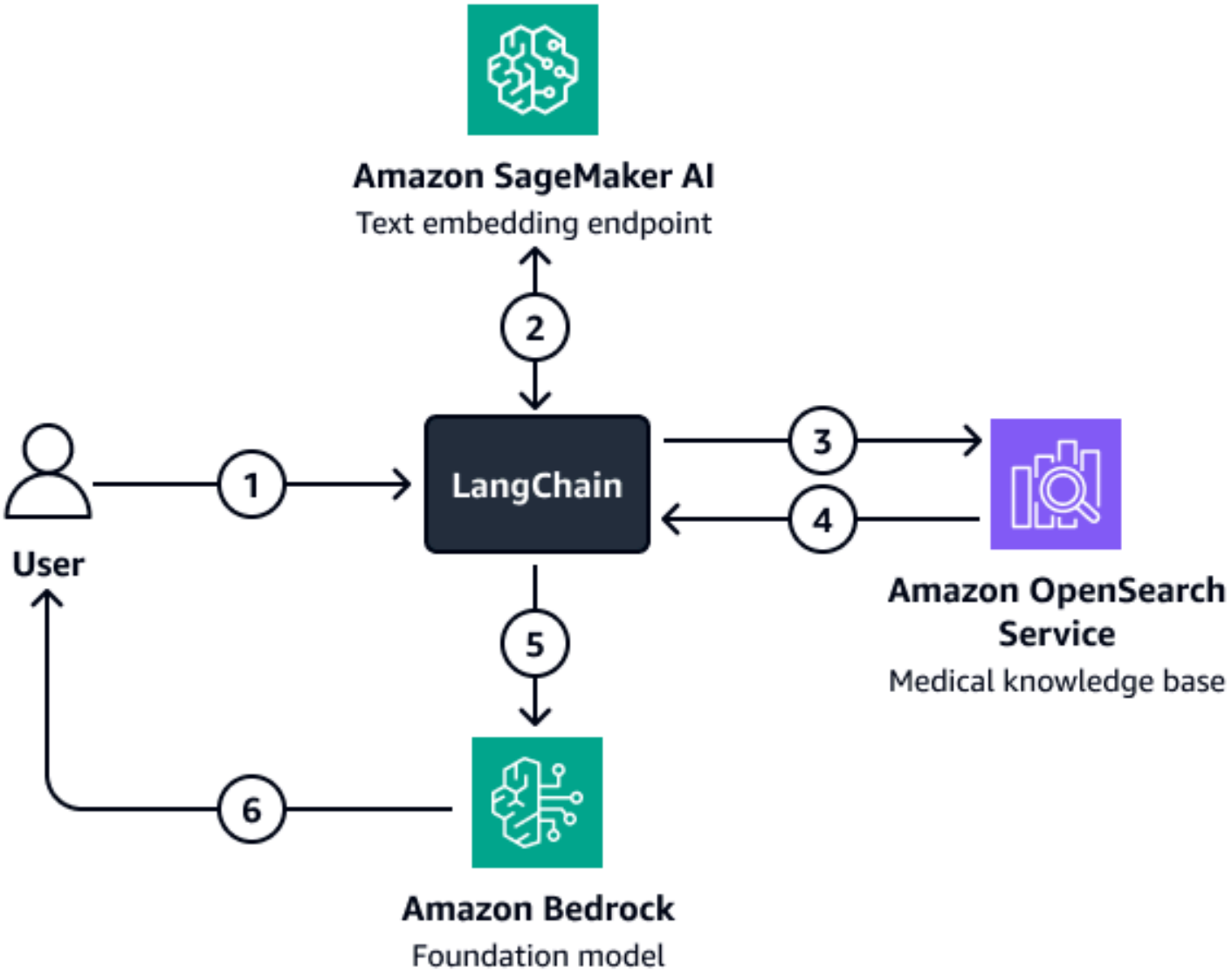


Das Diagramm zeigt den folgenden Workflow:

1. Ein Benutzer sendet eine Frage an den Amazon Bedrock-Agenten.
2. Der Amazon Bedrock-Agent wählt aus, welche Aktionsgruppe initiiert werden soll.
3. Der Amazon Bedrock-Agent initiiert eine AWS Lambda Funktion und übergibt ihr Parameter.

4. Die Lambda-Funktion initiiert das Amazon SageMaker AI-Texteinbettungsmodell, um die Benutzerfrage einzubetten.
5. Die Lambda-Funktion übergibt den eingebetteten Text und zusätzliche Parameter und Filter an Amazon OpenSearch Service. Amazon OpenSearch Service fragt die medizinische Wissensdatenbank ab und gibt die Ergebnisse an die Lambda-Funktion zurück.
6. Die Lambda-Funktion gibt die Ergebnisse an den Amazon Bedrock-Agenten zurück.
7. Das Basismodell im Amazon Bedrock-Agenten generiert eine Antwort auf der Grundlage der Ergebnisse und gibt die Antwort an den Benutzer zurück.

Für Situationen, in denen eine komplexere Filterung erforderlich ist, können Sie eine benutzerdefinierte Methode verwenden LangChain Retriever. Erstellen Sie diesen Retriever, indem Sie einen OpenSearch Service Vector Search Client einrichten, der direkt in LangChain. Diese Architektur ermöglicht es Ihnen, mehr Variablen zu übergeben, um die Filterparameter zu erstellen. Nachdem der Retriever eingerichtet ist, verwenden Sie das Amazon Bedrock-Modell und den Retriever, um eine Fragen-Antwort-Kette für den Abruf einzurichten. Diese Kette orchestriert die Interaktion zwischen dem Modell und dem Retriever, indem sie die Benutzereingaben und mögliche Filter an den Retriever weitergibt. Der Retriever gibt den relevanten Kontext zurück, der dem Foundation-Modell hilft, die Frage des Benutzers zu beantworten.



Das Diagramm zeigt den folgenden Workflow:

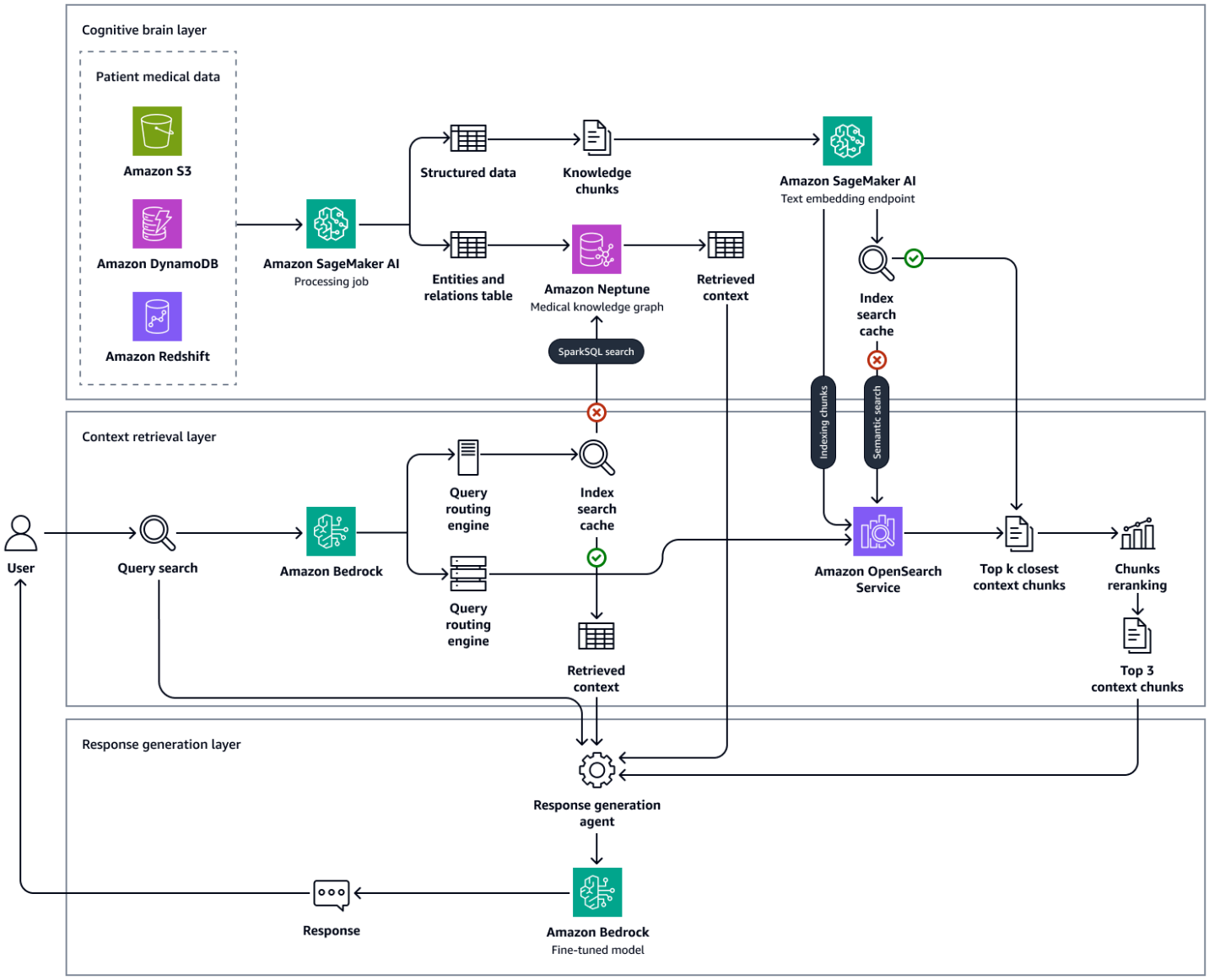
1. Ein Benutzer sendet eine Frage an den LangChain Retriever-Agent.
2. Das Tool LangChain Der Retriever-Agent sendet die Frage an den Amazon SageMaker AI-Texteinbettungsendpunkt, um die Frage einzubetten.
3. Das Tool LangChain Der Retriever-Agent leitet den eingebetteten Text an Amazon OpenSearch Service weiter.
4. Amazon OpenSearch Service sendet die abgerufenen Dokumente zurück an LangChain Retriever-Agent.
5. Das Tool LangChain Der Retriever-Agent leitet die Benutzerfrage und den abgerufenen Kontext an das Amazon Bedrock Foundation-Modell weiter.

6. Das Foundation-Modell generiert eine Antwort und sendet sie an den Benutzer.

Schritt 5: Verwendung LLMs zur Beantwortung medizinischer Fragen

Die vorherigen Schritte helfen Ihnen dabei, eine Anwendung für medizinische Intelligenz zu erstellen, mit der die Krankenakten eines Patienten abgerufen und relevante Medikamente und mögliche Diagnosen zusammengefasst werden können. Jetzt erstellen Sie die Generationsebene. Diese Ebene nutzt die generativen Funktionen eines LLM in Amazon Bedrock, wie Llama 3, um die Ausgabe der Anwendung zu verbessern.

Wenn ein Arzt eine Abfrage eingibt, führt die Kontext-Abruf-Ebene der Anwendung den Abrufvorgang aus dem Knowledge Graph durch und gibt die wichtigsten Datensätze zurück, die sich auf die Krankengeschichte, Demografie, Symptome, Diagnose und Behandlungsergebnisse des Patienten beziehen. Aus der Vektordatenbank werden außerdem in Echtzeit beschreibende Hinweise zur Interaktion zwischen Arzt und Patient, Erkenntnisse aus der diagnostischen Bildbeurteilung, Zusammenfassungen von Laboranalyseberichten und Erkenntnisse aus einer Vielzahl von medizinischen Forschungsarbeiten und wissenschaftlichen Büchern abgerufen. Diese am häufigsten abgerufenen Ergebnisse, die Anfrage des Klinikers und die Eingabeaufforderungen (die darauf zugeschnitten sind, Antworten auf der Grundlage der Art der Anfrage zu kuratieren) werden dann an das Foundation-Modell in Amazon Bedrock übergeben. Dies ist die Ebene zur Generierung von Antworten. Das LLM verwendet den abgerufenen Kontext, um eine Antwort auf die Anfrage des Klinikers zu generieren. Die folgende Abbildung zeigt den end-to-end Arbeitsablauf der Schritte in dieser Lösung.



Sie können ein vortrainiertes Basismodell in Amazon Bedrock, wie Llama 3, für eine Reihe von Anwendungsfällen verwenden, die die Anwendung für medizinische Intelligenz bewältigen muss. Das effektivste LLM für eine bestimmte Aufgabe hängt vom Anwendungsfall ab. Beispielsweise könnte ein vorab trainiertes Modell ausreichen, um Gespräche zwischen Patienten und Ärzten zusammenzufassen, Medikamente und Patientengeschichten zu durchsuchen und Erkenntnisse aus internen medizinischen Datensätzen und wissenschaftlichen Erkenntnissen abzurufen. Für andere komplexe Anwendungsfälle wie Laboruntersuchungen in Echtzeit, Empfehlungen für medizinische Verfahren und Prognosen von Behandlungsergebnissen könnte jedoch ein fein abgestimmtes LLM erforderlich sein. Sie können ein LLM verfeinern, indem Sie es anhand von Datensätzen aus dem medizinischen Bereich trainieren. Spezifische oder komplexe Anforderungen im Gesundheitswesen und in den Biowissenschaften treiben die Entwicklung dieser fein abgestimmten Modelle voran.

Weitere Informationen zur Feinabstimmung eines LLM oder zur Auswahl eines bestehenden LLM, das auf medizinischen Daten trainiert wurde, finden Sie unter Verwendung umfangreicher Sprachmodelle für Anwendungsfälle im Gesundheitswesen und in den Biowissenschaften.

Ausrichtung auf das AWS Well-Architected Framework

Die Lösung entspricht allen sechs Säulen des AWS Well-Architected Framework wie folgt:

- Operative Exzellenz — Die Architektur ist entkoppelt, um eine effiziente Überwachung und Aktualisierung zu gewährleisten. Amazon Bedrock-Agenten AWS Lambda helfen Ihnen dabei, Tools schnell bereitzustellen und rückgängig zu machen.
- Sicherheit — Diese Lösung wurde entwickelt, um Gesundheitsvorschriften wie HIPAA zu erfüllen. Sie können auch Verschlüsselung, detaillierte Zugriffskontrolle und Amazon Bedrock Guardrails implementieren, um Patientendaten zu schützen.
- Zuverlässigkeit — AWS Managed Services wie Amazon OpenSearch Service und Amazon Bedrock bieten die Infrastruktur für eine kontinuierliche Modellinteraktion.
- Leistungseffizienz — Die RAG-Lösung ruft mithilfe optimierter semantischer Suchen und Cypher-Abfragen schnell relevante Daten ab, während ein Agenten-Router optimale Routen für Benutzeranfragen identifiziert.
- Kostenoptimierung — Das pay-per-token Modell in der Amazon Bedrock- und RAG-Architektur reduziert die Kosten für Inferenzen und Vorschulungen.
- Nachhaltigkeit — Durch den Einsatz von serverloser Infrastruktur und pay-per-token Rechenleistung wird der Ressourcenverbrauch minimiert und die Nachhaltigkeit verbessert.

Anwendungsfall: Prognose von Behandlungsergebnissen und Wiederaufnahmequoten

KI-gestützte prädiktive Analysen bieten weitere Vorteile, da sie Patientenergebnisse prognostizieren und personalisierte Behandlungspläne ermöglichen. Dies kann die Patientenzufriedenheit und die Gesundheitsergebnisse verbessern. Durch die Integration dieser KI-Funktionen mit Amazon Bedrock und anderen Technologien können Gesundheitsdienstleister erhebliche Produktivitätssteigerungen erzielen, Kosten senken und die Gesamtqualität der Patientenversorgung verbessern.

Sie können medizinische Daten wie Krankengeschichten, klinische Notizen, Medikamente und Behandlungen in einem Wissensdiagramm speichern. Durch die Kombination des tiefen kontextuellen Verständnisses von LLMs mit den strukturierten, zeitlichen Daten in einem medizinischen Wissensgraphen können Gesundheitsdienstleister zusätzliche Einblicke in individuelle Patientenmuster gewinnen. Mithilfe von prädiktiven Analysen können Sie mögliche Fehleinhaltung oder Behandlungskomplikationen frühzeitig erkennen und personalisierte Werte für die Wahrscheinlichkeit einer erneuten Aufnahme erstellen.

Mit dieser Lösung können Sie die Wahrscheinlichkeit einer erneuten Zulassung vorhersagen. Diese Prognosen können die Behandlungsergebnisse verbessern und die Gesundheitskosten senken. Diese Lösung kann auch Krankenhausärzten und -verwaltern helfen, sich auf Patienten zu konzentrieren, bei denen ein höheres Risiko einer erneuten Aufnahme besteht. Sie hilft ihnen auch dabei, proaktive Interventionen für diese Patienten durch Warnmeldungen, Self-Service und datengestützte Maßnahmen einzuleiten.

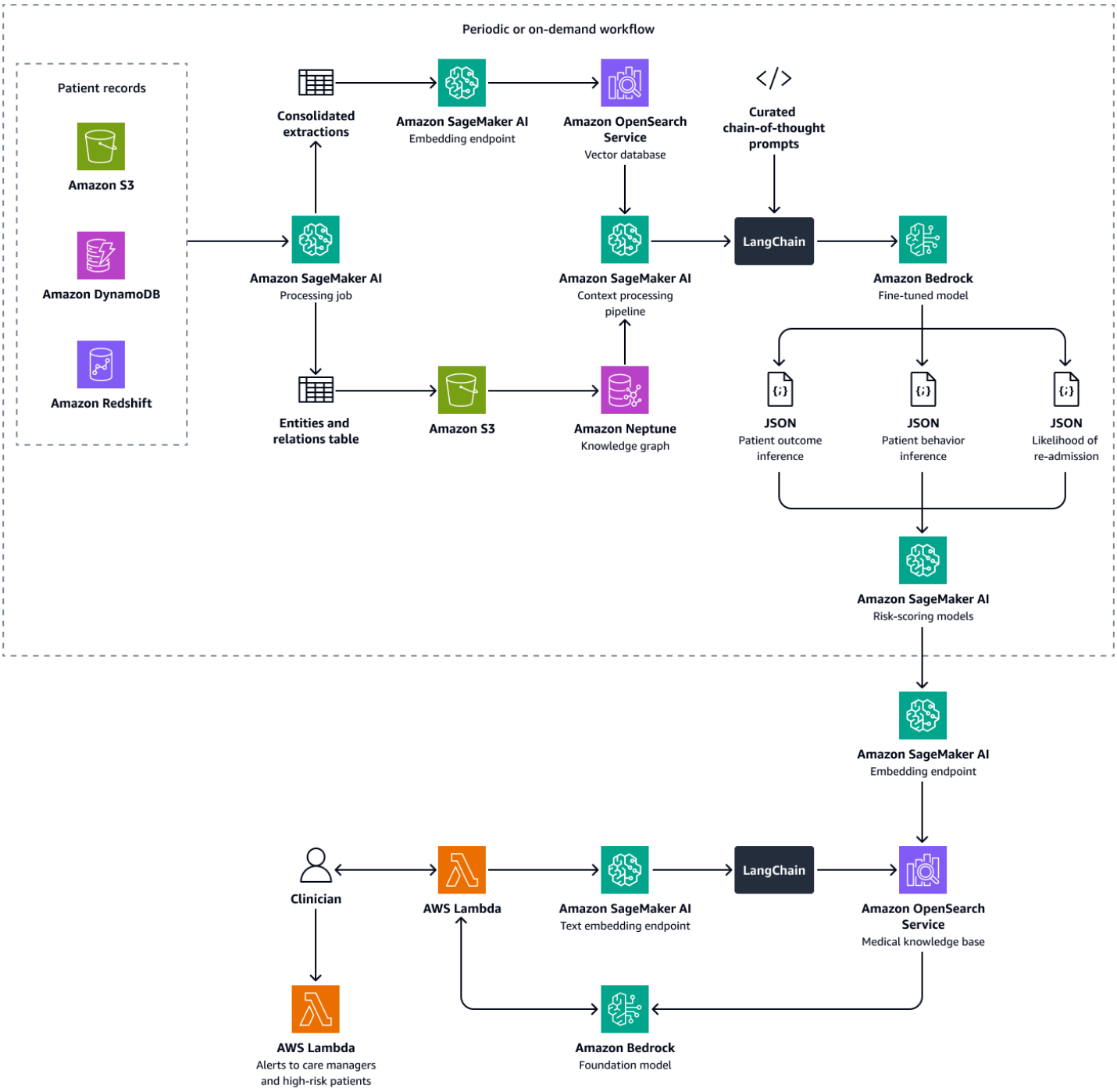
Übersicht über die Lösung

Diese Lösung verwendet ein RAG-Framework (Multi-Retriever Retrieval Augmented Generation) zur Analyse von Patientendaten. Es prognostiziert die Wahrscheinlichkeit einer erneuten Aufnahme in ein Krankenhaus für einzelne Patienten und hilft Ihnen bei der Berechnung eines Punkts für die Wahrscheinlichkeit einer Wiedereinweisung in ein Krankenhaus. Diese Lösung integriert die folgenden Funktionen:

- Wissensdiagramm — Speichert strukturierte, chronologische Patientendaten wie Krankenhausbesuche, frühere Wiedereinweisungen, Symptome, Laborergebnisse, verschriebene Behandlungen und die Historie der Medikamenteneinnahme

- Vektordatenbank — Speichert unstrukturierte klinische Daten wie Zusammenfassungen von Entlassungen, Arztnotizen und Aufzeichnungen über verpasste Termine oder gemeldete Arzneimittelnebenwirkungen
- Fein abgestimmtes LLM — Nutzt sowohl strukturierte Daten aus dem Wissensgraphen als auch unstrukturierte Daten aus der Vektordatenbank, um Rückschlüsse auf das Verhalten eines Patienten, die Therapietreue und die Wahrscheinlichkeit einer erneuten Aufnahme zu ziehen

Die Risikoeinstufungsmodelle quantifizieren die Schlussfolgerungen aus dem LLM in numerischen Werten. Sie können die Punktzahlen zu einer Bewertung der Wiedereinweisungsneigung auf Krankenhausebene zusammenfassen. Dieser Wert definiert die Risikoexposition jedes Patienten, und Sie können ihn regelmäßig oder nach Bedarf berechnen. Alle Schlussfolgerungen und Risikobewertungen werden indexiert und in Amazon OpenSearch Service gespeichert, sodass Pflegemanager und Kliniker sie abrufen können. Durch die Integration eines dialogorientierten KI-Agenten in diese Vektordatenbank können Ärzte und Pflegemanager nahtlos Erkenntnisse für einzelne Patienten, für die gesamte Einrichtung oder für einzelne medizinische Fachgebiete gewinnen. Sie können auch automatische Warnmeldungen auf der Grundlage von Risikobewertungen einrichten, was proaktive Interventionen fördert.



Der Aufbau dieser Lösung besteht aus den folgenden Schritten:

- Schritt 1: Vorhersage der Behandlungsergebnisse mithilfe eines medizinischen Wissensgraphen
- Schritt 2: Vorhersage des Verhaltens von Patienten gegenüber verschriebenen Medikamenten oder Behandlungen
- Schritt 3: Vorhersage der Wahrscheinlichkeit einer erneuten Aufnahme von Patienten

Schritt 1: Vorhersage der Behandlungsergebnisse mithilfe eines medizinischen Wissensgraphen

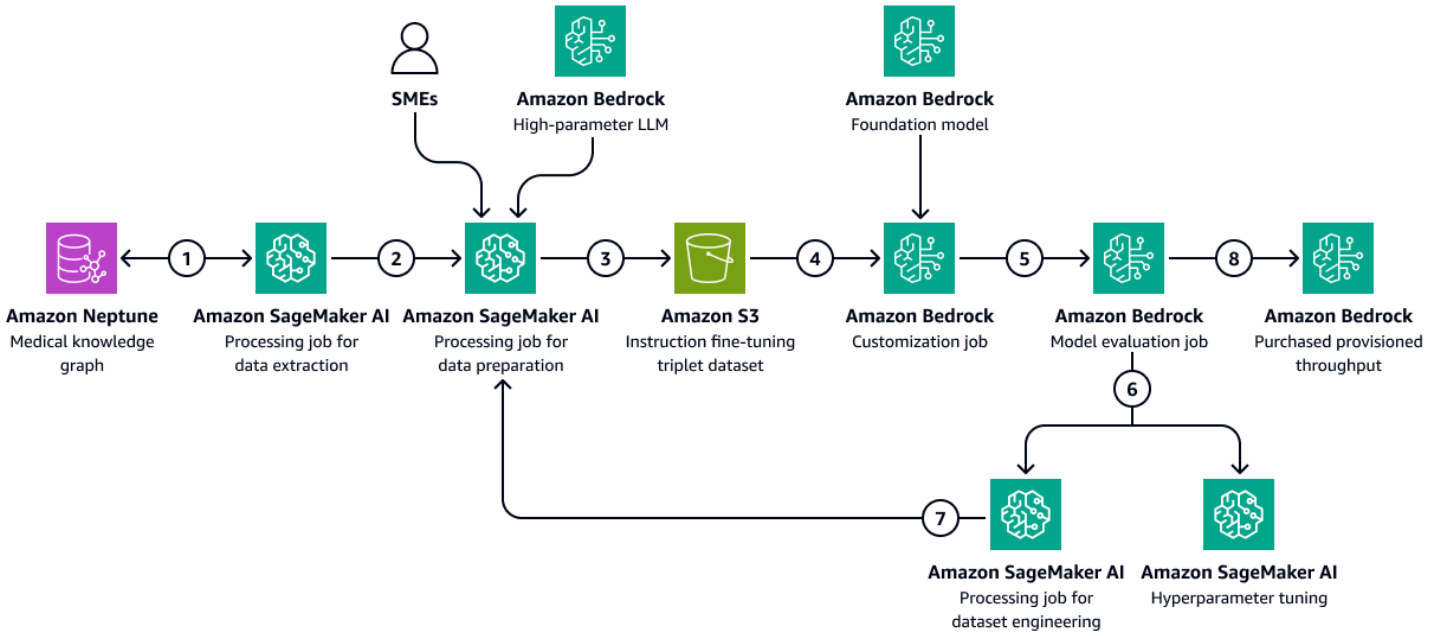
In Amazon Neptune können Sie einen Wissensgraphen verwenden, um zeitliches Wissen über Patientenbesuche und Behandlungsergebnisse im Zeitverlauf zu speichern. Die effektivste Methode, einen Wissensgraphen zu erstellen und zu speichern, ist die Verwendung eines Graphmodells und einer Graphdatenbank. Graphdatenbanken wurden speziell zum Speichern und Navigieren in Beziehungen entwickelt. Graphdatenbanken erleichtern die Modellierung und Verwaltung stark vernetzter Daten und verfügen über flexible Schemas.

Der Wissensgraph hilft Ihnen bei der Durchführung von Zeitreihenanalysen. Im Folgenden sind die wichtigsten Elemente der Graphdatenbank aufgeführt, die für die zeitliche Vorhersage von Behandlungsergebnissen verwendet werden:

- Historische Daten — Frühere Diagnosen, fortgesetzte Medikamente, zuvor verwendete Medikamente und Laborergebnisse für den Patienten
- Patientenbesuche (chronologisch) — Besuchstermine, Symptome, beobachtete Allergien, klinische Notizen, Diagnosen, Verfahren, Behandlungen, verschriebene Medikamente und Laborergebnisse
- Symptome und klinische Parameter — Klinische und symptombasierte Informationen, einschließlich Schweregrad, Verlaufsmuster und Reaktion des Patienten auf das Medikament

Sie können die Erkenntnisse aus dem Medical Knowledge Graph nutzen, um ein LLM in Amazon Bedrock, wie Llama 3, zu optimieren. Sie optimieren das LLM anhand sequentieller Patientendaten über die Reaktion des Patienten auf eine Reihe von Medikamenten oder Behandlungen im Laufe der Zeit. Verwenden Sie einen beschrifteten Datensatz, der eine Reihe von Medikamenten oder Behandlungen sowie Daten zur Interaktion zwischen Patient und Klinik in vordefinierte Kategorien unterteilt, die den Gesundheitszustand eines Patienten angeben. Beispiele für diese Kategorien sind Verschlechterung des Gesundheitszustands, Verbesserung oder stabiler Fortschritt. Wenn der Arzt einen neuen Kontext über den Patienten und seine Symptome eingibt, kann das fein abgestimmte LLM die Muster aus dem Trainingsdatensatz verwenden, um das mögliche Behandlungsergebnis vorherzusagen.

Die folgende Abbildung zeigt die aufeinanderfolgenden Schritte, die zur Feinabstimmung eines LLM in Amazon Bedrock mithilfe eines gesundheitsspezifischen Trainingsdatensatzes erforderlich sind. Diese Daten können den Gesundheitszustand der Patienten und die Reaktionen auf Behandlungen im Laufe der Zeit beinhalten. Dieser Trainingsdatensatz würde dem Modell helfen, allgemeine Vorhersagen über die Behandlungsergebnisse zu treffen.



Das Diagramm zeigt den folgenden Workflow:

1. Der Amazon SageMaker KI-Datenextraktionsjob fragt den Wissensgraphen ab, um chronologische Daten über die Reaktionen verschiedener Patienten auf eine Reihe von Medikamenten oder Behandlungen im Laufe der Zeit abzurufen.
2. Der Job zur SageMaker KI-Datenvorbereitung beinhaltet ein Amazon Bedrock LLM und Beiträge von Fachexperten ()SMEs. Der Job unterteilt die aus dem Knowledge Graph abgerufenen Daten in vordefinierte Kategorien (z. B. Verschlechterung des Gesundheitszustands, Verbesserung oder stabiler Verlauf), die den Gesundheitszustand jedes Patienten angeben.
3. Bei der Aufgabe wird ein Datensatz zur Feinabstimmung erstellt, der die aus dem Wissensgraphen extrahierten Informationen, die chain-of-thought Eingabeaufforderungen und die Kategorie der Behandlungsergebnisse enthält. Es lädt diesen Trainingsdatensatz in einen Amazon S3 S3-Bucket hoch.
4. Ein Amazon Bedrock-Anpassungsjob verwendet diesen Trainingsdatensatz zur Feinabstimmung eines LLM.

5. Der Amazon Bedrock-Anpassungsjob integriert das Amazon Bedrock-Basismodell der Wahl in die Schulungsumgebung. Er startet den Feinabstimmungsjob und verwendet den Trainingsdatensatz und die von Ihnen konfigurierten Trainingshyperparameter.
6. Ein Amazon Bedrock-Evaluierungsjob bewertet das fein abgestimmte Modell mithilfe eines vorgefertigten Frameworks für die Modellbewertung.
7. Wenn das Modell verbessert werden muss, wird der Trainingsjob nach sorgfältiger Prüfung des Trainingsdatensatzes erneut mit mehr Daten ausgeführt. Wenn das Modell keine inkrementelle Leistungsverbesserung zeigt, sollten Sie auch eine Änderung der Trainingshyperparameter in Betracht ziehen.
8. Nachdem die Modellevaluierung die von den Geschäftsbeteiligten definierten Standards erfüllt hat, hosten Sie das fein abgestimmte Modell für den von Amazon Bedrock bereitgestellten Durchsatz.

Schritt 2: Vorhersage des Verhaltens von Patienten gegenüber verschriebenen Medikamenten oder Behandlungen

LLMs Fine-Tune kann klinische Notizen, Entlassungszusammenfassungen und andere patientenspezifische Dokumente aus dem temporalen medizinischen Wissensdiagramm verarbeiten. Sie können beurteilen, ob der Patient wahrscheinlich verschriebene Medikamente oder Behandlungen einhält.

In diesem Schritt wird der in erstellte Wissensgraph verwendetSchritt 1: Vorhersage der Behandlungsergebnisse mithilfe eines medizinischen Wissensgraphen. Der Wissensgraph enthält Daten aus dem Patientenprofil, einschließlich der bisherigen Adhärenz des Patienten als Knoten. Als Merkmale solcher Knoten werden auch Fälle von Nichteinhaltung von Medikamenten oder Behandlungen, Nebenwirkungen von Medikamenten, mangelndem Zugang zu Medikamenten oder Kostenbarrieren oder komplexe Dosierungsschemata berücksichtigt.

Fine-tuned LLMs kann frühere Daten zur Verschreibungserfüllung aus dem Medical Knowledge Graph und beschreibende Zusammenfassungen der klinischen Notizen aus einer Amazon OpenSearch Service-Vektordatenbank nutzen. In diesen klinischen Notizen können häufig versäumte Termine oder die Nichteinhaltung von Behandlungen erwähnt werden. Das LLM kann diese Hinweise verwenden, um die Wahrscheinlichkeit einer future Nichteinhaltung vorherzusagen.

1. Bereiten Sie die Eingabedaten wie folgt vor:
   - Strukturierte Daten — Extrahieren Sie aktuelle Patientendaten, z. B. die letzten drei Besuche und die Laborergebnisse, aus dem medizinischen Wissensdiagramm.

- Unstrukturierte Daten — Rufen Sie die neuesten klinischen Notizen aus der Amazon OpenSearch Service-Vektordatenbank ab.

2. Erstellen Sie eine Eingabeaufforderung, die die Krankengeschichte und den aktuellen Kontext enthält. Im Folgenden finden Sie ein Beispiel für eine Eingabeaufforderung:

```
You are a highly specialized AI model trained in healthcare predictive analytics.
 Your task is to analyze a patient's historical medical records, adherence patterns,
 and clinical context to predict the **likelihood of future non-adherence** to
 prescribed medications or treatments.

### **Patient Details**
- **Patient ID:** {patient_id}
- **Age:** {age}
- **Gender:** {gender}
- **Medical Conditions:** {medical_conditions}
- **Current Medications:** {current_medications}
- **Prescribed Treatments:** {prescribed_treatments}

### **Chronological Medical History**
- **Visit Dates & Symptoms:** {visit_dates_symptoms}
- **Diagnoses & Procedures:** {diagnoses_procedures}
- **Prescribed Medications & Treatments:** {medications_treatments}
- **Past Adherence Patterns:** {historical_adherence}
- **Instances of Non-Adherence:** {past_non_adherence}
- **Side Effects Experienced:** {side_effects}
- **Barriers to Adherence (e.g., Cost, Access, Dosing Complexity):** {barriers}

### **Patient-Specific Insights**
- **Clinical Notes & Discharge Summaries:** {clinical_notes}
- **Missed Appointments & Non-Compliance Patterns:** {missed_appointments}

### **Let's think Step-by-Step to predict the patient behaviour**
1. You should first analyze past adherence trends and patterns of non-adherence.
2. Identify potential barriers, such as financial constraints, medication side
 effects, or complex dosing regimens.
3. Thoroughly examine clinical notes and documented patient behaviors that may hint
 at non-adherence.
4. Correlate adherence history with prescribed treatments and patient conditions.
5. Finally predict the likelihood of non-adherence based on these contextual
 insights.

### **Output Format (JSON)**
```

```
Return the prediction in the following structured format:
```json
{
  "patient_id": "{patient_id}",
  "likelihood_of_non_adherence": "{low | moderate | high}",
  "reasoning": "{detailed_explanation_based_on_patient_history}"
}
```

3. Übergeben Sie die Aufforderung an das fein abgestimmte LLM. Das LLM verarbeitet die Aufforderung und prognostiziert das Ergebnis. Im Folgenden finden Sie ein Beispiel für eine Antwort des LLM:

```
{
  "patient_id": "P12345",
  "likelihood_of_non_adherence": "high",
  "reasoning": "The patient has a history of missed appointments, has reported side
 effects to previous medications. Additionally, clinical notes indicate difficulty
 following complex dosing schedules."
}
```

4. Analysieren Sie die Antwort des Modells, um die Kategorie der prognostizierten Ergebnisse zu extrahieren. Bei der Kategorie für die Beispielantwort im vorherigen Schritt könnte es sich beispielsweise um eine hohe Wahrscheinlichkeit einer Nichteinhaltung handeln.
5. (Optional) Verwenden Sie Modelllogits oder zusätzliche Methoden, um Konfidenzwerte zuzuweisen. Logits sind die nicht normalisierten Wahrscheinlichkeiten, dass das Objekt zu einer bestimmten Klasse oder Kategorie gehört.

Schritt 3: Vorhersage der Wahrscheinlichkeit einer erneuten Aufnahme von Patienten

Wiedereinweisungen in Krankenhäuser sind aufgrund der hohen Kosten der Gesundheitsverwaltung und ihrer Auswirkungen auf das Wohlbefinden der Patienten ein großes Problem. Die Berechnung der Wiedereinweisungsraten in Krankenhäuser ist eine Möglichkeit, die Qualität der Patientenversorgung und die Leistung eines Gesundheitsdienstleisters zu messen.

Um die Wiederaufnahmequote zu berechnen, haben Sie einen Indikator definiert, z. B. eine Wiederaufnahmequote von 7 Tagen. Dieser Indikator gibt den Prozentsatz der aufgenommenen Patienten an, die innerhalb von sieben Tagen nach der Entlassung zu einem ungeplanten
```

Besuch ins Krankenhaus zurückkehren. Um die Wahrscheinlichkeit einer erneuten Aufnahme eines Patienten vorherzusagen, kann ein fein abgestimmtes LLM Zeitdaten aus dem medizinischen Wissensdiagramm verwenden, das Sie in erstellt haben. Schritt 1: Vorhersage der Behandlungsergebnisse mithilfe eines medizinischen Wissensgraphen In diesem Wissensdiagramm werden chronologische Aufzeichnungen über Begegnungen, Behandlungen, Medikamente und Symptome mit Patienten geführt. Diese Datensätze enthalten Folgendes:

- Dauer seit der letzten Entlassung des Patienten
- Reaktion des Patienten auf frühere Behandlungen und Medikamente
- Das Fortschreiten von Symptomen oder Zuständen im Laufe der Zeit

Sie können diese Zeitreihenereignisse verarbeiten, um anhand einer kuratierten Systemaufforderung die Wahrscheinlichkeit einer erneuten Aufnahme eines Patienten vorherzusagen. Die Aufforderung gibt die Prognoselogik an das fein abgestimmte LLM weiter.

1. Bereiten Sie die Eingabedaten wie folgt vor:
   - Verlauf der Therapietreue — Extrahieren Sie Daten zur Medikamentenabholung, Häufigkeit der Medikamentennachfüllung, Diagnose- und Medikamentendetails, chronologische Krankengeschichte und andere Informationen aus der medizinischen Wissensgrafik.
   - Verhaltensindikatoren — Rufen Sie klinische Notizen zu verpassten Terminen und von Patienten gemeldeten Nebenwirkungen ab und fügen Sie sie hinzu.
2. Erstellen Sie eine Eingabeaufforderung, die den Verlauf der Einhaltung und die Verhaltensindikatoren enthält. Im Folgenden finden Sie ein Beispiel für eine Eingabeaufforderung:

```
You are a highly specialized AI model trained in healthcare predictive analytics.
 Your task is to analyze a patient's historical medical records, clinical events, and
 adherence patterns to predict the **likelihood of hospital readmission** within the
 next few days.

### **Patient Details**
- **Patient ID:** {patient_id}
- **Age:** {age}
- **Gender:** {gender}
- **Primary Diagnoses:** {diagnoses}
- **Current Medications:** {current_medications}
- **Prescribed Treatments:** {prescribed_treatments}

### **Chronological Medical History**
```

```
- **Recent Hospital Encounters:** {encounters}
- **Time Since Last Discharge:** {time_since_last_discharge}
- **Previous Readmissions:** {past_readmissions}
- **Recent Lab Results & Vital Signs:** {recent_lab_results}
- **Procedures Performed:** {procedures_performed}
- **Prescribed Medications & Treatments:** {medications_treatments}
- **Past Adherence Patterns:** {historical_adherence}
- **Instances of Non-Adherence:** {past_non_adherence}

### **Patient-Specific Insights**
- **Clinical Notes & Discharge Summaries:** {clinical_notes}
- **Missed Appointments & Non-Compliance Patterns:** {missed_appointments}
- **Patient-Reported Side Effects & Complications:** {side_effects}

### **Reasoning Process – You have to analyze this use case step-by-step.**
1. First assess **time since last discharge** and whether recent hospital encounters
 suggest a pattern of frequent readmissions.
2. Second examine **recent lab results, vital signs, and procedures performed** to
 identify clinical deterioration.
3. Third analyze **adherence history**, checking if past non-adherence to medications
 or treatments correlates with readmissions.
4. Then identify **missed appointments, self-reported side effects, or symptoms
 worsening** from clinical notes.
5. Finally predict the **likelihood of readmission** based on these contextual
 insights.

### **Output Format (JSON)**
Return the prediction in the following structured format:
```json
{
  "patient_id": "{patient_id}",
  "likelihood_of_readmission": "{low | moderate | high}",
  "reasoning": "{detailed_explanation_based_on_patient_history}"
}
```
```

3. Übergeben Sie die Aufforderung an das fein abgestimmte LLM. Das LLM bearbeitet die Aufforderung und prognostiziert die Wahrscheinlichkeit einer erneuten Zulassung sowie die Gründe dafür. Im Folgenden finden Sie ein Beispiel für eine Antwort des LLM:

```
{
  "patient_id": "P67890",
  "likelihood_of_readmission": "high",
```

```
  "reasoning": "The patient was discharged only 5 days ago, has a history of more
 than two readmissions to hospitals where the patient received treatment. Recent
 lab results indicate abnormal kidney function and high liver enzymes. These factors
 suggest a medium risk of readmission."
}
```

4. Ordnen Sie die Vorhersage in eine standardisierte Skala ein, z. B. niedrig, mittel oder hoch.
5. Überprüfen Sie die vom LLM vorgelegten Überlegungen und identifizieren Sie die wichtigsten Faktoren, die zur Vorhersage beitragen.
6. Ordnen Sie die qualitativen Ergebnisse quantitativen Werten zu. Ein sehr hoher Wert könnte beispielsweise einer Wahrscheinlichkeit von 0,9 entsprechen.
7. Verwenden Sie Validierungsdatensätze, um die Modellergebnisse anhand der tatsächlichen Wiederzulassungsraten zu kalibrieren.

Schritt 4: Berechnung des Punkts für die Wahrscheinlichkeit einer Wiedereinweisung ins Krankenhaus

Als Nächstes berechnen Sie einen Wert für die Wahrscheinlichkeit einer erneuten Aufnahme in ein Krankenhaus pro Patient. Dieser Wert spiegelt die Nettoauswirkung der drei Analysen wider, die in den vorherigen Schritten durchgeführt wurden: potenzielle Behandlungsergebnisse, Verhalten der Patienten gegenüber Medikamenten und Behandlungen sowie Wahrscheinlichkeit einer erneuten Aufnahme von Patienten. Indem Sie den Wert der Wiedereinweisungsneigung auf Patientenebene nach Fachgebieten und dann auf Krankenhausebene aggregieren, können Sie Erkenntnisse für Kliniker, Pflegemanager und Administratoren gewinnen. Der Wert für die Wahrscheinlichkeit einer Wiederaufnahme in ein Krankenhaus hilft Ihnen dabei, die Gesamtleistung nach Einrichtung, Fachgebiet oder Erkrankung zu beurteilen. Anschließend können Sie diesen Wert verwenden, um proaktive Maßnahmen zu ergreifen.

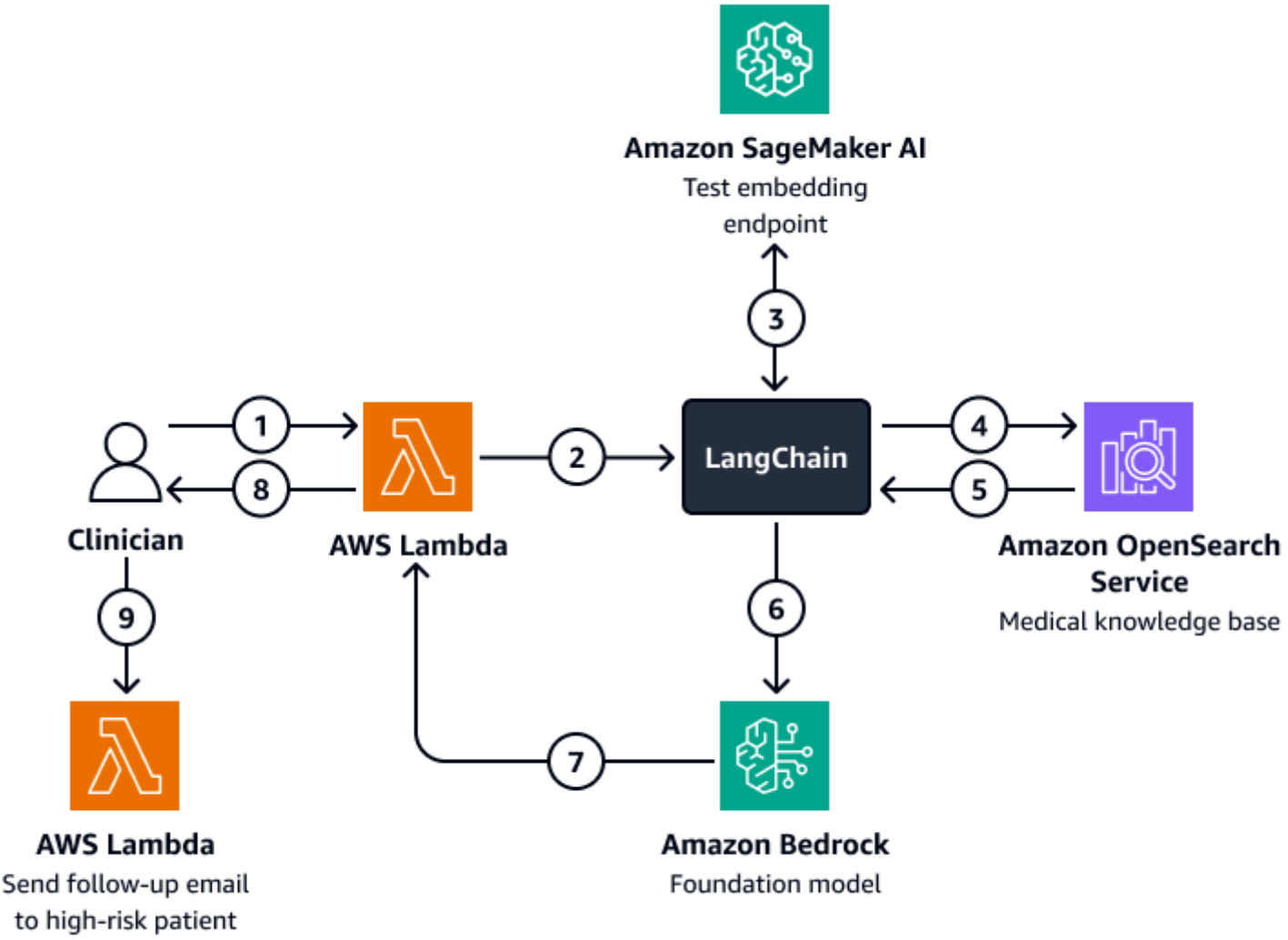
1. Weisen Sie jedem der verschiedenen Faktoren Gewichte zu (Prognose des Ergebnisses, Wahrscheinlichkeit der Einhaltung der Vorschriften, Wiederzulassung). Im Folgenden sind Beispielgewichte aufgeführt:
   - Gewicht für die Vorhersage des Ergebnisses: 0,4
   - Gewicht bei der Vorhersage der Einhaltung der Regeln: 0,3
   - Gewicht der Wahrscheinlichkeit einer erneuten Zulassung: 0,3
2. Verwenden Sie die folgende Berechnung, um die Gesamtpunktzahl zu berechnen:

```
ReadadmissionPropensityScore = (OutcomeScore × OutcomeWeight) +
(AdherenceScore × AdherenceWeight) +
(ReadmissionLikelihoodScore × ReadmissionLikelihoodWeight)
```

3. Stellen Sie sicher, dass sich alle Einzelwerte auf derselben Skala befinden, z. B. 0 bis 1.
4. Definieren Sie die Schwellenwerte für Maßnahmen. Beispielsweise lösen Werte über 0,7 Warnmeldungen aus.

Auf der Grundlage der oben genannten Analysen und der Bewertung der Wiederaufnahmebereitschaft eines Patienten können Ärzte oder Pflegemanager Warnmeldungen einrichten, um ihre einzelnen Patienten auf der Grundlage des berechneten Scores zu überwachen. Liegt der Wert über einem vordefinierten Schwellenwert, werden sie benachrichtigt, wenn dieser Schwellenwert erreicht ist. Dies hilft Pflegemanagern, bei der Erstellung von Entlassungsplänen für ihre Patienten proaktiv und nicht reaktiv vorzugehen. Speichern Sie die Ergebnisse, das Verhalten und die Wahrscheinlichkeit der Wiederaufnahme von Patienten in indexierter Form in einer Amazon OpenSearch Service-Vektordatenbank, sodass das Pflegepersonal sie mithilfe eines Konversations-AI-Agenten problemlos abrufen kann.

Das folgende Diagramm zeigt den Arbeitsablauf eines KI-Konversationsassistenten, den ein Arzt oder Pflegemanager nutzen kann, um Erkenntnisse über die Behandlungsergebnisse, das erwartete Verhalten und die Wahrscheinlichkeit einer erneuten Aufnahme von Patienten abzurufen. Benutzer können Erkenntnisse auf Patienten-, Abteilungs- oder Krankenhausebene abrufen. Der KI-Agent ruft diese Erkenntnisse ab, die in indizierter Form in einer Amazon OpenSearch Service-Vektordatenbank gespeichert werden. Der Mitarbeiter verwendet die Abfrage, um relevante Daten abzurufen, und gibt maßgeschneiderte Antworten, einschließlich Handlungsempfehlungen für Patienten, bei denen ein hohes Risiko besteht, erneut aufgenommen zu werden. Je nach Risikograd kann der Mitarbeiter auch Erinnerungen für Patienten und Pflegepersonal einrichten.



Das Diagramm zeigt den folgenden Workflow:

1. Der Kliniker stellt eine Frage an einen KI-Konversationsagenten, der eine Funktion beherbergt. AWS Lambda
2. Die Lambda-Funktion initiiert eine LangChain Agent.
3. Das Tool LangChain Der Agent sendet die Frage des Benutzers an einen Amazon SageMaker AI-Texteinbettungsendpunkt. Der Endpunkt bettet die Frage ein.
4. Das Tool LangChain Der Mitarbeiter leitet die eingebettete Frage an eine medizinische Wissensdatenbank in Amazon OpenSearch Service weiter.
5. Amazon OpenSearch Service gibt die spezifischen Erkenntnisse, die für die Benutzeranfrage am relevantesten sind, an die LangChain Agent.
6. Das Tool LangChain Agenten senden die Anfrage und den abgerufenen Kontext aus der Wissensdatenbank an ein Amazon Bedrock Foundation-Modell.

7. Das Amazon Bedrock Foundation-Modell generiert eine Antwort und sendet sie an die Lambda-Funktion.
8. Die Lambda-Funktion gibt die Antwort an den Arzt zurück.
9. Der Arzt initiiert eine Lambda-Funktion, die eine Folge-E-Mail an einen Patienten sendet, bei dem ein hohes Risiko einer erneuten Aufnahme besteht.

Ausrichtung auf das AWS Well-Architected Framework

Die Architektur zur Erfassung des Patientenverhaltens und zur Prognose der Wiedereinweisungsraten von Krankenhäusern integriert AWS-Services medizinische Wissensdiagramme und LLMs verbessert die Behandlungsergebnisse und orientiert sich gleichzeitig an den sechs Säulen des AWS Well-Architected Framework:

- Operational Excellence — Bei der Lösung handelt es sich um ein entkoppeltes, automatisiertes System, das Amazon Bedrock verwendet und AWS Lambda Warnmeldungen in Echtzeit ausgibt.
- Sicherheit — Diese Lösung wurde entwickelt, um Gesundheitsvorschriften wie HIPAA zu erfüllen. Sie können auch Verschlüsselung, detaillierte Zugriffskontrolle und Amazon Bedrock Guardrails implementieren, um Patientendaten zu schützen.
- Zuverlässigkeit — Die Architektur verwendet fehlertolerante, serverlose Systeme. AWS-Services
- Leistungseffizienz — Amazon OpenSearch Service und die Feinabstimmung LLMs können schnelle und genaue Prognosen liefern.
- Kostenoptimierung — Serverlose Technologien und pay-per-inference Modelle tragen zur Kostenminimierung bei. Die Verwendung eines fein abgestimmten LLM kann zwar mit zusätzlichen Kosten verbunden sein, das Modell verwendet jedoch einen RAG-Ansatz, der den Daten- und Rechenaufwand für den Feinabstimmungsprozess reduziert.
- Nachhaltigkeit — Die Architektur minimiert den Ressourcenverbrauch durch den Einsatz einer serverlosen Infrastruktur. Sie unterstützt auch effiziente, skalierbare Abläufe im Gesundheitswesen.

Anwendungsfall: Verwaltung und Weiterbildung Ihres Gesundheitspersonals

Die Umsetzung von Strategien zur Transformation und Weiterbildung von Talenten hilft den Mitarbeitern, neue Technologien und Praktiken im medizinischen Bereich und im Gesundheitswesen weiterhin kompetent einzusetzen. Proaktive Weiterbildungsinitiativen stellen sicher, dass medizinisches Fachpersonal eine qualitativ hochwertige Patientenversorgung bieten, die betriebliche Effizienz optimieren und die gesetzlichen Standards einhalten kann. Darüber hinaus fördert die Transformation von Talenten eine Kultur des kontinuierlichen Lernens. Dies ist entscheidend für die Anpassung an die sich verändernde Gesundheitslandschaft und die Bewältigung neuer Herausforderungen im Bereich der öffentlichen Gesundheit. Traditionelle Schulungsansätze wie Präsenzunterricht und statische Lernmodule bieten einheitliche Inhalte für ein breites Publikum. Ihnen mangelt es oft an personalisierten Lernwegen, die entscheidend sind, um auf die spezifischen Bedürfnisse und das Qualifikationsniveau der einzelnen Praktiker einzugehen. Diese one-size-fits-all Strategie kann zu mangelndem Engagement und einer suboptimalen Wissenserhaltung führen.

Folglich müssen Gesundheitsorganisationen innovative, skalierbare und technologiegestützte Lösungen einsetzen, mit denen sich die Unterschiede für jeden ihrer Mitarbeiter in ihrem aktuellen Zustand und in ihrem potenziellen future Zustand ermitteln lassen. Diese Lösungen sollten hyperpersonalisierte Lernpfade und die richtigen Lerninhalte empfehlen. Dies bereitet die Belegschaft effektiv auf die future des Gesundheitswesens vor.

In der Gesundheitsbranche können Sie generative KI einsetzen, um Ihre Belegschaft besser zu verstehen und weiterzubilden. Durch die Verbindung von großen Sprachmodellen (LLMs) und fortgeschrittenen Retrievern können Unternehmen verstehen, über welche Fähigkeiten sie derzeit verfügen, und Schlüsselkompetenzen identifizieren, die in future erforderlich sein könnten. Diese Informationen helfen Ihnen, diese Lücke zu schließen, indem Sie neue Mitarbeiter einstellen und die aktuelle Belegschaft weiterbilden. Mithilfe von Amazon Bedrock und Knowledge Graphs können Organisationen im Gesundheitswesen domänenspezifische Anwendungen entwickeln, die kontinuierliches Lernen und die Weiterentwicklung von Fähigkeiten ermöglichen.

Das durch diese Lösung bereitgestellte Wissen hilft Ihnen dabei, Talente effektiv zu verwalten, die Leistung Ihrer Belegschaft zu optimieren, den Unternehmenserfolg zu fördern, vorhandene Fähigkeiten zu identifizieren und eine Talentstrategie zu entwickeln. Diese Lösung kann Ihnen helfen, diese Aufgaben innerhalb von Wochen statt Monaten zu erledigen.

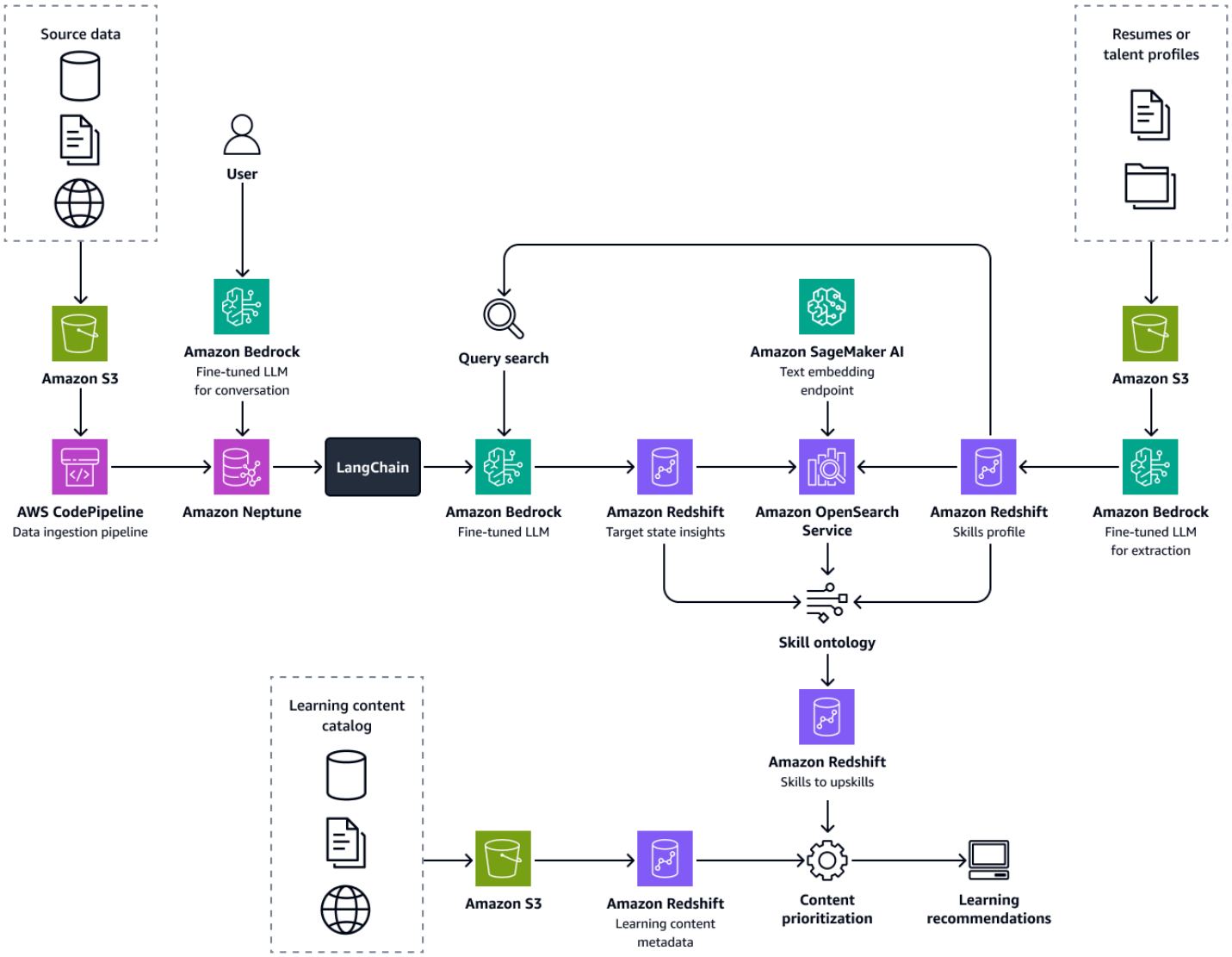
Übersicht über die Lösung

Bei dieser Lösung handelt es sich um ein Framework zur Transformation von Talenten im Gesundheitswesen, das aus den folgenden Komponenten besteht:

- Intelligenter Lebenslauf-Parser — Diese Komponente kann den Lebenslauf eines Kandidaten lesen und Kandidateninformationen, einschließlich Fähigkeiten, präzise extrahieren. Intelligente Lösung zur Informationsextraktion, die auf dem fein abgestimmten Llama 2-Modell in Amazon Bedrock basiert und auf einem firmeneigenen Schulungsdatensatz basiert, der Lebensläufe und Talentprofile aus mehr als 19 Branchen umfasst. Dieser LLM-basierte Prozess spart Hunderte von Stunden, da er die manuelle Überprüfung von Lebensläufen und die Zuordnung von Top-Kandidaten zu offenen Stellen automatisiert.
- Knowledge Graph — Ein Wissensgraph, der auf Amazon Neptune basiert, einer vereinheitlichten Sammlung von Talentinformationen, einschließlich der Rollen- und Qualifikationstaxonomie des Unternehmens sowie der Branche. Er erfasst die Semantik von Talenten im Gesundheitswesen anhand von Definitionen von Fähigkeiten, Rollen und deren Eigenschaften, Beziehungen und logischen Einschränkungen.
- Qualifikationsontologie — Die Entdeckung von Qualifikationsnähe zwischen den Fähigkeiten der Kandidaten und den Fähigkeiten im Idealzustand oder in der future (abgerufen mithilfe eines Wissensgraphen) wird durch Ontologiealgorithmen erreicht, die die semantische Ähnlichkeit zwischen Kandidatenkompetenzen und Fähigkeiten im Zielstatus messen.
- Lernweg und Lerninhalte — Bei dieser Komponente handelt es sich um eine Engine für Lernempfehlungen, die auf Grundlage der identifizierten Qualifikationslücken die richtigen Lerninhalte aus einem Katalog von Lernmaterialien beliebiger Anbieter empfehlen kann. Identifizierung der optimalen Weiterbildungswege für jeden Kandidaten durch Analyse der Qualifikationslücken und Empfehlung priorisierter Lerninhalte, um jedem Kandidaten eine reibungslose und kontinuierliche berufliche Entwicklung während des Übergangs zu einer neuen Rolle zu ermöglichen.

Diese cloudbasierte, automatisierte Lösung basiert auf Diensten für maschinelles Lernen LLMs, Wissensgraphen und Retrieval Augmented Generation (RAG). Es kann skaliert werden, um Zehntausende von Lebensläufen in kürzester Zeit zu verarbeiten, sofortige Kandidatenprofile zu erstellen, Lücken in ihrem aktuellen oder potenziellen future Status zu identifizieren und dann effizient die richtigen Lerninhalte zu empfehlen, um diese Lücken zu schließen.

Die folgende Abbildung zeigt den end-to-end Ablauf des Frameworks. Die Lösung basiert auf der Feinabstimmung LLMs in Amazon Bedrock. Diese LLMs rufen Daten aus der Wissensdatenbank für Talente im Gesundheitswesen in Amazon Neptune ab. Datengestützte Algorithmen geben Empfehlungen für optimale Lernwege für jeden Kandidaten.



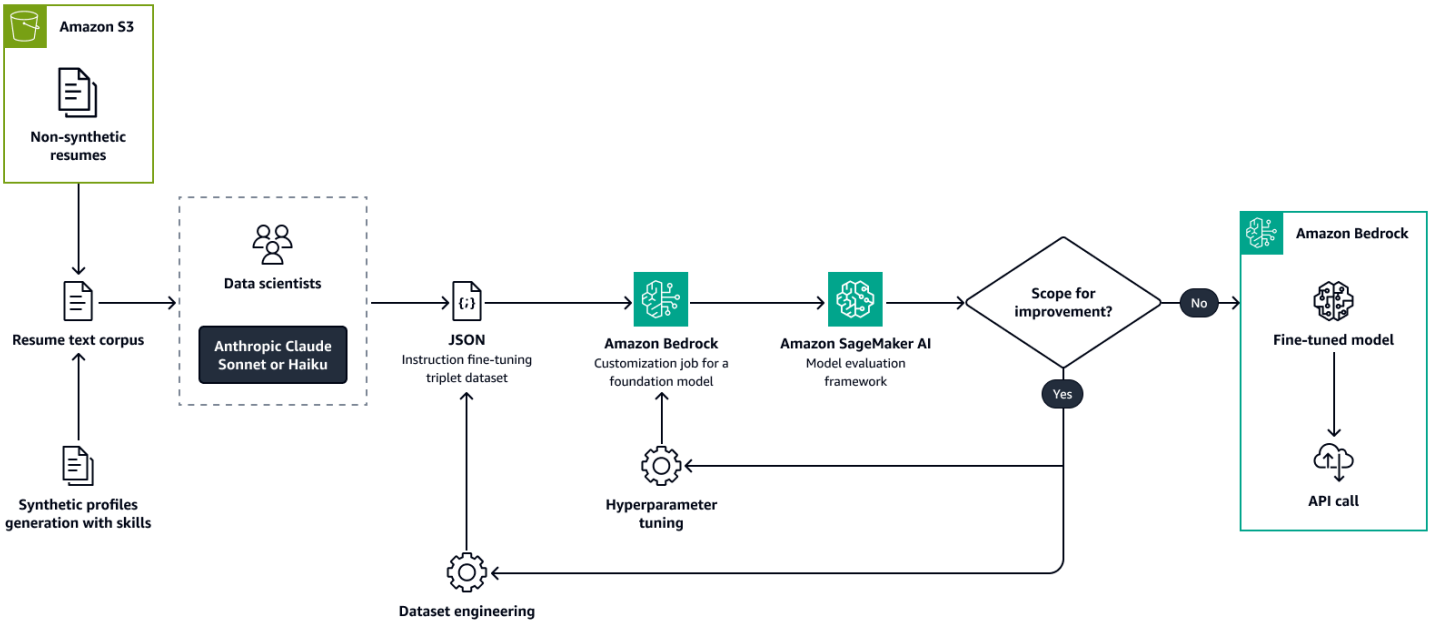
Der Aufbau dieser Lösung besteht aus den folgenden Schritten:

- Schritt 1: Extraktion von Talentinformationen und Erstellung eines Kompetenzprofils
- Schritt 2: Anhand eines role-to-skill Wissensgraphen die Relevanz ermitteln
- Schritt 3: Identifizierung von Qualifikationslücken und Empfehlung von Schulungen

Schritt 1: Extraktion von Talentinformationen und Erstellung eines Kompetenzprofils

Zunächst optimieren Sie ein umfangreiches Sprachmodell wie Llama 2 in Amazon Bedrock mit einem benutzerdefinierten Datensatz. Dadurch wird das LLM an den Anwendungsfall angepasst. Während der Schulung extrahieren Sie präzise und konsistent wichtige Talentattribute aus Lebensläufen von Kandidaten oder ähnlichen Talentprofilen. Zu diesen Talentattributen gehören Fähigkeiten, aktuelle Berufsbezeichnung, Berufsbezeichnungen mit Zeiträumen, Ausbildung und Zertifizierungen. Weitere Informationen finden Sie in der Amazon Bedrock-Dokumentation unter Passen Sie Ihr Modell an, um seine Leistung für Ihren Anwendungsfall zu verbessern.

Die folgende Abbildung zeigt den Prozess zur Feinabstimmung eines Modells zur Analyse von Lebensläufen mithilfe von Amazon Bedrock. Sowohl echte als auch synthetisch erstellte Lebensläufe werden an ein LLM weitergeleitet, um wichtige Informationen zu extrahieren. Eine Gruppe von Datenwissenschaftlern validiert die extrahierten Informationen anhand des ursprünglichen Rohtextes. Die extrahierten Informationen werden dann mithilfe von chain-of-thoughtEingabeaufforderungen und dem Originaltext verkettet, um einen Trainingsdatensatz für die Feinabstimmung abzuleiten. Dieser Datensatz wird dann an einen Amazon Bedrock-Anpassungsjob übergeben, der das Modell verfeinert. Ein Amazon SageMaker AI-Batch-Job führt ein Modellevaluierungs-Framework aus, das das fein abgestimmte Modell bewertet. Wenn das Modell verbessert werden muss, wird der Job erneut mit mehr Daten oder anderen Hyperparametern ausgeführt. Nachdem die Evaluierung den Standards entspricht, hosten Sie das benutzerdefinierte Modell über den von Amazon Bedrock bereitgestellten Durchsatz.



Schritt 2: Anhand eines role-to-skill Wissensgraphen die Relevanz ermitteln

Als Nächstes erstellen Sie ein Wissensdiagramm, das die Fähigkeiten und die Rollentaxonomie Ihrer Organisation und anderer Organisationen in der Gesundheitsbranche zusammenfasst. Diese erweiterte Wissensdatenbank basiert auf aggregierten Talent- und Unternehmensdaten in Amazon Redshift. Sie können Talentdaten von einer Reihe von Anbietern von Arbeitsmarktdaten sowie aus unternehmensspezifischen strukturierten und unstrukturierten Datenquellen wie ERP-Systemen (Enterprise Resource Planning), einem Personalinformationssystem (HRIS), Lebensläufen von Mitarbeitern, Stellenbeschreibungen und Dokumenten zur Talentarchitektur sammeln.

Erstellen Sie den Wissensgraphen auf Amazon Neptune. Knoten stehen für Fähigkeiten und Rollen, und Kanten stehen für die Beziehungen zwischen ihnen. Reichern Sie dieses Diagramm mit Metadaten an, um Details wie den Namen der Organisation, die Branche, die Berufsgruppe, die Art der Qualifikation, den Rollentyp und Branchenkennzeichnungen aufzunehmen.

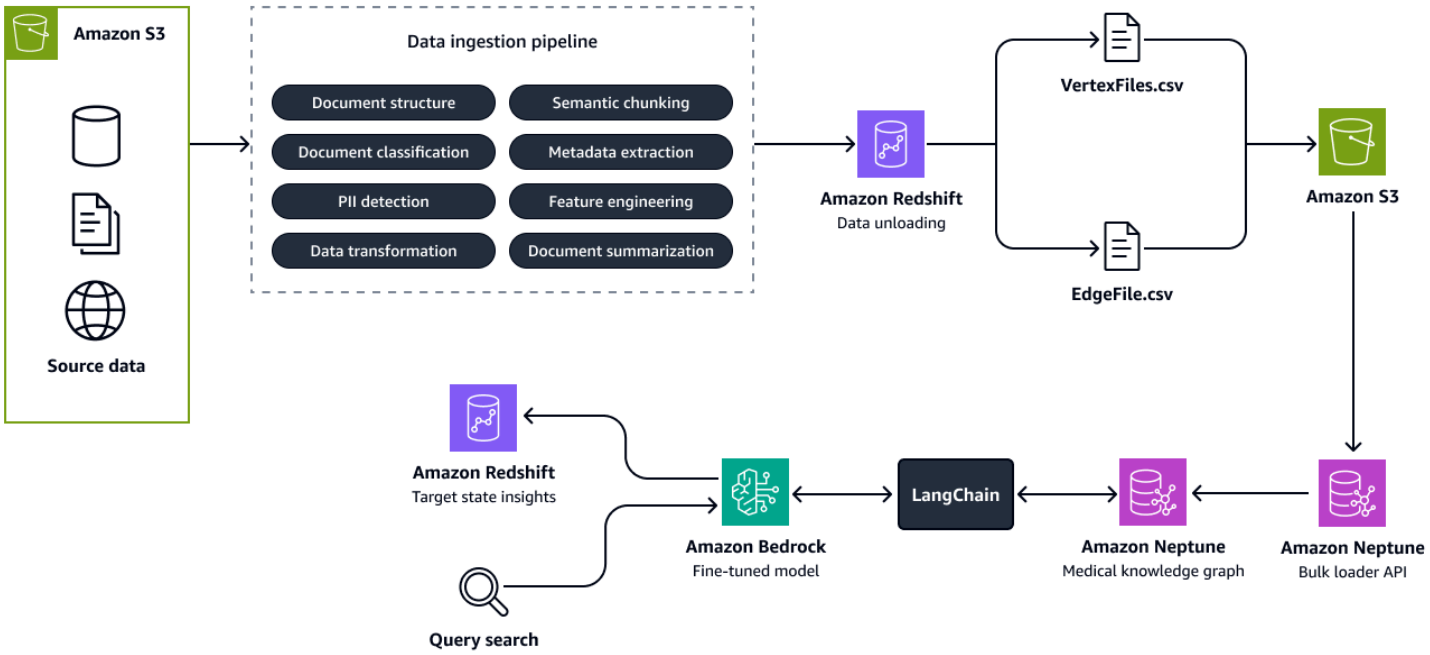
Als Nächstes entwickeln Sie eine Graph Retrieval Augmented Generation-Anwendung (Graph RAG). Graph RAG ist ein RAG-Ansatz, der Daten aus einer Graphdatenbank abruft. Im Folgenden sind die Komponenten der Graph RAG-Anwendung aufgeführt:

- Integration mit einem LLM in Amazon Bedrock — Die Anwendung verwendet ein LLM in Amazon Bedrock für das Verständnis natürlicher Sprache und die Generierung von Abfragen. Benutzer

können mithilfe natürlicher Sprache mit dem System interagieren. Dadurch ist es auch für technisch nicht versierte Akteure zugänglich.

- Orchestrierung und Informationsabruf — Verwendung oder LlamaIndexLangChainOrchestratoren, um die Integration zwischen dem LLM und dem Neptune Knowledge Graph zu erleichtern. Sie verwalten den Prozess der Konvertierung von Abfragen in natürlicher Sprache in OpenCypher-Abfragen. Anschließend führen sie die Abfragen im Knowledge Graph aus. Verwenden Sie Prompt Engineering, um das LLM über bewährte Methoden für die Erstellung von OpenCypher-Abfragen zu informieren. Dies hilft bei der Optimierung der Abfragen, um den entsprechenden Untergraphen abzurufen, der alle relevanten Entitäten und Beziehungen zu den abgefragten Rollen und Fähigkeiten enthält.
- Generierung von Erkenntnissen — Das LLM in Amazon Bedrock verarbeitet die abgerufenen Grafikdaten. Es generiert detaillierte Einblicke in den aktuellen Status und prognostiziert future Zustände für die abgefragte Rolle und die damit verbundenen Fähigkeiten.

Die folgende Abbildung zeigt die Schritte zum Erstellen eines Wissensgraphen aus Quelldaten. Sie übergeben die strukturierten und unstrukturierten Quelldaten an die Datenerfassungspipeline. Die Pipeline extrahiert Informationen und wandelt sie in eine CSV-Bulk-Load-Formation um, die mit Amazon Neptune kompatibel ist. Die Bulk-Loader-API lädt die CSV-Dateien, die in einem Amazon S3 S3-Bucket gespeichert sind, in den Neptune Knowledge Graph hoch. Bei Benutzeranfragen in Bezug auf den future Status von Talenten, relevanten Rollen oder Fähigkeiten interagiert das fein abgestimmte LLM in Amazon Bedrock mit dem Knowledge Graph über eine LangChain Orchestrator. Der Orchestrator ruft den relevanten Kontext aus dem Knowledge Graph ab und überträgt die Antworten in die Insights-Tabelle in Amazon Redshift. Das Tool LangChain Der Orchestrator konvertiert wie Graph QAChain die natürliche Sprachabfrage des Benutzers in eine OpenCypher-Abfrage, um den Knowledge Graph abzufragen. Das fein abgestimmte Modell von Amazon Bedrock generiert eine Antwort auf der Grundlage des abgerufenen Kontextes.



Schritt 3: Identifizierung von Qualifikationslücken und Empfehlung von Schulungen

In diesem Schritt berechnen Sie genau die Nähe zwischen dem aktuellen Status eines medizinischen Fachpersonals und potenziellen future Funktionen in diesem Bundesstaat. Zu diesem Zweck führen Sie eine Qualifikationsaffinitätsanalyse durch, indem Sie die Fähigkeiten der Person mit der beruflichen Rolle vergleichen. In einer Amazon OpenSearch Service-Vektordatenbank speichern Sie Informationen zur Skill-Taxonomie und Skill-Metadaten, wie z. B. die Beschreibung der Fähigkeiten, den Fertigkeitstyp und die Skill-Cluster. Verwenden Sie ein Amazon Bedrock-Einbettungsmodell, z. B. Amazon Titan Text Embeddings-Modelle, um die identifizierte Schlüsselkompetenz in Vektoren einzubetten. Mithilfe einer Vektorsuche rufen Sie die Beschreibungen der Fähigkeiten im aktuellen Status und der Fähigkeiten im Zielstatus ab und führen eine Ontologieanalyse durch. Die Analyse liefert Näherungswerte zwischen den Qualifikationspaaren im aktuellen Bundesstaat und im Zielstaat. Für jedes Paar verwenden Sie die berechneten Ontologiewerte, um die Lücken in den Qualifikationsaffinitäten zu identifizieren. Anschließend empfehlen Sie den optimalen Weiterbildungsweg, den der Kandidat bei Rollenwechseln in Betracht ziehen kann.

Für jede Rolle beinhaltet die Empfehlung der richtigen Lerninhalte für die Weiterbildung oder Umschulung einen systematischen Ansatz, der mit der Erstellung eines umfassenden Katalogs von Lerninhalten beginnt. Dieser Katalog, den Sie in einer Amazon Redshift Redshift-Datenbank speichern, fasst Inhalte verschiedener Anbieter zusammen und enthält Metadaten wie Inhaltsdauer,

Schwierigkeitsgrad und Lernmodus. Der nächste Schritt besteht darin, die in den einzelnen Inhalten enthaltenen Schlüsselkompetenzen zu extrahieren und sie dann den individuellen Fähigkeiten zuzuordnen, die für die Zielrolle erforderlich sind. Sie erreichen diese Zuordnung, indem Sie die Reichweite der Inhalte anhand einer Analyse der Nähe zu den Fähigkeiten analysieren. Bei dieser Analyse wird bewertet, inwieweit die im Inhalt vermittelten Fähigkeiten mit den für die Rolle angestrebten Fähigkeiten übereinstimmen. Die Metadaten spielen eine entscheidende Rolle bei der Auswahl der für jede Fähigkeit am besten geeigneten Inhalte und stellen sicher, dass die Lernenden maßgeschneiderte Empfehlungen erhalten, die ihren Lernbedürfnissen entsprechen. Verwenden Sie es LLMs in Amazon Bedrock, um Fähigkeiten aus den Inhaltsmetadaten zu extrahieren, Feature-Engineering durchzuführen und die Inhaltsempfehlungen zu validieren. Dies verbessert die Genauigkeit und Relevanz des Weiterbildungs- oder Umschulungsprozesses.

Ausrichtung auf das AWS Well-Architected Framework

Die Lösung entspricht allen sechs Säulen des AWS Well-Architected Framework:

- Operative Exzellenz — Eine modulare, automatisierte Pipeline verbessert die betriebliche Exzellenz. Die wichtigsten Komponenten der Pipeline sind entkoppelt und automatisiert, was schnellere Modellaktualisierungen und eine einfachere Überwachung ermöglicht. Darüber hinaus unterstützen automatisierte Trainingspipelines eine schnellere Veröffentlichung von fein abgestimmten Modellen.
- Sicherheit — Diese Lösung verarbeitet sensible und persönlich identifizierbare Informationen (PII), wie z. B. Daten in Lebensläufen und Talentprofilen. Implementieren Sie in AWS Identity and Access Management (IAM) detaillierte Richtlinien zur Zugriffskontrolle und stellen Sie sicher, dass nur autorisiertes Personal Zugriff auf diese Daten hat.
- Zuverlässigkeit — Die Lösung verwendet Neptune AWS-Services, Amazon Bedrock und OpenSearch Service, die Fehlertoleranz, hohe Verfügbarkeit und ununterbrochenen Zugriff auf Erkenntnisse auch bei hoher Nachfrage bieten.
- Leistungseffizienz — Die LLMs in Amazon Bedrock und OpenSearch Service fein abgestimmten Vektordatenbanken sind darauf ausgelegt, große Datenmengen schnell und präzise zu verarbeiten, um zeitnahe, personalisierte Lernempfehlungen zu liefern.
- Kostenoptimierung — Diese Lösung verwendet einen RAG-Ansatz, der den Bedarf an kontinuierlichem Vortraining von Modellen reduziert. Anstatt das gesamte Modell wiederholt zu verfeinern, optimiert das System nur bestimmte Prozesse, wie das Extrahieren von Informationen aus Lebensläufen und die Strukturierung der Ergebnisse. Dies führt zu erheblichen Kosteneinsparungen. Durch die Minimierung der Häufigkeit und des Umfangs ressourcenintensiver

Modellschulungen und durch die Nutzung von pay-per-use Cloud-Diensten können Organisationen im Gesundheitswesen ihre Betriebskosten optimieren und gleichzeitig eine hohe Leistung aufrechterhalten.

- Nachhaltigkeit — Diese Lösung nutzt skalierbare, Cloud-native Dienste, die Rechenressourcen dynamisch zuweisen. Dies reduziert den Energieverbrauch und die Umweltbelastung und unterstützt gleichzeitig groß angelegte, datenintensive Initiativen zur Talenttransformation.

Entwicklung und Orchestrierung generativer KI-Lösungen für das Gesundheitswesen

Um die Lösungen in diesem Leitfaden zu entwickeln, müssen Sie eine RAG-Architektur aufbauen, die mithilfe von Feineinstellungen erweiterte LLMs Patientendaten, klinische und diagnostische Erkenntnisse und prognostizierte Patientenergebnisse für Gesundheitsdienstleister bereitstellt. Dies erfordert die Integration mehrerer Tools, um einen kohärenten AWS-Services und effizienten Arbeitsablauf zu schaffen. In diesem Abschnitt wird Folgendes behandelt:

- Amazon Q Developer— Verwenden Sie Amazon Q Developer, um technische Fragen und Codefehler während des Entwicklungsprozesses zu lösen.
- RAG-Design mit mehreren Retrievern— Entwerfen und implementieren Sie RAG-Lösungen, die mehrere Retriever verwenden, um den richtigen medizinischen Kontext für die Frage des Benutzers abzurufen.
- ReAct Agenten— Implementieren Sie Agenten, die Argumentation mit dynamischem Handeln kombinieren.

Amazon Q Developer

Beim Aufbau einer generativen KI-Lösung kann es schwierig sein, KI-Agenten zu erstellen und die wichtigsten Dienste miteinander zu verbinden. Amazon Q Developer hilft Datenwissenschaftlern und KI-Ingenieuren jedoch, indem es Zugriff auf einen fortschrittlichen generativen KI-Assistenten bietet. Amazon Q kann schnell und präzise auf Benutzerfragen und Codefehler eingehen, was Ihnen helfen kann, den LLM-Entwicklungsprozess zu optimieren. Amazon Q bietet Entwicklern, die Anwendungen erstellen, die Amazon Bedrock Foundation-Modelle verwenden, erhebliche Vorteile. Es kann Arbeitsabläufe rationalisieren und die Codequalität verbessern. Es automatisiert die Generierung von Python-Skripten und Infrastructure-as-Code-Konfigurationen (IaC) und reduziert so die Entwicklungszeit und den Entwicklungsaufwand erheblich. Durch erweiterte Refactoring-Funktionen kann Amazon Q die Codeleistung verbessern, Sicherheitslücken identifizieren und sicherstellen, dass Entwickler sich an bewährte Methoden halten. Darüber hinaus erleichtert es Anfängern das Lernen und die Einführung, indem es kontextsensitive Vorschläge und Erklärungen bietet, wodurch komplexe Codierungsaufgaben leichter zugänglich und effizienter werden.

RAG-Design mit mehreren Retrievern

In einer generativen KI-Anwendung kann eine Multi-Retriever-RAG-Pipeline effizient Informationen aus mehreren Datenquellen abrufen, um Gesundheitsdienstleistern und Klinikern bei der Beantwortung medizinischer Fragen zu helfen. Diese Pipeline verwendet verschiedene Arten von Retrievern, um relevante Daten aus verschiedenen Wissensdatenbanken abzurufen. Jeder Retriever ist darauf spezialisiert, eine bestimmte Art von Informationen abzurufen, z. B. Patientenanamnese, diagnostische Erkenntnisse, klinische Notizen oder Inhalte aus medizinischer Forschung und akademischen Texten.

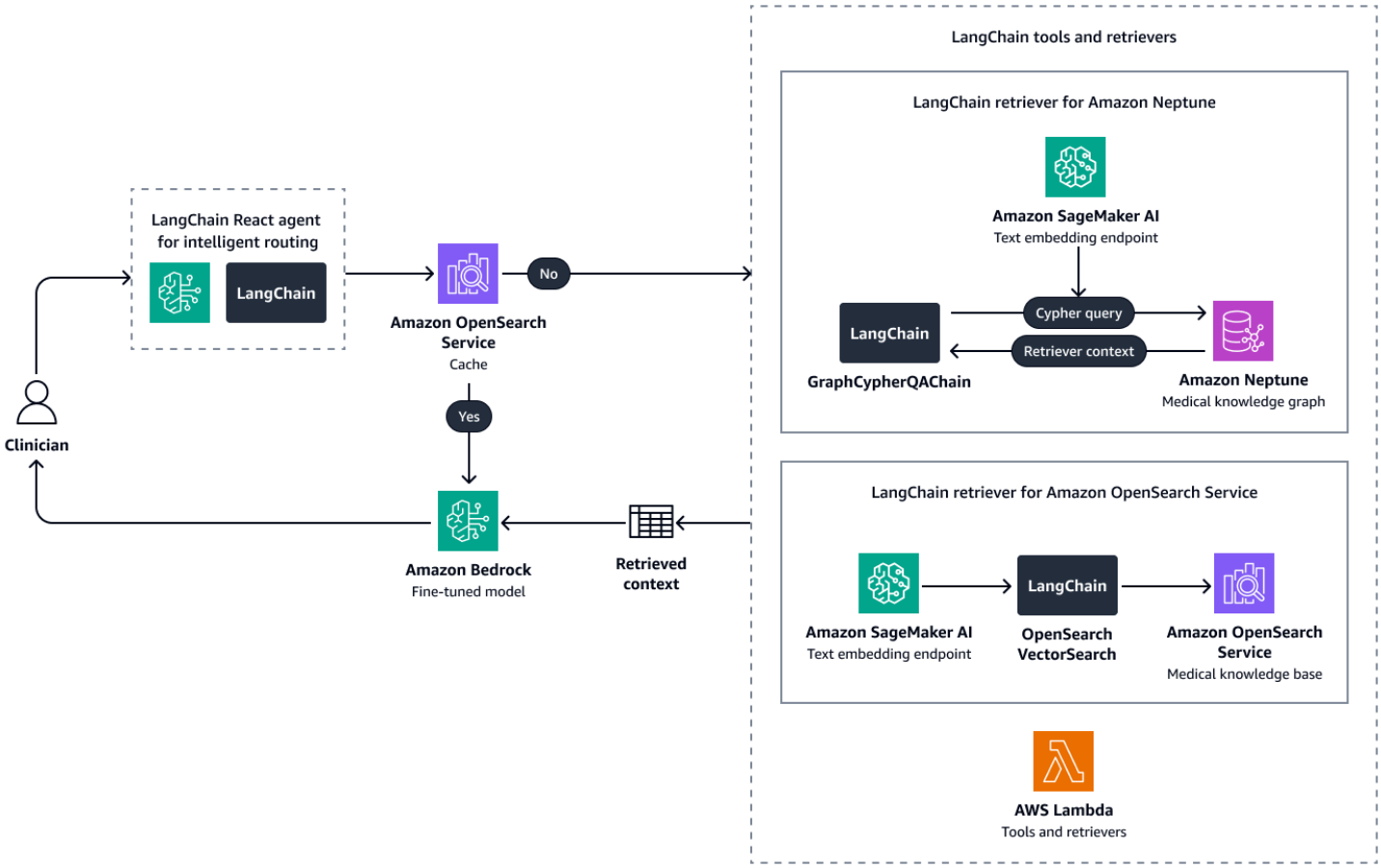
Ermitteln Sie anhand der Art der Daten und der spezifischen Anwendungsanforderungen, welche Backend-Wissensdatenbank für Ihren Anwendungsfall geeignet ist. Eine Amazon OpenSearch Service-Vektordatenbank eignet sich gut für große Mengen unstrukturierter oder halbstrukturierter Gesundheitsdaten, einschließlich Zusammenfassungen von Bilddiagnosen, Entlassungszusammenfassungen, klinischen Berichten, medizinischen Forschungsergebnissen und akademischen Textinhalten. Andererseits kann ein Graphdatenbank-Service wie Amazon Neptune ideal für Anwendungsfälle im Gesundheitswesen sein, die eine gründliche Untersuchung der zeitlichen Beziehungen zwischen Entitäten erfordern, wie z. B. Patient, Patientengeschichte, Gesundheitsdienstleister, Medikamente, Symptome und Behandlungen.

Eine wichtige Komponente dieser Pipeline ist die Vorhersage der Absicht von Benutzeranfragen. Dadurch wird sichergestellt, dass das System die Abfrage an die richtige Retrieverkette weiterleitet. Wenn ein Arzt beispielsweise nach der Behandlungsgeschichte, den Symptomen, der Interaktion mit dem Krankenhaus, der Wahrscheinlichkeit einer erneuten Aufnahme in das Krankenhaus oder möglichen Behandlungsergebnissen eines Patienten fragt, identifiziert das Modul zur Vorhersage der Abfrageabsicht diese Absicht. Es leitet die Anfrage an die Retriever-Kette weiter, die Patientenakten oder chronologische Behandlungsdaten aus dem Medical Knowledge Graph abrufen kann. Wenn es sich bei der Frage um die Entdeckung von Krankheiten, spezifische diagnostische Beurteilungen oder Einzelheiten bestimmter klinischer Verfahren aus akademischen Lehrbüchern handelt, wird die Anfrage alternativ an die Retriever-Kette weitergeleitet, die diese Informationen aus der Service-Vektordatenbank abrufen kann. OpenSearch Sie können die Funktionen zum Aufrufen von Tools verwenden LangChain um ein benutzerdefiniertes Tool an das Amazon Bedrock LLM zu binden, das eine Benutzerfrage in vordefinierte Absichten klassifizieren kann.

Dieses Multi-Retriever-RAG-System umfasst LangChain Agenten, die für die Verwaltung des Zugriffs auf die jeweilige Wissensdatenbank konzipiert sind. Sie können Folgendes verwenden … LangChain um die Interaktion zwischen dem Amazon Bedrock LLM, den verschiedenen Retrievern und Tools

zu orchestrieren. LangChain enthält eine Klasse zum Aufrufen von Tools, mit deren Hilfe Sie benutzerdefinierte Tools erstellen können, z. B. einen Intent Classifier, einen Retriever für Neptune, einen Retriever für OpenSearch Service oder jedes andere Tool, das entwickelt werden kann, um die Benutzerabsicht zu klassifizieren und auf Daten aus einer bestimmten Wissensdatenbank in einem strukturierten Format zuzugreifen. Anschließend geben Sie diese Tools an die Klasse weiter, um einen Agenten für Reasoning and Acting () zu erstellen. ReAct Der ReAct Agent verarbeitet die Benutzerfrage, plant die sequentiellen Schritte zur Beantwortung der Frage und führt dann iterativ die verfügbaren Tools aus und verarbeitet die Antworten der Tools, um schließlich die Benutzeranfrage zu beantworten.

Die folgende Abbildung zeigt, wie ein Multi-Retriever-RAG-System funktioniert, das für effizienten Wissensabruf und intelligente Abfrageauflösung konzipiert ist. A LangChain ReAct Der Agent analysiert die Absicht des Benutzers, formuliert einen strukturierten Ausführungsplan und wählt die relevantesten Tools zum Abrufen aus. Das System fragt einen Cache für frühere Fragen ab und sucht anhand von Schlüsselattributen wie Patienten-ID, Gesundheitszustand und Besuchsdatum nach ähnlichen Abfragen. Wenn eine sehr ähnliche Frage gefunden wird, wird die entsprechende Antwort direkt abgerufen. Andernfalls führt der Agent den entsprechenden Retriever aus. Für den Abruf patientenzentrierter Informationen wie Behandlungsanamnese, Symptome, Krankenhausinteraktionen oder Wahrscheinlichkeit einer erneuten Aufnahme verwendet das System einen Graph Retriever. Für diagnostische Untersuchungen, klinische Verfahren und strukturierte medizinische Befunde verwendet der Agent einen Vektor-Datenbankabruf. In Szenarien, die eine Kombination von Kontextwissen aus beiden Datenspeichern erfordern, um eine umfassende Antwort zu generieren, verwendet das System eine hybride Abrufstrategie, die Ergebnisse sowohl aus dem Wissensgraphen als auch aus der Vektordatenbank integriert.



ReAct Agenten

Die Agenten Reasoning and Acting (ReAct) sind für vielfältige RAG-Anwendungen konzipiert. Diese Agenten bieten eine leistungsstarke Kombination aus Argumentation und dynamischem Handeln, insbesondere für komplexe Anwendungen, die logische Workflows zum Abrufen von step-by-step Informationen beinhalten. Weitere Informationen finden Sie unter ReActSynergisierung von Argumentation und Handeln in Sprachmodellen.

In medizinischen und medizinischen Kontexten sind die Anfragen eines Klinikers oder Arztes oft vielschichtig. Ein Kliniker könnte beispielsweise fragen: „Welche Behandlungen wurden ähnlichen Patienten mit Bluthochdruck und Typ-2-Diabetes verabreicht?“ Nach der Identifizierung der Benutzerabsicht, die darin besteht, die Behandlungen für Bluthochdruck und Typ-2-Diabetes abzurufen, muss der KI-Agent diese Abfrage in Unteraufgaben unterteilen und dann die effizienteste Abrufstrategie auswählen. In diesem Fall sollte der KI-Agent die relevantesten Knoten (wie Alter, Geschlecht, Erkrankungen, Behandlungen und Medikamente) des Patienten identifizieren und dann das Diagramm nach diesen Entitäten und ihren Attributen und Beziehungen abfragen. ReAct Agenten

sind sehr hilfreich, da sie die Fähigkeit eines LLM zum Denken (logische Folgerungen) mit einer Aktion (Abfragen oder Interaktion mit externen Ressourcen oder Wissensdatenbanken) kombinieren.

Um die Benutzerfrage „Welche Behandlungen wurden ähnlichen Patienten mit Bluthochdruck und Typ-2-Diabetes verabreicht?“ zu beantworten , das folgende Beispiel veranschaulicht, wie ein ReAct Agent funktioniert:

1. Argumentation des Agenten — Der ReAct Mitarbeiter schließt daraus, dass es sich bei der Frage um das Abrufen von Informationen über Erkrankungen (Diabetes und Bluthochdruck) handelt. Dabei werden das Alter des Patienten, die Behandlungen, die Medikamente und der zu analysierende Zeitraum berücksichtigt.
2. Aktion des Agenten — Der Agent verwendet OpenCypher, um den Wissensgraphen nach Behandlungen abzufragen, die spezifisch für Typ-2-Diabetes und Bluthochdruck sind. Außerdem werden verabreichte Medikamente, Daten von Krankenhausbesuchen, Nebenwirkungen von Medikamenten, bekannte Behandlungsergebnisse und Querverweisdaten für ähnliche Patienten (z. B. Patienten gleichen Geschlechts und Alters) abgerufen.
3. Beobachtung der Arzneimittelwirkstoffe — Aus dem Knowledge Graph ruft der Agent tabellarische Daten der letzten sechs Monate über Behandlungen von Patienten ab, die sowohl an Bluthochdruck als auch an Typ-2-Diabetes leiden.
4. Argumentation des Behandlers — Um die Ergebnisse der abgerufenen Datensätze in eine Rangfolge einzuordnen, identifiziert der Experte wichtige Merkmale, wie z. B. die Behandlungsdauer, Nebenwirkungen von Medikamenten oder bekannte Behandlungsergebnisse.
5. Aktion des Agenten — Der Agent ordnet die Datensätze anhand von identifizierten Attributen und vordefinierter Logik, die ihm durch die Systemaufforderung vermittelt wird, neu.
6. Generierung von Antworten — Das LLM in Amazon Bedrock generiert eine Antwort, die auf dem Kontext basiert, den der ReAct Agent vorbereitet hat.

Evaluierung generativer KI-Lösungen für das Gesundheitswesen

Die Bewertung der von Ihnen entwickelten KI-Lösungen für das Gesundheitswesen ist entscheidend, um sicherzustellen, dass sie in realen medizinischen Umgebungen effektiv, zuverlässig und skalierbar sind. Verwenden Sie einen systematischen Ansatz, um die Leistung der einzelnen Komponenten der Lösung zu bewerten. Im Folgenden finden Sie eine Zusammenfassung der Methoden und Kennzahlen, die Sie zur Bewertung Ihrer Lösung verwenden können.

Themen

Bewertung der Extraktion von Informationen

Evaluieren Sie die Leistung von Informationsextraktionslösungen wie dem intelligenten Resume-Parser und dem benutzerdefinierten Entitäten-Extraktor. Sie können die Ausrichtung der Antworten dieser Lösungen anhand eines Testdatensatzes messen. Wenn Sie nicht über einen Datensatz verfügen, der vielseitige Talentprofile im Gesundheitswesen und Patientenakten abdeckt, können Sie mithilfe der Argumentationsfähigkeit eines LLM einen benutzerdefinierten Testdatensatz erstellen. Sie könnten beispielsweise ein Modell mit großen Parametern verwenden, wie Anthropic Claude Modelle, um einen Testdatensatz zu generieren.

Im Folgenden sind drei wichtige Kennzahlen aufgeführt, die Sie für die Bewertung der Modelle zur Informationsextraktion verwenden können:

- Genauigkeit und Vollständigkeit — Mit diesen Kennzahlen wird bewertet, inwieweit die Ergebnisse die korrekten und vollständigen Informationen aus den Ground-Truth-Daten erfasst haben. Dabei wird sowohl die Richtigkeit der extrahierten Informationen als auch das Vorhandensein aller relevanten Details in den extrahierten Informationen überprüft.
- Ähnlichkeit und Relevanz — Mit diesen Metriken werden die semantischen, strukturellen und kontextuellen Ähnlichkeiten zwischen den Ergebnissen und den Ground-Truth-Daten (die Ähnlichkeit) sowie der Grad bewertet, in dem das Ergebnis mit dem Inhalt, dem Kontext und der Absicht der Ground-Truth-Daten übereinstimmt und diese berücksichtigt (die Relevanz).

- Angepasste Erinnerungs- oder Erfassungsrate — Diese Raten bestimmen empirisch, wie viele der aktuellen Werte in den Ground-Truth-Daten vom Modell korrekt identifiziert wurden. Die Rate sollte eine Strafe für alle falschen Werte beinhalten, die das Modell extrahiert.
- Genauigkeitswert — Mithilfe des Präzisionswerts können Sie ermitteln, wie viele falsch positive Ergebnisse in den Prognosen im Vergleich zu den echten positiven Ergebnissen enthalten sind. Sie können beispielsweise Präzisionskennzahlen verwenden, um die Richtigkeit der extrahierten Fertigkeiten zu messen.

Evaluierung von RAG-Lösungen mit mehreren Retrievern

Um zu beurteilen, wie gut das System relevante Informationen abruft und wie effektiv es diese Informationen verwendet, um genaue und kontextbezogene Antworten zu generieren, können Sie die folgenden Kennzahlen verwenden:

- Relevanz der Antwort — Messen Sie, wie relevant die generierte Antwort, die den abgerufenen Kontext verwendet, für die ursprüngliche Abfrage ist.
- Kontextgenauigkeit — Beurteilen Sie anhand der insgesamt abgerufenen Ergebnisse den Anteil der abgerufenen Dokumente oder Textfragmente, die für die Anfrage relevant sind. Eine höhere Kontextgenauigkeit weist darauf hin, dass der Abrufmechanismus bei der Auswahl relevanter Informationen wirksam ist.
- Zuverlässigkeit — Beurteilt, wie genau die generierte Antwort die Informationen im abgerufenen Kontext widerspiegelt. Mit anderen Worten, messen Sie, ob die Antwort den Quellinformationen entspricht.

Evaluierung einer Lösung mithilfe eines LLM

Sie können eine Technik namens LLM- verwendenas-a-judge, um die Textantworten Ihrer generativen KI-Lösung auszuwerten. Es beinhaltet die Verwendung LLMs zur Bewertung und Bewertung der Leistung von Modellergebnissen. Diese Technik nutzt die Funktionen von Amazon Bedrock, um Urteile zu verschiedenen Attributen wie Antwortqualität, Kohärenz, Einhaltung, Genauigkeit und Vollständigkeit menschlicher Präferenzen oder Ground-Truth-Daten abzugeben. Für eine umfassende Bewertung verwenden Sie chain-of-thought Techniken (CoT) und wenige Eingabeaufforderungen. Die Aufforderung weist das LLM an, die generierte Antwort anhand einer Bewertungsrubrik zu bewerten, und die wenigen Stichproben in der Eingabeaufforderung veranschaulichen den tatsächlichen Bewertungsprozess. Die Aufforderung enthält auch Richtlinien,

die der LLM-Evaluator befolgen muss. Sie könnten beispielsweise erwägen, eine oder mehrere der folgenden Bewertungstechniken zu verwenden, bei denen ein LLM zur Beurteilung der generierten Antworten verwendet wird:

- Paarweiser Vergleich — Geben Sie dem LLM-Gutachter eine medizinische Frage und mehrere Antworten, die durch verschiedene, iterative Versionen der von Ihnen erstellten RAG-Systeme generiert wurden. Bitten Sie den LLM-Evaluator, die beste Antwort auf der Grundlage von Antwortqualität, Kohärenz und Übereinstimmung mit der ursprünglichen Frage zu ermitteln.
- Einstufung mit einer einzigen Antwort — Diese Technik eignet sich gut für Anwendungsfälle, in denen Sie die Genauigkeit der Kategorisierung bewerten müssen, z. B. bei der Klassifizierung von Behandlungsergebnissen, der Kategorisierung des Patientenverhaltens, der Wahrscheinlichkeit einer erneuten Aufnahme von Patienten und der Risikokategorisierung. Verwenden Sie den LLM-Evaluator, um die individuelle Kategorisierung oder Klassifikation isoliert zu analysieren und die darin enthaltenen Argumentation anhand von Ground-Truth-Daten zu bewerten.
- Benotung anhand von Referenzen — Stellen Sie dem LLM-Gutachter eine Reihe von medizinischen Fragen zur Verfügung, die aussagekräftige Antworten erfordern. Erstellen Sie Beispielantworten auf diese Fragen, z. B. Referenzantworten oder ideale Antworten. Bitten Sie den LLM-Evaluator, die vom LLM generierte Antwort mit den Referenzantworten oder idealen Antworten zu vergleichen, und fordern Sie den LLM-Evaluator auf, die generierte Antwort auf Richtigkeit, Vollständigkeit, Ähnlichkeit, Relevanz oder andere Merkmale zu bewerten. Mit dieser Technik können Sie beurteilen, ob die generierten Antworten einer klar definierten Standardantwort oder einer beispielhaften Antwort entsprechen.

Ressourcen

AWS Dokumentation

- Dokumentation zu Amazon Bedrock
- Dokumentation zu Amazon Neptune
- Amazon OpenSearch Service-Dokumentation
- Anwendung des AWS Well-Architected Frameworks für Amazon Neptune (Prescriptive Guidance)AWS
- Best Practices für den Betrieb von Amazon OpenSearch Service (OpenSearch Servicedokumentation)
- Verwendung von Amazon Comprehend Medical und LLMs für das Gesundheitswesen und die Biowissenschaften (AWS Prescriptive Guidance)

AWS Blog-Beiträge

- Erstellen Sie RAG- und agentenbasierte generative KI-Anwendungen mit dem neuen Amazon Titan Text Premier-Modell, das in Amazon Bedrock verfügbar ist
- Ergänzen Sie Commercial Intelligence, indem Sie mit Amazon Neptune einen Knowledge Graph aus einem Data Warehouse erstellen
- Verwendung von Wissensgraphen zur Erstellung von GraphRag-Anwendungen mit Amazon Bedrock und Amazon Neptune

Sonstige Ressourcen

- Integration von Retrieval-Augmented Generation mit großen Sprachmodellen in der Nephrologie: Förderung praktischer Anwendungen (Central, National Library of Medicine) PubMed
- Einführung in LangChain (LangChain Dokumentation)

Mitwirkende

Inhaltserstellung

- Nitu Nivedita, Geschäftsführer — Leiter für künstliche Intelligenz, Daten und KI, Accenture
- Manoj Appully, Gründer und CTO von Cadiem
- Conor Folan, Berater für Daten und KI, Accenture
- Deepak Krishna AR, Berater — Daten und KI, Accenture
- Almore Cato, Manager — Daten und KI, Accenture
- Soonam Kurian, leitender Lösungsarchitekt, AWS

Überprüfend

- Sally Lin, Senior Managerin für Datenwissenschaft — Daten und KI, Accenture
- Terry Huang, Manager für Datenwissenschaft — Daten und KI, Accenture
- William Lorenz, Lösungsarchitekt bei Partners, AWS

Technisches Schreiben

- Lilly AbouHarb, leitende technische Redakteurin, AWS

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	—	14. März 2025

AWS Glossar zu präskriptiven Leitlinien

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

Zahlen

7 Rs

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

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

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

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

A

ABAC

Siehe attributbasierte Zugriffskontrolle.

abstrahierte Dienste

Weitere Informationen finden Sie unter Managed Services.

ACID

Siehe Atomarität, Konsistenz, Isolierung und Haltbarkeit.

Aktiv-Aktiv-Migration

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

Aktiv-Passiv-Migration

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

Aggregatfunktion

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

AI

Siehe künstliche Intelligenz.

AIOps

Siehe Operationen im Bereich künstliche Intelligenz.

Anonymisierung

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

Anti-Muster

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

Anwendungssteuerung

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

Anwendungsportfolio

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

künstliche Intelligenz (KI)

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

Operationen mit künstlicher Intelligenz (AIOps)

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

Asymmetrische Verschlüsselung

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

Atomizität, Konsistenz, Isolierung, Haltbarkeit (ACID)

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

Attributbasierte Zugriffskontrolle (ABAC)

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

autoritative Datenquelle

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

Availability Zone

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

AWS Framework für die Cloud-Einführung (AWS CAF)

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

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

AWS Workload-Qualifizierungsrahmen (AWS WQF)

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

B

schlechter Bot

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

BCP

Siehe Planung der Geschäftskontinuität.

Verhaltensdiagramm

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

Big-Endian-System

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

Binäre Klassifikation

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

Bloom-Filter

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

Blau/Grün-Bereitstellung

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

Bot

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

Botnetz

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

branch

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

Zugang durch Glasbruch

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

Brownfield-Strategie

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

Puffer-Cache

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

Geschäftsfähigkeit

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

Planung der Geschäftskontinuität (BCP)

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

C

CAF

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

Bereitstellung auf Kanaren

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

CCoE

Weitere Informationen finden Sie im Cloud Center of Excellence.

CDC

Siehe Erfassung von Änderungsdaten.

Erfassung von Datenänderungen (CDC)

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

Chaos-Technik

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

CI/CD

Siehe Continuous Integration und Continuous Delivery.

Klassifizierung

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

clientseitige Verschlüsselung

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

Cloud-Exzellenzzentrum (CCoE)

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

Cloud Computing

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

Cloud-Betriebsmodell

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

Phasen der Einführung der Cloud

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

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

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

CMDB

Siehe Datenbank für das Konfigurationsmanagement.

Code-Repository

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

Kalter Cache

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

Kalte Daten

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

Computer Vision (CV)

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

Drift in der Konfiguration

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

Verwaltung der Datenbankkonfiguration (CMDB)

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

Konformitätspaket

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

Kontinuierliche Bereitstellung und kontinuierliche Integration (CI/CD)

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

CV

Siehe Computer Vision.

D

Daten im Ruhezustand

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

Datenklassifizierung

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

Datendrift

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

Daten während der Übertragung

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

Datennetz

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

Datenminimierung

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

Datenperimeter

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

Vorverarbeitung der Daten

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

Herkunft der Daten

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

betroffene Person

Eine Person, deren Daten gesammelt und verarbeitet werden.

Data Warehouse

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

Datenbankdefinitionssprache (DDL)

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

Datenbankmanipulationssprache (DML)

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

DDL

Siehe Datenbankdefinitionssprache.

Deep-Ensemble

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

Deep Learning

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

defense-in-depth

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

delegierter Administrator

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

Bereitstellung

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

Entwicklungsumgebung

Siehe Umgebung.

Detektivische Kontrolle

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

Abbildung des Wertstroms in der Entwicklung (DVSM)

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

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

digitaler Zwilling

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

Maßtabelle

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

Katastrophe

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

Disaster Recovery (DR)

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

DML

Siehe Sprache zur Datenbankmanipulation.

Domainorientiertes Design

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

DR

Siehe Disaster Recovery.

Erkennung von Driften

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

DVSM

Siehe Abbildung des Wertstroms in der Entwicklung.

E

EDA

Siehe explorative Datenanalyse.

EDI

Siehe elektronischer Datenaustausch.

Edge-Computing

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

elektronischer Datenaustausch (EDI)

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

Verschlüsselung

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

Verschlüsselungsschlüssel

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

Endianismus

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

Endpunkt

Siehe Service-Endpunkt.

Endpunkt-Services

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

Unternehmensressourcenplanung (ERP)

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

Envelope-Verschlüsselung

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

Umgebung

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

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

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

Epics

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

ERP

Siehe Enterprise Resource Planning.

Explorative Datenanalyse (EDA)

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

F

Faktentabelle

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

schnell scheitern

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

Grenze zur Fehlerisolierung

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

Feature-Zweig

Siehe Zweig.

Features

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

Bedeutung der Feature

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

Featuretransformation

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

Eingabeaufforderung mit wenigen Klicks

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

FGAC

Siehe detaillierte Zugriffskontrolle.

Feinkörnige Zugriffskontrolle (FGAC)

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

Flash-Cut-Migration

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

FM

Siehe Fundamentmodell.

Fundamentmodell (FM)

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

G

generative KI

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

Geoblocking

Siehe geografische Einschränkungen.

Geografische Einschränkungen (Geoblocking)

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

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

Gitflow-Workflow

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

goldenes Bild

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

Greenfield-Strategie

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

Integritätsschutz

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

H

HEKTAR

Siehe Hochverfügbarkeit.

Heterogene Datenbankmigration

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

hohe Verfügbarkeit (HA)

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

historische Modernisierung

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

Holdout-Daten

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

Homogene Datenbankmigration

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

heiße Daten

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

Hotfix

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

Hypercare-Phase

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

I

IaC

Sehen Sie sich Infrastruktur als Code an.

Identitätsbasierte Richtlinie

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

Leerlaufanwendung

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

IIoT

Siehe Industrielles Internet der Dinge.

unveränderliche Infrastruktur

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

Eingehende (ingress) VPC

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

Inkrementelle Migration

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

Industrie 4.0

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

Infrastruktur

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

Infrastructure as Code (IaC)

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

industrielles Internet der Dinge (T) IIo

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

Inspektions-VPC

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

Internet of Things (IoT)

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

Interpretierbarkeit

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

IoT

Siehe Internet der Dinge.

IT information library (ITIL, IT-Informationsbibliothek)

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

T service management (ITSM, IT-Servicemanagement)

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

BIS

Weitere Informationen finden Sie in der IT-Informationsbibliothek.

ITSM

Siehe IT-Servicemanagement.

L

Labelbasierte Zugangskontrolle (LBAC)

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

Landing Zone

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

großes Sprachmodell (LLM)

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

Große Migration

: Eine Migration von 300 oder mehr Servern.

SCHWARZ

: Weitere Informationen finden Sie unter Label-basierte Zugriffskontrolle.

Geringste Berechtigung

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

Lift and Shift

: Siehe 7 Rs.

Little-Endian-System

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

LLM

Siehe großes Sprachmodell.

Niedrigere Umgebungen

Siehe Umgebung.

M

Machine Learning (ML)

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

Hauptzweig

Siehe Filiale.

Malware

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

verwaltete Dienste

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

Manufacturing Execution System (MES)

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

MAP

Siehe Migration Acceleration Program.

Mechanismus

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

Mitgliedskonto

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

DURCHEINANDER

Siehe Manufacturing Execution System.

Message Queuing-Telemetrietransport (MQTT)

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

Microservice

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

Microservices-Architekturen

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

Migration Acceleration Program (MAP)

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

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

Migration in großem Maßstab

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

Migrationsfabrik

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

Migrationsmetadaten

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

Migrationsmuster

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

Migration Portfolio Assessment (MPA)

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

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

Migration Readiness Assessment (MRA)

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

Migrationsstrategie

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

ML

Siehe maschinelles Lernen.

Modernisierung

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

Bewertung der Modernisierungsfähigkeit

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

Monolithische Anwendungen (Monolithen)

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

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

MPA

Siehe Bewertung des Migrationsportfolios.

MQTT

Siehe Message Queuing-Telemetrietransport.

Mehrklassen-Klassifizierung

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

veränderbare Infrastruktur

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

O

OAC

Siehe Origin Access Control.

OAI

Siehe Zugriffsidentität von Origin.

COM

Siehe organisatorisches Change-Management.

Offline-Migration

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

OI

Siehe Betriebsintegration.

OLA

Siehe Vereinbarung auf operativer Ebene.

Online-Migration

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

OPC-UA

Siehe Open Process Communications — Unified Architecture.

Offene Prozesskommunikation — Einheitliche Architektur (OPC-UA)

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

Vereinbarung auf Betriebsebene (OLA)

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

Überprüfung der Betriebsbereitschaft (ORR)

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

Betriebstechnologie (OT)

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

Betriebsintegration (OI)

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

Organisationspfad

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

Organisatorisches Veränderungsmanagement (OCM)

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

Ursprungszugriffskontrolle (OAC)

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

Ursprungszugriffsidentität (OAI)

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

ORR

Weitere Informationen finden Sie unter Überprüfung der Betriebsbereitschaft.

NICHT

Siehe Betriebstechnologie.

Ausgehende (egress) VPC

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

P

Berechtigungsgrenze

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

persönlich identifizierbare Informationen (PII)

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

Personenbezogene Daten

Siehe persönlich identifizierbare Informationen.

Playbook

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

PLC

Siehe programmierbare Logiksteuerung.

PLM

Siehe Produktlebenszyklusmanagement.

policy

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

Polyglotte Beharrlichkeit

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

Portfoliobewertung

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

predicate

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

Prädikat Pushdown

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

Präventive Kontrolle

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

Prinzipal

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

Datenschutz von Natur aus

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

Privat gehostete Zonen

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

proaktive Steuerung

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

Produktlebenszyklusmanagement (PLM)

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

Produktionsumgebung

Siehe Umgebung.

Speicherprogrammierbare Steuerung (SPS)

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

schnelle Verkettung

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

Pseudonymisierung

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

publish/subscribe (pub/sub)

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

Q

Abfrageplan

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

Abfrageplanregression

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

R

RACI-Matrix

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

LAPPEN

Siehe Erweiterte Generierung beim Abrufen.

Ransomware

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

RASCI-Matrix

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

RCAC

Siehe Zugriffskontrolle für Zeilen und Spalten.

Read Replica

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

neu strukturieren

Siehe 7 Rs.

Recovery Point Objective (RPO)

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

Wiederherstellungszeitziel (RTO)

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

Refaktorierung

Siehe 7 Rs.

Region

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

Regression

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

rehosten

Siehe 7 Rs.

Veröffentlichung

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

umziehen

Siehe 7 Rs.

neue Plattform

Siehe 7 Rs.

Rückkauf

Siehe 7 Rs.

Ausfallsicherheit

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

Ressourcenbasierte Richtlinie

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

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

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

Reaktive Kontrolle

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

Beibehaltung

Siehe 7 Rs.

zurückziehen

Siehe 7 Rs.

Retrieval Augmented Generation (RAG)

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

Drehung

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

Zugriffskontrolle für Zeilen und Spalten (RCAC)

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

RPO

Siehe Recovery Point Objective.

RTO

Siehe Ziel der Wiederherstellungszeit.

Runbook

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

S

SAML 2.0

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

SCADA

Siehe Aufsichtskontrolle und Datenerfassung.

SCP

Siehe Richtlinie zur Dienstkontrolle.

Secret

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

Sicherheit durch Design

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

Sicherheitskontrolle

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

Härtung der Sicherheit

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

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

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

Automatisierung von Sicherheitsreaktionen

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

Serverseitige Verschlüsselung

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

Service-Kontrollrichtlinie (SCP)

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

Service-Endpunkt

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

Service Level Agreement (SLA)

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

Service-Level-Indikator (SLI)

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

Service-Level-Ziel (SLO)

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

Modell der geteilten Verantwortung

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

SIEM

Siehe Sicherheitsinformations- und Event-Management-System.

Single Point of Failure (SPOF)

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

SLA

Siehe Service Level Agreement.

SLI

Siehe Service-Level-Indikator.

ALSO

Siehe Service-Level-Ziel.

split-and-seed Modell

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

SPOTTEN

Siehe Single Point of Failure.

Sternschema

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

Strangler-Fig-Muster

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

Subnetz

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

Aufsichtskontrolle und Datenerfassung (SCADA)

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

Symmetrische Verschlüsselung

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

synthetisches Testen

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

Systemaufforderung

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

T

tags

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

Zielvariable

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

Aufgabenliste

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

Testumgebungen

Siehe Umgebung.

Training

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

Transit-Gateway

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

Stammbasierter Workflow

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

Vertrauenswürdiger Zugriff

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

Optimieren

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

Zwei-Pizzen-Team

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

U

Unsicherheit

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

undifferenzierte Aufgaben

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

höhere Umgebungen

Siehe Umgebung.

V

Vacuuming

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

Versionskontrolle

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

VPC-Peering

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

Schwachstelle

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

W

Warmer Cache

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

warme Daten

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

Fensterfunktion

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

Workload

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

Workstream

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

WURM

Sehen Sie einmal schreiben, viele lesen.

WQF

Siehe AWS Workload-Qualifizierungsrahmen.

einmal schreiben, viele lesen (WORM)

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

Z

Zero-Day-Exploit

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

Zero-Day-Sicherheitslücke

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

Zero-Shot-Aufforderung

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

Zombie-Anwendung

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

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