Verwendung von Amazon Comprehend Medical und LLMs für das Gesundheitswesen und die Biowissenschaften

AWS Präskriptive Leitlinien

AWS Präskriptive Leitlinien: Verwendung von Amazon Comprehend Medical und LLMs für das Gesundheitswesen und die Biowissenschaften

Table of Contents

Verwendung von Amazon Comprehend Medical und LLMs für das Gesundheitswesen und die Biowissenschaften

Amazon Web Services (???Mitwirkende)

Dezember 2025 (Verlauf der Dokumente)

-Übersicht

Das ständig wachsende Volumen an medizinischen Daten und der Bedarf an effizienter und genauer Verarbeitung haben die Einführung von natürlicher Sprachverarbeitung (NLP) mit Technologien für künstliche Intelligenz und maschinelles Lernen (KI/ML) vorangetrieben. Vortrainierte Klassifikatormodelle und umfangreiche Sprachmodelle (LLMs) haben sich als leistungsstarke Tools für verschiedene medizinische NLP-Aufgaben erwiesen, darunter die Beantwortung klinischer Fragen, die Zusammenfassung von Berichten und die Generierung von Erkenntnissen. Der Bereich Gesundheitswesen und Biowissenschaften stellt jedoch aufgrund der Komplexität der medizinischen Terminologie, des fachspezifischen Wissens und der regulatorischen Anforderungen besondere Herausforderungen dar. Die effektive Verwendung von vortrainierten Klassifikatoren oder LLMs in diesem Bereich erfordert einen gut durchdachten Ansatz, der die Stärken dieser Modelle mit domänenspezifischen Ressourcen und Techniken kombiniert.

Die Branchenpraktiken im Gesundheitswesen und in den Biowissenschaften stützten sich traditionell auf regelbasierte Systeme, manuelle Kodierung und Verfahren zur Überprüfung durch Experten. Diese Systeme und Prozesse sind zeitaufwändig und fehleranfällig. Die Integration von KI- und NLP-Technologien wie Amazon Comprehend Medical und den Foundation-Modellen in Amazon Bedrock bietet effiziente und skalierbare Lösungen für die Verarbeitung medizinischer Daten und verbessert gleichzeitig die Genauigkeit und Konsistenz.

In diesem Leitfaden wird der Einsatz von Amazon Comprehend Medical und LLMs die intelligente Automatisierung im Gesundheitswesen untersucht. Es beschreibt bewährte Verfahren, Herausforderungen und praktische Ansätze zur Rationalisierung der Prozesse für die medizinische Kodierung, die Extraktion von Patienteninformationen und die Zusammenfassung von Aufzeichnungen. Durch die Nutzung der Funktionen von Amazon Comprehend Medical und LLMs können Organisationen im Gesundheitswesen ein neues Maß an betrieblicher Effizienz erreichen, Kosten senken und potenziell die Patientenversorgung verbessern.

Der Leitfaden beschreibt die besonderen Aspekte des Gesundheitswesens, wie z. B. das Verständnis der medizinischen Terminologie, die Verwendung domänenspezifischer LLMs Anwendungen und die Beseitigung der Einschränkungen von Systemen. AI/ML Er bietet IT-Managern, Architekten und technischen Experten im Gesundheitswesen einen umfassenden Entscheidungspfad, mit dem sie beurteilen können, ob die Organisation bereit ist, Implementierungsoptionen zu bewerten und die geeigneten AWS-Services Tools für eine erfolgreiche Automatisierung einzusetzen.

Indem sie die in diesem Leitfaden beschriebenen Richtlinien und bewährten Verfahren befolgen, können Organisationen im Gesundheitswesen das Potenzial von AI/ML Technologien nutzen und gleichzeitig die Komplexität des medizinischen Bereichs bewältigen. Dieser Ansatz unterstützt die Einhaltung ethischer und regulatorischer Richtlinien und fördert den verantwortungsvollen Einsatz von KI-Systemen im Gesundheitswesen. Es wurde entwickelt, um genaue und private Erkenntnisse zu generieren.

Zielgruppe

Dieser Leitfaden richtet sich an Technologievertreter, Architekten, technische Leiter und Entscheidungsträger, die KI-gestützte Lösungen zur Verarbeitung natürlicher Sprache für die Analyse und Automatisierung medizinischer Daten implementieren möchten.

Ziele

Organisationen im Gesundheitswesen und in den Biowissenschaften können mithilfe von Amazon Comprehend Medical und mehrere Geschäftsziele erreichen. LLMs Zu diesen Ergebnissen gehören in der Regel die Steigerung der betrieblichen Effizienz, die Senkung der Kosten und die Verbesserung der Patientenversorgung. In diesem Abschnitt werden die wichtigsten Geschäftsziele und die damit verbundenen Vorteile beschrieben, die sich aus der Umsetzung der in diesem Leitfaden beschriebenen Strategien und bewährten Verfahren ergeben.

Im Folgenden sind einige der Ziele aufgeführt, die Unternehmen durch die Umsetzung der Richtlinien und bewährten Verfahren in diesem Leitfaden erreichen können:

- Verkürzung der Entwicklungszeit — Das ultimative Ziel dieses Leitfadens besteht darin, die Entwicklungszeit und damit verbundene Kosten zu reduzieren, technische Schulden zu verringern und mögliche Projektausfälle aufgrund von POC zu verhindern. Wenn Unternehmen wichtige AI/ML Dienste wie Amazon Comprehend Medical und die Vor- und Nachteile der LLM-Nutzung für Aufgaben im Gesundheitswesen verstehen, können sie die Markteinführungszeit verkürzen und ihre Geschäftsziele schneller erreichen.

- Extrahieren Sie Informationen, um medizinische Kodierungsaufgaben zu automatisieren — Nach Patientenbesuchen können Kodierungsspezialisten und -dienstleister Erkenntnisse aus medizinischen Texten gewinnen, z. B. subjektive, objektive Notizen, Beurteilungs- und Plannotizen (SOAP). Dies kann den manuellen Dokumentationsaufwand reduzieren und dem Anbieter helfen, sich auf die Bedürfnisse des Patienten zu konzentrieren. Durch die Kombination der Funktionen zur Erkennung von Entitäten von Amazon Comprehend Medical mit LLMs können Unternehmen relevante medizinische Informationen aus Patientenakten, klinischen Notizen und anderen Gesundheitsdatenquellen extrahieren. Dadurch können menschliche Fehler minimiert und einheitliche Verfahren gefördert werden.
- Zusammenfassung von Patientenakten und klinischen Unterlagen — Durch die automatisierte Zusammenfassung der Krankengeschichte, der Behandlungspläne und der medizinischen Ergebnisse können Gesundheitsdienstleister wertvolle Zeit sparen. LLMs kann bei der Erstellung umfassender und strukturierter klinischer Unterlagen helfen. Sie können mit Amazon Comprehend Medical zusätzlichen Kontext abrufen, ein LLM für medizinische Domänen verwenden oder ein LLM mit medizinischen Daten verfeinern. Diese Ansätze können dazu beitragen, genaue Zusammenfassungen zu erstellen und sicherzustellen, dass die Dokumentation den Compliance-Anforderungen und -Standards entspricht.
- Support klinischer Entscheidungen und Patientenversorgung — Mithilfe der Ontologie-Verknüpfung in Amazon Comprehend Medical können Anbieter medizinische Fragen beantworten oder Empfehlungen zur Patientenversorgung einholen. LLMs Auf diese Weise können medizinische Fachkräfte fundierte Entscheidungen treffen, die die Behandlungsergebnisse verbessern und das Risiko von medizinischen Fehlern verringern.

Generative KI- und NLP-Ansätze für das Gesundheitswesen und die Biowissenschaften

Natural Language Processing (NLP) ist eine Technologie für maschinelles Lernen, die es Computern ermöglicht, menschliche Sprache zu interpretieren, zu manipulieren und zu verstehen. Organisationen im Gesundheitswesen und in den Biowissenschaften verfügen über große Datenmengen aus Patientenakten. Sie können NLP-Software verwenden, um diese Daten automatisch zu verarbeiten. Sie können beispielsweise NLP mit generativer KI kombinieren, um die medizinische Kodierung zu optimieren, Patienteninformationen zu extrahieren und Aufzeichnungen zusammenzufassen.

Abhängig von der NLP-Aufgabe, die Sie ausführen möchten, sind unterschiedliche Architekturen möglicherweise am besten für Ihren Anwendungsfall geeignet. Dieser Leitfaden befasst sich mit den folgenden generativen KI- und NLP-Optionen für Anwendungen im Gesundheitswesen und in den Biowissenschaften in folgenden Bereichen: AWS

- Verwenden von Amazon Comprehend Medical— Erfahren Sie, wie Sie Amazon Comprehend Medical unabhängig verwenden können, ohne es in ein großes Sprachmodell (LLM) zu integrieren.
- Kombination von Amazon Comprehend Medical mit großen Sprachmodellen— Erfahren Sie, wie Sie Amazon Comprehend Medical mit einem LLM in einer Retrieval Augment Generation (RAG) - Architektur kombinieren können.
- Verwendung umfangreicher Sprachmodelle für Anwendungsfälle im Gesundheitswesen und in den Biowissenschaften— Erfahren Sie, wie Sie ein LLM für Anwendungen im Gesundheitswesen und in den Biowissenschaften einsetzen können, entweder mithilfe einer fein abgestimmten LLM- oder einer RAG-Architektur.

Verwenden von Amazon Comprehend Medical

Amazon Comprehend Medical erkennt und sendet nützliche Informationen in unstrukturiertem klinischem Text wie Arztnotizen, Zusammenfassungen von Entlassungen, Testergebnissen und Fallnotizen. AWS-Service Es verwendet Modelle zur Verarbeitung natürlicher Sprache (NLP), um Entitäten zu erkennen. Entitäten sind Textverweise auf medizinische Informationen wie Erkrankungen, Medikamente oder geschützte Gesundheitsinformationen (PHI).

> **Important**
>
> Amazon Comprehend Medical ist kein Ersatz für professionelle medizinische Beratung, Diagnose oder Behandlung. Amazon Comprehend Medical bietet Konfidenzwerte, die das Maß an Vertrauen in die Genauigkeit der erkannten Entitäten angeben. Identifizieren Sie den richtigen Konfidenzschwellenwert für Ihren Anwendungsfall, und verwenden Sie hohe Konfidenzschwellenwerte in Situationen, die eine hohe Genauigkeit erfordern. In bestimmten Anwendungsfällen sollten die Ergebnisse von entsprechend geschulten menschlichen Prüfern überprüft und verifiziert werden. Amazon Comprehend Medical sollte beispielsweise nur in Patientenversorgungsszenarien verwendet werden, nachdem es von geschultem medizinischem Fachpersonal auf Richtigkeit und fundiertes medizinisches Urteilsvermögen überprüft wurde.

Sie können auf Amazon Comprehend Medical über die AWS-Managementkonsole, die AWS Command Line Interface (AWS CLI) oder über die zugreifen. AWS SDKs Sie AWS SDKs sind für verschiedene Programmiersprachen und Plattformen wie Java, Python, Ruby, .NET, iOS und Android verfügbar. Sie können den verwenden SDKs , um programmgesteuert von Ihrer Client-Anwendung aus auf Amazon Comprehend Medical zuzugreifen.

In diesem Abschnitt werden die wichtigsten Funktionen von Amazon Comprehend Medical beschrieben. Außerdem werden die Vorteile dieses Dienstes im Vergleich zu einem Large Language Model (LLM) erörtert.

Funktionen von Amazon Comprehend Medical

Amazon Comprehend Medical bietet APIs nahezu Echtzeit- und Batch-Inferenzen. Diese APIs können medizinischen Text aufnehmen und Ergebnisse für medizinische NLP-Aufgaben liefern, indem sie medizinische Entitäten erkennen und Entitätsbeziehungen identifizieren. Sie können Analysen sowohl für einzelne Dateien als auch als Batch-Analyse für mehrere Dateien durchführen, die in einem Amazon Simple Storage Service (Amazon S3) -Bucket gespeichert sind. Amazon Comprehend Medical bietet die folgenden Textanalyse-API-Operationen für die synchrone Erkennung von Entitäten:

- Entitäten erkennen — Erkennt allgemeine medizinische Kategorien wie Anatomie, Gesundheitszustand, PHI-Kategorie, Verfahren und Zeitausdrücke.

- PHI erkennen — Erkennt bestimmte Entitäten wie Alter, Datum, Name und ähnliche persönliche Informationen.

Amazon Comprehend Medical umfasst auch mehrere API-Operationen, mit denen Sie Batch-Textanalysen für klinische Dokumente durchführen können. Weitere Informationen zur Verwendung dieser API-Operationen finden Sie unter Batch zur Textanalyse. APIs

Verwenden Sie Amazon Comprehend Medical, um Entitäten in klinischem Text zu erkennen und diese Entitäten mit Konzepten in standardisierten medizinischen Ontologien zu verknüpfen, einschließlich der RxNorm Wissensdatenbanken ICD-10-CM und SNOMED CT. Sie können Analysen sowohl für einzelne Dateien als auch als Batch-Analyse für große Dokumente oder mehrere Dateien durchführen, die in einem Amazon S3 S3-Bucket gespeichert sind. Amazon Comprehend Medical bietet die folgende ontologische Verknüpfung von API-Operationen:

- Infer ICD10 CM — Die Infer ICD10 CM-Operation erkennt potenzielle Erkrankungen und verknüpft sie mit Codes aus der Version 2019 der Internationalen Klassifikation der Krankheiten, 10. Revision, Klinische Änderung (ICD-10-CM). Für jede festgestellte potenzielle Erkrankung listet Amazon Comprehend Medical die entsprechenden ICD-10-CM-Codes und Beschreibungen auf. Zu den in den Ergebnissen aufgelisteten Erkrankungen gehört ein Konfidenzwert, der das Vertrauen angibt, das Amazon Comprehend Medical in die Genauigkeit der Entitäten und der übereinstimmenden Konzepte in den Ergebnissen hat.
- InferRxNorm— Die InferRxNormOperation identifiziert Medikamente, die in einer Patientenakte als Entitäten aufgeführt sind. Dabei werden Entitäten mit Konzeptkennungen (RxCUI) aus der RxNorm Datenbank der National Library of Medicine verknüpft. Jeder RxCUI ist für unterschiedliche Stärken und Darreichungsformen einzigartig. Die in den Ergebnissen aufgelisteten Medikamente enthalten einen Konfidenzwert, der das Vertrauen von Amazon Comprehend Medical in die Genauigkeit der Entitäten angibt, die den Konzepten aus der RxNorm Wissensdatenbank zugeordnet wurden. Amazon Comprehend Medical listet die Top Rx aufCUIs , die möglicherweise für jedes erkannte Medikament in absteigender Reihenfolge auf, basierend auf dem Vertrauenswert.
- InfersnoMedCT — Die Operation InfersnoMedCT identifiziert mögliche medizinische Konzepte als Entitäten und verknüpft sie mit Codes aus der Version 2021-03 der Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT). SNOMED CT bietet ein umfassendes Vokabular medizinischer Konzepte, einschließlich Erkrankungen und Anatomie sowie medizinischer Tests, Behandlungen und Verfahren. Für jede übereinstimmende Konzept-ID gibt Amazon Comprehend Medical die fünf besten medizinischen Konzepte zurück, jeweils mit einem Konfidenzwert und Kontextinformationen wie Merkmalen und Attributen. Das SNOMED-CT-Konzept IDs kann dann

verwendet werden, um klinische Patientendaten für die medizinische Kodierung, Berichterstattung oder klinische Analysen zu strukturieren, wenn es mit der SNOMED-CT-Polyhierarchie verwendet wird.

Weitere Informationen finden Sie unter Textanalyse APIs und Ontologie-Verknüpfung APIs in der Amazon Comprehend Medical Medical-Dokumentation.

Anwendungsfälle für Amazon Comprehend Medical

Als eigenständiger Service kann Amazon Comprehend Medical auf den Anwendungsfall Ihres Unternehmens zugeschnitten sein. Amazon Comprehend Medical kann Aufgaben wie die folgenden ausführen:

- Hilfe bei der medizinischen Kodierung in Patientenakten
- Erkennen Sie geschützte Gesundheitsinformationsdaten (PHI)
- Validierung von Medikamenten, einschließlich Eigenschaften wie Dosierung, Häufigkeit und Form

Die Ergebnisse von Amazon Comprehend Medical sind für die meisten Arztpraxen leicht verdaulich. Möglicherweise müssen Sie jedoch Alternativen in Betracht ziehen, wenn Sie Einschränkungen wie die folgenden haben:

- Verschiedene Entitätsdefinitionen — Ihre Definition einer `FREQUENCY` Medikamentenentität kann beispielsweise unterschiedlich sein. Für die Häufigkeit prognostiziert Amazon Comprehend Medical nach Bedarf, aber Ihre Organisation verwendet möglicherweise den Begriff Pro re Nata (PRN).
- Überwältigende Menge an Ergebnissen — Patientennotizen enthalten beispielsweise häufig mehrere Symptome und Stichwörter, die mehreren ICD-10-CM-Codes zugeordnet sind. Einige der Schlüsselwörter sind jedoch nicht für die Diagnose geeignet. In diesem Fall muss der Anbieter zahlreiche ICD-10-CM-Entitäten und ihre Konfidenzwerte bewerten, was eine manuelle Verarbeitungszeit erfordert.
- Benutzerdefinierte Entitäten oder NLP-Aufgaben — Anbieter könnten zum Beispiel PRN-Beweise extrahieren wollen, etwa bei Bedarf gegen Schmerzen. Da dies nicht über Amazon Comprehend Medical erhältlich ist, ist ein anderes AI/ML Modell garantiert. Eine andere AI/ML Lösung ist erforderlich, wenn die NLP-Aufgabe außerhalb der Entitätenerkennung liegt, z. B. Zusammenfassung, Beantwortung von Fragen und Stimmungsanalyse.

Kombination von Amazon Comprehend Medical mit großen Sprachmodellen

Eine Studie von NEJM AI aus dem Jahr 2024 zeigte, dass die Verwendung eines LLM mit Null-Shot-Aufforderung für medizinische Codierungsaufgaben im Allgemeinen zu einer schlechten Leistung führt. Die Verwendung von Amazon Comprehend Medical mit einem LLM kann dazu beitragen, diese Leistungsprobleme zu verringern. Die Ergebnisse von Amazon Comprehend Medical sind ein hilfreicher Kontext für ein LLM, das NLP-Aufgaben ausführt. Wenn Sie beispielsweise Kontext von Amazon Comprehend Medical zum großen Sprachmodell bereitstellen, können Sie Folgendes erreichen:

- Verbessern Sie die Genauigkeit der Entitätsauswahl, indem Sie die ersten Ergebnisse von Amazon Comprehend Medical als Kontext für das LLM verwenden
- Implementieren Sie benutzerdefinierte Entitätenerkennung, Zusammenfassung, Beantwortung von Fragen und weitere Anwendungsfälle

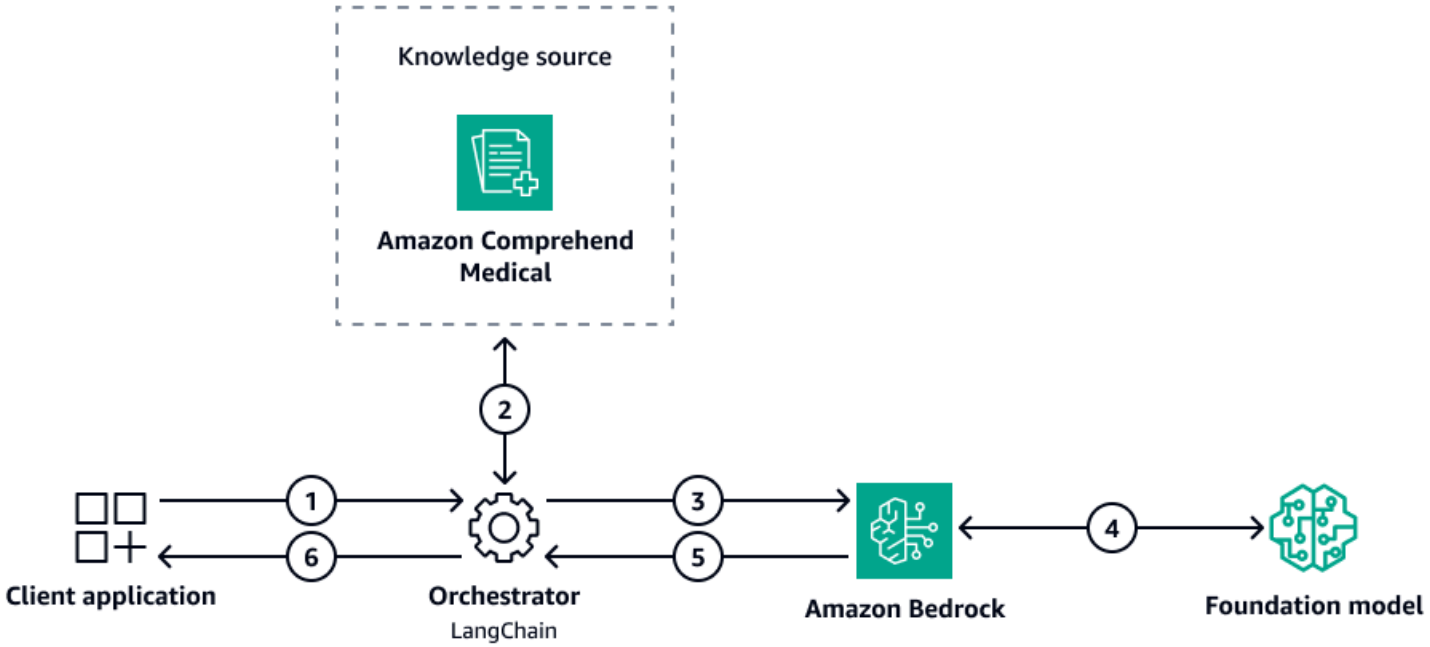
In diesem Abschnitt wird beschrieben, wie Sie Amazon Comprehend Medical mit einem LLM kombinieren können, indem Sie einen Retrieval Augmented Generation (RAG) -Ansatz verwenden. Retrieval Augmented Generation (RAG) ist eine generative KI-Technologie, bei der ein LLM auf eine maßgebliche Datenquelle verweist, die sich außerhalb seiner Trainingsdatenquellen befindet, bevor es eine Antwort generiert. Weitere Informationen finden Sie unter Was ist RAG.

Zur Veranschaulichung dieses Ansatzes wird in diesem Abschnitt das Beispiel der medizinischen (Diagnose-) Kodierung im Zusammenhang mit ICD-10-CM verwendet. Es enthält eine Beispielarchitektur und schnelle technische Vorlagen, mit denen Sie Ihre Innovation beschleunigen können. Es enthält auch bewährte Methoden für die Verwendung von Amazon Comprehend Medical innerhalb eines RAG-Workflows.

RAG-basierte Architektur mit Amazon Comprehend Medical

Das folgende Diagramm veranschaulicht einen RAG-Ansatz zur Identifizierung von ICD-10-CM-Diagnosecodes anhand von Patientennotizen. Es verwendet Amazon Comprehend Medical als Wissensquelle. Bei einem RAG-Ansatz ruft die Abrufmethode üblicherweise Informationen aus einer Vektordatenbank ab, die anwendbares Wissen enthält. Anstelle einer Vektordatenbank verwendet diese Architektur Amazon Comprehend Medical für die Abrufaufgabe. Der Orchestrator sendet die Patientennotizen an Amazon Comprehend Medical und ruft die ICD-10-CM-Codeinformationen ab.

Der Orchestrator sendet diesen Kontext über Amazon Bedrock an das Downstream Foundation Model (LLM). Das LLM generiert mithilfe der ICD-10-CM-Codeinformationen eine Antwort, und diese Antwort wird an die Client-Anwendung zurückgesendet.



Das Diagramm zeigt den folgenden RAG-Workflow:

1. Die Client-Anwendung sendet die Patientennotizen als Anfrage an den Orchestrator. Ein Beispiel für diese Patientennotizen könnte lauten: „Bei der Patientin handelt es sich um eine 71-jährige Patientin von Dr. X. Die Patientin wurde gestern Abend in die Notaufnahme gebracht und hatte in der Anamnese etwa 7 bis 8 Tage lang anhaltende Bauchschmerzen. Sie hatte kein definitives Fieber oder Schüttelfrost und keine Gelbsucht in der Vorgeschichte. Die Patientin bestreitet, in letzter Zeit signifikant abgenommen zu haben.“
2. Der Orchestrator verwendet Amazon Comprehend Medical, um ICD-10-CM-Codes abzurufen, die für die medizinischen Informationen in der Abfrage relevant sind. Er verwendet die Infer ICD10 CM-API, um die ICD-10-CM-Codes aus den Patientennotizen zu extrahieren und abzuleiten.
3. Der Orchestrator erstellt eine Aufforderung, die die Eingabeaufforderungsvorlage, die ursprüngliche Abfrage und die von Amazon Comprehend Medical abgerufenen ICD-10-CM-Codes enthält. Es sendet diesen erweiterten Kontext an Amazon Bedrock.
4. Amazon Bedrock verarbeitet die Eingabe und generiert anhand eines Fundamentmodells eine Antwort, die die ICD-10-CM-Codes und die entsprechenden Beweise aus der Abfrage enthält. Die generierte Antwort umfasst die identifizierten ICD-10-CM-Codes und Nachweise aus den Patientenakten, die jeden Code belegen. Das folgende Beispiel zeigt eine mögliche Antwort:

```
<response>
<icd10>
<code>R10.9</code>
<evidence>history of abdominal pain</evidence>
</icd10>
<icd10>
<code>R10.30</code>
<evidence>history of abdominal pain</evidence>
</icd10>
</response>
```

5. Amazon Bedrock sendet die generierte Antwort an den Orchestrator.
6. Der Orchestrator sendet die Antwort zurück an die Client-Anwendung, wo der Benutzer die Antwort überprüfen kann.

Anwendungsfälle für die Verwendung von Amazon Comprehend Medical in einem RAG-Workflow

Amazon Comprehend Medical kann bestimmte NLP-Aufgaben ausführen. Weitere Informationen finden Sie unter Anwendungsfälle für Amazon Comprehend Medical.

Möglicherweise möchten Sie Amazon Comprehend Medical in einen RAG-Workflow für erweiterte Anwendungsfälle integrieren, z. B. für die folgenden:

- Generieren Sie detaillierte klinische Zusammenfassungen, indem Sie extrahierte medizinische Entitäten mit Kontextinformationen aus Patientenakten kombinieren
- Automatisieren Sie die medizinische Kodierung für komplexe Fälle, indem Sie extrahierte Entitäten mit ontologiebezogenen Informationen für die Codezuweisung verwenden
- Automatisieren Sie die Erstellung strukturierter klinischer Notizen aus unstrukturiertem Text mithilfe extrahierter medizinischer Entitäten
- Analysieren Sie Nebenwirkungen von Medikamenten anhand der Namen und Eigenschaften der extrahierten Medikamente
- Entwickeln Sie intelligente klinische Unterstützungssysteme, die extrahierte medizinische Informationen mit up-to-date Forschungsergebnissen und Leitlinien kombinieren

Bewährte Methoden für die Verwendung von Amazon Comprehend Medical in einem RAG-Workflow

Bei der Integration der Ergebnisse von Amazon Comprehend Medical in eine Aufforderung zur Einreichung eines LLM ist es wichtig, die Best Practices zu befolgen. Dies kann die Leistung und Genauigkeit verbessern. Im Folgenden finden Sie die wichtigsten Empfehlungen:

- Verstehen Sie die Vertrauenswerte von Amazon Comprehend Medical — Amazon Comprehend Medical bietet Konfidenzwerte für jede erkannte Entität und Ontologieverknüpfung. Es ist wichtig, die Bedeutung dieser Werte zu verstehen und geeignete Schwellenwerte für Ihren spezifischen Anwendungsfall festzulegen. Konfidenzwerte helfen dabei, Entitäten mit geringer Zuverlässigkeit herauszufiltern, das Rauschen zu reduzieren und die Qualität der LLM-Eingaben zu verbessern.
- Verwenden Sie Konfidenzwerte in der Prompt-Technik — Ziehen Sie bei der Erstellung von Aufforderungen für das LLM in Betracht, Amazon Comprehend Medical Confidence Scores als zusätzlichen Kontext einzubeziehen. Dies hilft dem LLM, Unternehmen anhand ihres Konfidenzniveaus zu priorisieren oder abzuwägen, wodurch möglicherweise die Qualität der Ergebnisse verbessert wird.
- Bewerten Sie die Ergebnisse von Amazon Comprehend Medical mit Ground-Truth-Daten — Ground-Truth-Daten sind Informationen, von denen bekannt ist, dass sie wahr sind. Sie können verwendet werden, um zu überprüfen, ob eine AI/ML Anwendung genaue Ergebnisse liefert. Bevor Sie die Ergebnisse von Amazon Comprehend Medical in Ihren LLM-Workflow integrieren, bewerten Sie die Leistung des Services anhand einer repräsentativen Stichprobe Ihrer Daten. Vergleichen Sie die Ergebnisse mit Ground-Truth-Anmerkungen, um mögliche Unstimmigkeiten oder Verbesserungsmöglichkeiten zu identifizieren. Diese Bewertung hilft Ihnen, die Stärken und Grenzen von Amazon Comprehend Medical für Ihren Anwendungsfall zu verstehen.
- Strategisch relevante Informationen auswählen — Amazon Comprehend Medical kann eine große Menge an Informationen bereitstellen, aber möglicherweise sind nicht alle für Ihre Aufgabe relevant. Wählen Sie sorgfältig die Entitäten, Attribute und Metadaten aus, die für Ihren Anwendungsfall am relevantesten sind. Wenn Sie dem LLM zu viele irrelevante Informationen zur Verfügung stellen, kann dies zu Störungen und möglicherweise zu Leistungseinbußen führen.
- Entitätsdefinitionen aufeinander abstimmen — Stellen Sie sicher, dass die Definitionen von Entitäten und Attributen, die von Amazon Comprehend Medical verwendet werden, mit Ihrer Interpretation übereinstimmen. Wenn es Unstimmigkeiten gibt, sollten Sie erwägen, dem LLM zusätzlichen Kontext oder Erläuterungen zur Verfügung zu stellen, um die Lücke zwischen den Ergebnissen von Amazon Comprehend Medical und Ihren Anforderungen zu schließen. Wenn

die Amazon Comprehend Medical Medical-Einheit Ihre Erwartungen nicht erfüllt, können Sie eine benutzerdefinierte Entitätenerkennung implementieren, indem Sie zusätzliche Anweisungen (und mögliche Beispiele) in die Aufforderung einfügen.

- Stellen Sie domänenspezifisches Wissen bereit — Amazon Comprehend Medical bietet zwar wertvolle medizinische Informationen, erfasst aber möglicherweise nicht alle Nuancen Ihres spezifischen Fachgebiets. Erwägen Sie, die Ergebnisse von Amazon Comprehend Medical durch zusätzliche domänenspezifische Wissensquellen wie Ontologien, Terminologien oder von Experten kuratierte Datensätze zu ergänzen. Dies bietet einen umfassenderen Kontext für das LLM.
- Einhaltung ethischer und regulatorischer Richtlinien — Beim Umgang mit medizinischen Daten ist es wichtig, ethische Grundsätze und regulatorische Richtlinien einzuhalten, beispielsweise in Bezug auf Datenschutz, Sicherheit und verantwortungsvollen Umgang mit KI-Systemen im Gesundheitswesen. Stellen Sie sicher, dass Ihre Implementierung den geltenden Gesetzen und bewährten Verfahren der Branche entspricht.

Durch die Befolgung dieser bewährten Methoden können AI/ML Praktiker die Stärken von Amazon Comprehend Medical und effektiv nutzen. LLMs Bei medizinischen NLP-Aufgaben tragen diese bewährten Methoden dazu bei, potenzielle Risiken zu minimieren und die Leistung zu verbessern.

Promptes Engineering für den medizinischen Kontext von Amazon Comprehend

Promptes Engineering ist der Prozess, bei dem Eingabeaufforderungen entworfen und verfeinert werden, um eine generative KI-Lösung zur Generierung der gewünschten Ergebnisse zu leiten. Sie wählen die am besten geeigneten Formate, Ausdrücke, Wörter und Symbole aus, anhand derer die KI sinnvoller mit Ihren Benutzern interagieren kann.

Abhängig von der API-Operation, die Sie ausführen, gibt Amazon Comprehend Medical die erkannten Entitäten, Ontologiecodes und Beschreibungen sowie Konfidenzwerte zurück. Diese Ergebnisse werden innerhalb der Aufforderung zum Kontext, wenn Ihre Lösung das Ziel-LLM aufruft. Sie müssen die Eingabeaufforderung so gestalten, dass der Kontext in der Eingabeaufforderungsvorlage dargestellt wird.

> **Note**
>
> Die Beispiel-Eingabeaufforderungen in diesem Abschnitt folgen anthropischen Leitlinien. Wenn Sie einen anderen LLM-Anbieter verwenden, befolgen Sie die Empfehlungen dieses Anbieters.

Im Allgemeinen fügen Sie sowohl den medizinischen Originaltext als auch die Ergebnisse von Amazon Comprehend Medical in die Aufforderung ein. Im Folgenden finden Sie eine allgemeine Eingabeaufforderungsstruktur:

```
<medical_text>
medical text
</medical_text>

<comprehend_medical_text_results>
comprehend medical text results
</comprehend_medical_text_results>

<prompt_instructions>
prompt instructions
</prompt_instructions>
```

Dieser Abschnitt enthält Strategien, um die Ergebnisse von Amazon Comprehend Medical als sofortigen Kontext für die folgenden allgemeinen medizinischen NLP-Aufgaben einzubeziehen:

- Amazon Comprehend Medical Medical-Ergebnisse filtern
- Erweitern Sie medizinische NLP-Aufgaben mit Amazon Comprehend Medical
- Wenden Sie Leitplanken mit Amazon Comprehend Medical an

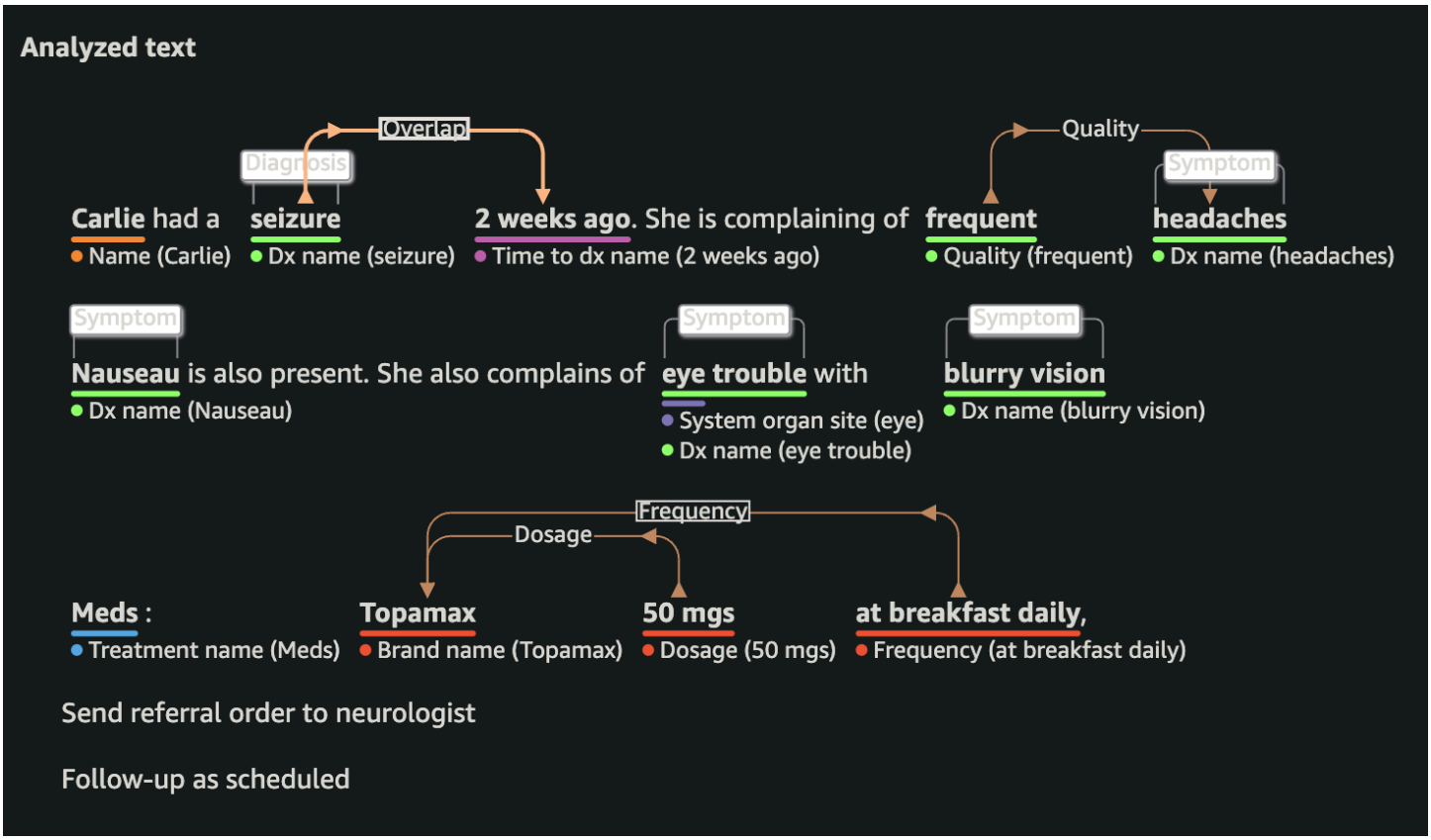
Amazon Comprehend Medical Medical-Ergebnisse filtern

Amazon Comprehend Medical stellt in der Regel eine große Menge an Informationen bereit. Möglicherweise möchten Sie die Anzahl der Ergebnisse reduzieren, die der Arzt überprüfen muss. In diesem Fall können Sie ein LLM verwenden, um diese Ergebnisse zu filtern. Die Entitäten von Amazon Comprehend Medical enthalten einen Konfidenzwert, den Sie bei der Gestaltung der Aufforderung als Filtermechanismus verwenden können.

Im Folgenden finden Sie ein Beispiel für eine Patientennotiz:

```
Carlie had a seizure 2 weeks ago. She is complaining of frequent headaches
Nausea is also present. She also complains of eye trouble with blurry vision
Meds : Topamax 50 mgs at breakfast daily,
Send referral order to neurologist
Follow-up as scheduled
```

In dieser Patientennotiz erkennt Amazon Comprehend Medical die folgenden Entitäten.



Die Entitäten verweisen auf die folgenden ICD-10-CM-Codes für Anfälle und Kopfschmerzen.

Kategorie	ICD-10-CM-Code	ICD-10-CM-Beschreibung	Zuverlässigkeitswert
Anfall	R56.9	Nicht näher bezeichnete Krämpfe	0,8348
Anfall	G40.909	Epilepsie, nicht näher bezeichnet, nicht	0,5424

		hartnäckig, ohne Status epilepticus	
Anfall	R56,00	Einfache Fieberkrämpfe	0,4937
Anfall	G40.09	Andere Anfälle	0,4397
Anfall	G40.409	Sonstige generalisierte Epilepsie und epileptische Syndrome, nicht behandelbar, ohne Status epilepticus	0,4138
Kopfschmerzen	R51	Kopfschmerzen	0,4067
Kopfschmerzen	R51.9	Kopfschmerzen, nicht näher bezeichnet	0,3844
Kopfschmerzen	G44.52	Neue tägliche anhaltende Kopfschmerzen (NDPH)	0,3005
Kopfschmerzen	G44	Anderes Kopfschmerzsyndrom	0,2670
Kopfschmerzen	G44.8	Andere näher bezeichnete Kopfschmerzsyndrome	0,2542

Sie können ICD-10-CM-Codes an die Eingabeaufforderung übergeben, um die LLM-Präzision zu erhöhen. Um das Rauschen zu reduzieren, können Sie die ICD-10-CM-Codes anhand des Konfidenzwerts filtern, der in den Ergebnissen von Amazon Comprehend Medical enthalten ist. Im Folgenden finden Sie ein Beispiel für eine Eingabeaufforderung, die nur ICD-10-CM-Codes mit einem Konfidenzwert von mehr als 0,4 enthält:

```
<patient_note>
Carlie had a seizure 2 weeks ago. She is complaining of frequent headaches
Nausea is also present. She also complains of eye trouble with blurry vision
Meds : Topamax 50 mgs at breakfast daily,
Send referral order to neurologist
Follow-up as scheduled
</patient_note>

<comprehend_medical_results>
<icd-10>
  <entity>
    <text>seizure</text>
    <code>
      <description>Unspecified convulsions</description>
      <code_value>R56.9</code_value>
      <score>0.8347607851028442</score>
    </code>
    <code>
      <description>Epilepsy, unspecified, not intractable, without status epilepticus</
description>
      <code_value>G40.909</code_value>
      <score>0.542376697063446</score>
    </code>
    <code>
      <description>Other seizures</description>
      <code_value>G40.89</code_value>
      <score>0.43966275453567505</score>
    </code>
    <code>
      <description>Other generalized epilepsy and epileptic syndromes, not intractable,
 without status epilepticus</description>
      <code_value>G40.409</code_value>
      <score>0.41382506489753723</score>
    </code>
  </entity>
  <entity>
    <text>headaches</text>
    <code>
      <description>Headache</description>
      <code_value>R51</code_value>
      <score>0.4066613018512726</score>
    </code>
  </entity>
```

```
  <entity>
    <text>Nausea</text>
    <code>
      <description>Nausea</description>
      <code_value>R11.0</code_value>
      <score>0.6460834741592407</score>
    </code>
  </entity>
  <entity>
    <text>eye trouble</text>
    <code>
      <description>Unspecified disorder of eye and adnexa</description>
      <code_value>H57.9</code_value>
      <score>0.6780954599380493</score>
    </code>
    <code>
      <description>Unspecified visual disturbance</description>
      <code_value>H53.9</code_value>
      <score>0.5871203541755676</score>
    </code>
    <code>
      <description>Unspecified disorder of binocular vision</description>
      <code_value>H53.30</code_value>
      <score>0.5539672374725342</score>
    </code>
  </entity>
  <entity>
    <text>blurry vision</text>
    <code>
      <description>Other visual disturbances</description>
      <code_value>H53.8</code_value>
      <score>0.9001834392547607</score>
    </code>
  </entity>
</icd-10>
</comprehend_medical_results>

<prompt>
Given the patient note and Amazon Comprehend Medical ICD-10-CM code results above,
 please select the most relevant ICD-10-CM diagnosis codes for the patient.
For each selected code, provide a brief explanation of why it is relevant based on the
 information in the patient note.
</prompt>
```

Erweitern Sie medizinische NLP-Aufgaben mit Amazon Comprehend Medical

Bei der Verarbeitung von medizinischem Text kann der Kontext von Amazon Comprehend Medical dem LLM helfen, bessere Token auszuwählen. In diesem Beispiel möchten Sie die Diagnosesymptome den Medikamenten zuordnen. Sie möchten auch Text finden, der sich auf medizinische Tests bezieht, z. B. Begriffe, die sich auf einen Bluttest beziehen. Sie können Amazon Comprehend Medical verwenden, um die Entitäten und Medikamentennamen zu ermitteln. In diesem Fall würden Sie DetectEntitiesV2 und InferRxNorm APIs Amazon Comprehend Medical verwenden.

Im Folgenden finden Sie ein Beispiel für eine Patientennotiz:

```
Carlie had a seizure 2 weeks ago. She is complaining of increased frequent headaches
Given lyme disease symptoms such as muscle ache and stiff neck will order prescription.
Meds : Topamax 50 mgs at breakfast daily. Amoxicillan 25 mg by mouth twice a day
Place MRI radiology order at RadNet
```

Um den Fokus auf den Diagnosecode zu legen, werden in der Eingabeaufforderung nur die Entitäten verwendet, die `DX_NAME` sich auf den Typ beziehen. `MEDICAL_CONDITION` Andere Metadaten sind aufgrund ihrer Irrelevanz ausgeschlossen. Bei Medikamenten-Entitäten ist der Name des Medikaments zusammen mit den extrahierten Attributen enthalten. Andere Metadaten von Arzneimittelentitäten von Amazon Comprehend Medical sind aufgrund ihrer Irrelevanz ausgeschlossen. Die folgende Beispielaufforderung verwendet gefilterte Amazon Comprehend Medical Medical-Ergebnisse. Die Eingabeaufforderung konzentriert sich auf `MEDICAL_CONDITION` Entitäten mit dem `DX_NAME` Typ. Diese Aufforderung dient dazu, Diagnosecodes genauer mit Medikamenten zu verknüpfen und medizinische Tests genauer zu extrahieren:

```
<patient_note>
Carlie had a seizure 2 weeks ago. She is complaining of increased freqeunt headaches
Given lyme disease symptoms such as muscle ache and stiff neck will order
 prescription.
Meds : Topamax 50 mgs at breakfast daily. Amoxicillan 25 mg by mouth twice a day
Place MRI radiology order at RadNet
</patient_note>

<detect_entity_results>
<entity>
    <text>seizure</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
```

```
<entity>
    <text>headaches</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>lyme disease</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>muscle ache</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>stiff neck</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
</detect_entity_results>

<rx_results>
<entity>
    <text>Topamax</text>
    <category>MEDICATION</category>
    <type>BRAND_NAME</type>
    <attributes>
        <attribute>
            <type>FREQUENCY</type>
            <text>at breakfast daily</text>
        </attribute>
        <attribute>
            <type>DOSAGE</type>
            <text>50 mgs</text>
        </attribute>
        <attribute>
            <type>ROUTE_OR_MODE</type>
            <text>by mouth</text>
        </attribute>
    </attributes>
</entity>
<entity>
```

```
    <text>Amoxicillan</text>
    <category>MEDICATION</category>
    <type>GENERIC_NAME</type>
    <attributes>
        <attribute>
            <type>ROUTE_OR_MODE</type>
            <text>by mouth</text>
        </attribute>
        <attribute>
            <type>DOSAGE</type>
            <text>25 mg</text>
        </attribute>
        <attribute>
            <type>FREQUENCY</type>
            <text>twice a day</text>
        </attribute>
    </attributes>
</entity>
</rx_results>

<prompt>
Based on the patient note and the detected entities, can you please:
1. Link the diagnosis symptoms with the medications prescribed.
Provide your reasoning for the linkages.
2. Extract any entities related to medical order tests mentioned in the note.
</prompt>
```

Wenden Sie Leitplanken mit Amazon Comprehend Medical an

Sie können ein LLM und Amazon Comprehend Medical verwenden, um Leitplanken zu erstellen, bevor die generierte Antwort verwendet wird. Sie können diesen Workflow entweder für unveränderten oder für nachbearbeiteten medizinischen Text ausführen. Zu den Anwendungsfällen gehören der Umgang mit geschützten Gesundheitsinformationen (PHI), die Erkennung von Halluzinationen oder die Implementierung benutzerdefinierter Richtlinien für die Veröffentlichung von Ergebnissen. Sie können beispielsweise den Kontext von Amazon Comprehend Medical verwenden, um PHI-Daten zu identifizieren, und dann das LLM verwenden, um diese PHI-Daten zu entfernen.

Im Folgenden finden Sie ein Beispiel für Informationen aus einer Patientenakte, die PHI enthalten:

```
Patient name: John Doe
Patient SSN: 123-34-5678
Patient DOB: 01/01/2024
```

```
Patient address: 123 Main St, Anytown USA
Exam details: good health. Pulse is 60 bpm. needs to work on diet with BMI of 190
```

Im Folgenden finden Sie eine Beispielaufforderung, die die Ergebnisse von Amazon Comprehend Medical als Kontext enthält:

```
<original_text>
Patient name: John Doe
Patient SSN: 123-34-5678 Patient DOB: 01/01/2024
Patient address: 123 Main St, Anytown USA
Exam details: good health. Pulse is 60 bpm. needs to work on diet with BMI of 190
</original_text>

<comprehend_medical_phi_entities>
<entity>
  <text>John Doe</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9967944025993347</score>
  <type>NAME</type>
</entity>
<entity>
  <text>123-34-5678</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9998034834861755</score>
  <type>ID</type>
</entity>
<entity>
  <text>01/01/2000</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9964448809623718</score>
  <type>DATE</type>
</entity>
</comprehend_medical_phi_entities>

<instructions>
Using the provided original text and the Amazon Comprehend Medical PHI entities
 detected, please analyze the text to determine if it contains any additional protected
 health information (PHI) beyond the entities already identified. If additional PHI is
 found, please list and categorize it. If no additional PHI is found, please state that
 explicitly.
In addition if PHI is found, generate updated text with the PHI removed.
</instructions>
```

Verwendung umfangreicher Sprachmodelle für Anwendungsfälle im Gesundheitswesen und in den Biowissenschaften

Hier wird beschrieben, wie Sie große Sprachmodelle (LLMs) für Anwendungen im Gesundheitswesen und in den Biowissenschaften verwenden können. Einige Anwendungsfälle erfordern die Verwendung eines großen Sprachmodells für generative KI-Funktionen. Selbst für die meisten gibt es Vor- und Nachteile state-of-the-art LLMs, und die Empfehlungen in diesem Abschnitt sollen Ihnen helfen, Ihre Zielergebnisse zu erreichen.

Sie können den Entscheidungspfad verwenden, um die geeignete LLM-Lösung für Ihren Anwendungsfall zu ermitteln und dabei Faktoren wie Fachwissen und verfügbare Trainingsdaten zu berücksichtigen. Darüber hinaus werden in diesem Abschnitt beliebte vortrainierte medizinische Fachkräfte LLMs sowie bewährte Verfahren für deren Auswahl und Anwendung beschrieben. Außerdem werden die Kompromisse zwischen komplexen, leistungsstarken Lösungen und einfacheren, kostengünstigeren Ansätzen erörtert.

Anwendungsfälle für ein LLM

Amazon Comprehend Medical kann bestimmte NLP-Aufgaben ausführen. Weitere Informationen finden Sie unter Anwendungsfälle für Amazon Comprehend Medical.

Die logischen und generativen KI-Fähigkeiten eines LLM können für fortgeschrittene Anwendungsfälle im Gesundheitswesen und in den Biowissenschaften erforderlich sein, z. B. für die folgenden:

- Klassifizierung von benutzerdefinierten medizinischen Entitäten oder Textkategorien
- Beantwortung klinischer Fragen
- Zusammenfassung der medizinischen Berichte
- Generierung und Erfassung von Erkenntnissen aus medizinischen Informationen

Anpassungsansätze

Es ist wichtig zu verstehen, wie LLMs sie umgesetzt werden. LLMs werden üblicherweise mit Milliarden von Parametern trainiert, einschließlich Trainingsdaten aus vielen Bereichen. Dieses Training ermöglicht es dem LLM, die meisten allgemeinen Aufgaben zu bewältigen. Herausforderungen treten jedoch häufig auf, wenn domänenspezifisches Wissen erforderlich ist. Beispiele für Fachwissen im Gesundheitswesen und in den Biowissenschaften sind Klinikkodizes,

medizinische Terminologie und Gesundheitsinformationen, die zur Generierung genauer Antworten erforderlich sind. Daher führt die Verwendung des LLM unverändert (Zero-Shot Prompting ohne Ergänzung des Fachwissens) für diese Anwendungsfälle wahrscheinlich zu ungenauen Ergebnissen. Es gibt mehrere beliebte Ansätze, mit denen Sie diese Herausforderung bewältigen können: Prompt Engineering, Retrieval Augmented Generation (RAG) und Feinabstimmung.

Prompt-Engineering

Prompt Engineering ist der Prozess, bei dem Sie generative KI-Lösungen anleiten, um die gewünschten Ergebnisse zu erzielen, indem Sie die Eingaben an das LLM anpassen. Durch die Erstellung präziser Eingabeaufforderungen mit relevantem Kontext ist es möglich, das Modell zur Erledigung spezialisierter Aufgaben im Gesundheitswesen zu führen, die Argumentation erfordern. Effektives Prompt-Engineering kann die Modellleistung für Anwendungsfälle im Gesundheitswesen erheblich verbessern, ohne dass Modelländerungen erforderlich sind. Weitere Informationen zu Prompt Engineering finden Sie unter Implementieren von Advanced Prompt Engineering mit Amazon Bedrock (AWS Blogbeitrag). Few-Shot Prompting und chain-of-thought Prompting sind Techniken, die Sie beim Prompt-Engineering anwenden können.

Few Shot Prompting

Die Eingabeaufforderung ist eine Technik, bei der Sie dem LLM einige Beispiele für die gewünschte Eingabe und Ausgabe zur Verfügung stellen, bevor Sie es bitten, eine ähnliche Aufgabe auszuführen. Im Gesundheitswesen eignet sich dieser Ansatz besonders für spezielle Aufgaben wie die Erkennung medizinischer Entitäten oder die Zusammenfassung klinischer Notizen. Wenn Sie Ihrer Aufforderung 3—5 hochwertige Beispiele hinzufügen, können Sie das Verständnis des Modells für medizinische Terminologie und domänenspezifische Muster erheblich verbessern. Ein Beispiel für Few-Shot-Prompting finden Sie unter Few-Shot Prompting Engineering and Fine-Tuning for LLMs in Amazon Bedrock (Blogbeitrag).AWS

Wenn Sie beispielsweise Medikamentendosierungen aus klinischen Notizen extrahieren, können Sie Beispiele für verschiedene Schreibweisen angeben, anhand derer das Modell Abweichungen in der Art und Weise erkennen kann, wie medizinisches Fachpersonal Verschreibungen dokumentiert. Dieser Ansatz ist besonders effektiv, wenn mit standardisierten Dokumentationsformaten gearbeitet wird oder wenn die Daten konsistente Muster aufweisen.

Chain-of-thought auffordernd

Chain-of-thought Die Aufforderung (CoT) führt den LLM durch einen step-by-step Argumentationsprozess. Dies macht es für komplexe medizinische Entscheidungsunterstützung

und diagnostische Argumentationsaufgaben wertvoll. Indem Sie das Modell ausdrücklich anweisen, bei der Analyse klinischer Szenarien „Schritt für Schritt zu denken“, können Sie seine Fähigkeit verbessern, medizinische Argumentationsprotokolle zu befolgen und Diagnosefehler zu reduzieren.

Diese Technik eignet sich hervorragend, wenn klinisches Denken mehrere logische Schritte erfordert, wie z. B. Differentialdiagnose oder Behandlungsplanung. Dieser Ansatz hat jedoch Einschränkungen, wenn es um hochspezialisiertes medizinisches Wissen geht, das nicht auf den Trainingsdaten des Modells basiert, oder wenn bei Entscheidungen in der Intensivmedizin absolute Präzision erforderlich ist.

In diesen Fällen kann die Kombination von CoT mit einem anderen Ansatz zu besseren Ergebnissen führen. Eine Möglichkeit besteht darin, CoT mit der Aufforderung zur Selbstkonsistenz zu kombinieren. Weitere Informationen finden Sie unter Verbessern der Leistung generativer Sprachmodelle mit Selbstkonsistenzabfragen auf Amazon Bedrock (AWS Blogbeitrag). Eine weitere Option ist die Kombination von Argumentationsstrukturen, wie z. B. ReAct Aufforderungen, mit RAG. Weitere Informationen finden Sie unter Entwickeln fortschrittlicher generativer KI-Assistenten auf Chatbasis mithilfe von RAG und ReAct Prompting (Prescriptive Guidance).AWS

Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) ist 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-System kann medizinische Ontologieinformationen (wie internationale Klassifikationen von Krankheiten, nationale Arzneimittelakten und medizinische Fachüberschriften) aus einer Wissensquelle abrufen. Dies bietet zusätzlichen Kontext für das LLM zur Unterstützung der medizinischen NLP-Aufgabe.

Wie im Kombination von Amazon Comprehend Medical mit großen Sprachmodellen Abschnitt beschrieben, können Sie einen RAG-Ansatz verwenden, um Kontext aus Amazon Comprehend Medical abzurufen. Andere allgemeine Wissensquellen umfassen medizinische Domänendaten, die in einem Datenbankservice wie Amazon OpenSearch Service, Amazon Kendra oder Amazon Aurora gespeichert sind. Das Extrahieren von Informationen aus diesen Wissensquellen kann die Abrufleistung beeinträchtigen, insbesondere bei semantischen Abfragen, die eine Vektordatenbank verwenden.

Eine weitere Option zum Speichern und Abrufen von domänenspezifischem Wissen ist die Verwendung von Amazon Q Business in Ihrem RAG-Workflow. Amazon Q Business kann interne Dokumentenspeicher oder öffentlich zugängliche Websites (wie CMS.gov für ICD-10-Daten)

indizieren. Amazon Q Business kann dann relevante Informationen aus diesen Quellen extrahieren, bevor Ihre Anfrage an das LLM weitergeleitet wird.

Es gibt mehrere Möglichkeiten, einen benutzerdefinierten RAG-Workflow zu erstellen. Beispielsweise gibt es viele Möglichkeiten, Daten aus einer Wissensquelle abzurufen. Der Einfachheit halber empfehlen wir den gängigen Abrufansatz, bei dem Wissen in Form von Einbettungen mithilfe einer Vektordatenbank wie Amazon OpenSearch Service gespeichert wird. Dies erfordert, dass Sie ein Einbettungsmodell verwenden, z. B. einen Satztransformator, um Einbettungen für die Abfrage und für das in der Vektordatenbank gespeicherte Wissen zu generieren.

Weitere Informationen zu vollständig verwalteten und benutzerdefinierten RAG-Ansätzen finden Sie unter Retrieval Augmented Generation-Optionen und Architekturen auf. AWS

Feinabstimmung

Zur Feinabstimmung eines vorhandenen Modells müssen Sie ein LLM-Modell, z. B. ein Amazon Titan-, Mistral- oder Lama-Modell, erstellen und das Modell anschließend an Ihre benutzerdefinierten Daten anpassen. Es gibt verschiedene Techniken zur Feinabstimmung, von denen die meisten die Änderung nur einiger weniger Parameter beinhalten, anstatt alle Parameter im Modell zu ändern. Dies wird als parametereffizientes Feintuning (PEFT) bezeichnet. Weitere Informationen finden Sie unter Hugging Face PEFT on. GitHub

Im Folgenden sind zwei häufige Anwendungsfälle aufgeführt, in denen Sie sich für die Feinabstimmung eines LLM für eine medizinische NLP-Aufgabe entscheiden könnten:

- Generative Aufgabe — Decoder-basierte Modelle führen generative KI-Aufgaben aus. AI/ML Praktiker verwenden Ground-Truth-Daten, um ein bestehendes LLM zu verfeinern. Sie könnten das LLM beispielsweise mithilfe von MedQuAD, einem öffentlichen Datensatz zur Beantwortung medizinischer Fragen, schulen. Wenn Sie eine Abfrage für das fein abgestimmte LLM aufrufen, benötigen Sie keinen RAG-Ansatz, um dem LLM den zusätzlichen Kontext bereitzustellen.
- Einbettungen — Encoder-basierte Modelle erzeugen Einbettungen, indem sie Text in numerische Vektoren umwandeln. Diese auf Encodern basierenden Modelle werden in der Regel als Einbettungsmodelle bezeichnet. Ein Satztransformator-Modell ist eine bestimmte Art von Einbettungsmodell, das für Sätze optimiert ist. Ziel ist es, Einbettungen aus Eingabetext zu generieren. Die Einbettungen werden dann für semantische Analysen oder für Abruftasks verwendet. Zur Feinabstimmung des Einbettungsmodells benötigen Sie einen Korpus an medizinischem Wissen, z. B. Dokumenten, den Sie als Trainingsdaten verwenden können. Dies wird mit Textpaaren erreicht, die auf Ähnlichkeit oder Stimmung basieren, um ein

Satztransformator-Modell zu verfeinern. Weitere Informationen finden Sie unter Training und Feinabstimmung von Einbettungsmodellen mit Sentence Transformers v3 auf Hugging Face.

Sie können Amazon SageMaker Ground Truth verwenden, um einen hochwertigen, beschrifteten Trainingsdatensatz zu erstellen. Sie können die beschrifteten Datensatzausgabe aus Ground Truth verwenden, um Ihre eigenen Modelle zu trainieren. Sie können die Ausgabe auch als Trainingsdatensatz für ein Amazon SageMaker AI-Modell verwenden. Weitere Informationen zur Erkennung benannter Entitäten, zur Textklassifizierung mit einem Etikett und zur Textklassifizierung mit mehreren Bezeichnungen finden Sie unter Textkennzeichnung mit Ground Truth in der Amazon SageMaker AI-Dokumentation.

Weitere Informationen zur Feinabstimmung finden Sie Feinabstimmung großer Sprachmodelle im Gesundheitswesen in diesem Handbuch.

Einen LLM auswählen

Amazon Bedrock ist der empfohlene Ausgangspunkt für die Bewertung leistungsstarker LLMs Produkte. Weitere Informationen finden Sie unter Unterstützte Foundation-Modelle in Amazon Bedrock. Sie können Modellevaluierungsjobs in Amazon Bedrock verwenden, um die Ergebnisse mehrerer Ausgaben zu vergleichen und dann das Modell auszuwählen, das für Ihren Anwendungsfall am besten geeignet ist. Weitere Informationen finden Sie unter Wählen Sie das Modell mit der besten Leistung anhand von Amazon Bedrock-Bewertungen in der Amazon Bedrock-Dokumentation.

Einige LLMs verfügen nur über begrenzte Schulungen zu medizinischen Daten. Wenn Ihr Anwendungsfall die Feinabstimmung eines LLM oder eines LLM erfordert, das Amazon Bedrock nicht unterstützt, sollten Sie die Verwendung von Amazon AI in Betracht ziehen. SageMaker In SageMaker KI können Sie ein fein abgestimmtes LLM verwenden oder ein benutzerdefiniertes LLM wählen, das auf medizinischen Domänendaten trainiert wurde.

In der folgenden Tabelle sind beliebte Unternehmen aufgeführt LLMs , die auf Daten aus dem medizinischen Bereich trainiert wurden.

LLM	Aufgaben	Wissen	Architektur
BioBert	Informationsabruf, Textklassifizierung und Erkennung benannter Entitäten	Zusammenfassungen von PubMed, Volltextartikel von und PubMedCen	Encoder

		tral allgemeines Fachwissen	
Klinik Albert	Informationsabruf, Textklassifizierung und Erkennung benannter Entitäten	Großer, multizentrischer Datensatz zusammen mit über 3.000.000 Patientenakten aus elektronischen Patientendatensystemen (EHR)	Encoder
Klinisches GPT	Zusammenfassung, Beantwortung von Fragen und Textgenerierung	Umfangreiche und vielfältige medizinische Datensätze, darunter Krankenakten, fachspezifisches Wissen und Konsultationen in mehreren Gesprächsrunden	Decoder
GatorTron-GO	Zusammenfassung, Beantwortung von Fragen, Textgenerierung und Informationsabruf	Klinische Hinweise und biomedizinische Literatur	Encoder
Med-Bert	Informationsabruf, Textklassifizierung und Erkennung benannter Entitäten	Großer Datensatz mit medizinischen Texten, klinischen Notizen, Forschungsarbeiten und Dokumenten zum Gesundheitswesen	Encoder
Med-Palm	Beantwortung von Fragen für medizinische Zwecke	Datensätze mit medizinischem und biomedizinischem Text	Decoder

Medaille Alpaca	Aufgaben zur Beantwortung von Fragen und zum medizinischen Dialog	Eine Vielzahl von medizinischen Texten, die Ressourcen wie medizinische Karteikarten, Wikis und Dialogdatensätze umfassen	Decoder
BioMedbert	Informationsabruf, Textklassifizierung und Erkennung benannter Entitäten	Ausschließlich Kurzfassungen PubMed und Volltextartikel von PubMedCentral	Encoder
BioMedLM	Zusammenfassung, Beantwortung von Fragen und Textgenerierung	Biomedizinische Literatur aus Wissensquellen PubMed	Decoder

Im Folgenden finden Sie bewährte Methoden für den Einsatz von vortrainierten Ärzten: LLMs

- Machen Sie sich mit den Trainingsdaten und ihrer Relevanz für Ihre medizinische NLP-Aufgabe vertraut.
- Identifizieren Sie die LLM-Architektur und ihren Zweck. Encoder eignen sich für Einbettungen und NLP-Aufgaben. Decoder sind für Generierungsaufgaben vorgesehen.
- Evaluieren Sie die Infrastruktur-, Leistungs- und Kostenanforderungen für die Durchführung des vortrainierten medizinischen LLM.
- Wenn eine Feinabstimmung erforderlich ist, stellen Sie sicher, dass die Trainingsdaten korrekt sind. Stellen Sie sicher, dass Sie alle persönlich identifizierbaren Informationen (PII) oder geschützten Gesundheitsinformationen (PHI) maskieren oder unkenntlich machen.

Reale medizinische NLP-Aufgaben können sich LLMs in Bezug auf Wissen oder beabsichtigte Anwendungsfälle von vortrainierten Aufgaben unterscheiden. Wenn ein domänenspezifisches LLM Ihre Bewertungsmaßstäbe nicht erfüllt, können Sie ein LLM mit Ihrem eigenen Datensatz verfeinern oder ein neues Basismodell trainieren. Die Schulung eines neuen Basismodells ist ein ehrgeiziges

und oft teures Unterfangen. Für die meisten Anwendungsfälle empfehlen wir, ein vorhandenes Modell zu verfeinern.

Bei der Verwendung oder Feinabstimmung eines vortrainierten medizinischen LLMs ist es wichtig, die Infrastruktur, die Sicherheit und die Leitplanken zu berücksichtigen.

Infrastruktur

Im Vergleich zur Nutzung von Amazon Bedrock für On-Demand- oder Batch-Inferenzen erfordert das Hosten von vortrainierten medizinischen LLMs (üblicherweise von Hugging Face) erhebliche Ressourcen. Um vortrainierte medizinische LLMs zu hosten, wird üblicherweise ein Amazon SageMaker AI-Image verwendet, das auf einer Amazon Elastic Compute Cloud (Amazon EC2) -Instance mit einer oder mehreren Instances ausgeführt wird GPUs, z. B. ml.g5-Instances für beschleunigtes Rechnen oder ml.inf2-Instances für. AWS Inferentia Das liegt daran, dass sie eine große Menge an Arbeitsspeicher und Festplattenspeicher LLMs verbrauchen.

Sicherheit und Leitplanken

Je nach Ihren geschäftlichen Compliance-Anforderungen sollten Sie erwägen, Amazon Comprehend und Amazon Comprehend Medical zu verwenden, um personenbezogene Daten (PII) und geschützte Gesundheitsinformationen (PHI) aus Trainingsdaten zu maskieren oder zu redigieren. Dadurch wird verhindert, dass das LLM vertrauliche Daten verwendet, wenn es Antworten generiert.

Wir empfehlen Ihnen, Vorurteile, Fairness und Halluzinationen in Ihren generativen KI-Anwendungen zu berücksichtigen und zu bewerten. Unabhängig davon, ob Sie ein bereits vorhandenes LLM verwenden oder eines optimieren, implementieren Sie Leitplanken, um schädliche Reaktionen zu verhindern. Guardrails sind Schutzmaßnahmen, die Sie an Ihre generativen KI-Anwendungsanforderungen und verantwortungsvollen KI-Richtlinien anpassen. Sie können beispielsweise Amazon Bedrock Guardrails verwenden.

Feinabstimmung großer Sprachmodelle im Gesundheitswesen

Der in diesem Abschnitt beschriebene Ansatz zur Feinabstimmung unterstützt die Einhaltung ethischer und regulatorischer Richtlinien und fördert den verantwortungsvollen Einsatz von KI-Systemen im Gesundheitswesen. Es wurde entwickelt, um genaue und vertrauliche Erkenntnisse zu generieren. Generative KI revolutioniert die Gesundheitsversorgung, aber off-the-shelf Modelle sind in klinischen Umgebungen, in denen Genauigkeit entscheidend ist und Compliance nicht verhandelbar ist, oft unzureichend. Die Feinabstimmung von Basismodellen mit domänenspezifischen Daten schließt diese Lücke. Es hilft Ihnen dabei, KI-Systeme zu entwickeln, die die Sprache der

Medizin sprechen und gleichzeitig strenge regulatorische Standards einhalten. Der Weg zu einer erfolgreichen Feinabstimmung erfordert jedoch eine sorgfältige Bewältigung der einzigartigen Herausforderungen des Gesundheitswesens: Schutz sensibler Daten, Rechtfertigung von KI-Investitionen mit messbaren Ergebnissen und Wahrung der klinischen Relevanz in einem sich schnell entwickelnden medizinischen Umfeld.

Wenn leichtere Ansätze an ihre Grenzen stoßen, wird die Feinabstimmung zu einer strategischen Investition. Es wird davon ausgegangen, dass die Gewinne an Genauigkeit, Latenz oder betrieblicher Effizienz die erheblichen Rechen- und Engineering-Kosten ausgleichen werden. Es ist wichtig, sich daran zu erinnern, dass der Fortschritt bei Basismodellen schnell voranschreitet, sodass der Vorteil eines fein abgestimmten Modells möglicherweise nur bis zur nächsten großen Modellversion anhält.

In diesem Abschnitt wird die Diskussion anhand der folgenden zwei wichtigen Anwendungsfälle von Kunden aus dem AWS Gesundheitswesen behandelt:

- Systeme zur Unterstützung klinischer Entscheidungen — Verbessern Sie die diagnostische Genauigkeit durch Modelle, die komplexe Patientengeschichten verstehen und sich weiterentwickelnde Richtlinien entwickeln. Durch eine Feinabstimmung können Modelle dazu beitragen, komplexe Patientengeschichten besser zu verstehen und spezielle Richtlinien zu integrieren. Dadurch können Fehler bei der Modellvorhersage potenziell reduziert werden. Sie müssen diese Vorteile jedoch gegen die Kosten für Schulungen zu großen, sensiblen Datensätzen und der Infrastruktur abwägen, die für anspruchsvolle klinische Anwendungen erforderlich ist. Rechtfertigen die verbesserte Genauigkeit und die verbesserte Kontextsensitivität die Investition, insbesondere wenn häufig neue Modelle auf den Markt kommen?
- Analyse medizinischer Dokumente — Automatisieren Sie die Verarbeitung von klinischen Notizen, bildgebenden Berichten und Versicherungsdokumenten und wahren Sie gleichzeitig die Einhaltung des Health Insurance Portability and Accountability Act (HIPAA). Hier kann das Modell durch eine Feinabstimmung möglicherweise in der Lage sein, einzigartige Formate, spezielle Abkürzungen und regulatorische Anforderungen effektiver zu handhaben. Der Vorteil liegt häufig in einer Verkürzung der Zeit für manuelle Prüfungen und einer verbesserten Einhaltung von Vorschriften. Dennoch ist es wichtig zu beurteilen, ob diese Verbesserungen erheblich genug sind, um die Ressourcen für die Feinabstimmung zu rechtfertigen. Finden Sie heraus, ob zeitnahes Engineering und Workflow-Orchestrierung Ihren Anforderungen gerecht werden können.

Diese realen Szenarien veranschaulichen den Prozess der Feinabstimmung, von den ersten Experimenten bis zur Implementierung des Modells, und berücksichtigen gleichzeitig die individuellen Anforderungen des Gesundheitswesens in jeder Phase.

Schätzung der Kosten und der Kapitalrendite

Die folgenden Kostenfaktoren müssen Sie bei der Feinabstimmung eines LLM berücksichtigen:

- Modellgröße — Bei größeren Modellen ist die Feinabstimmung teurer
- Datensatzgröße — Die Rechenkosten und der Zeitaufwand steigen mit der Größe des Datensatzes für die Feinabstimmung
- Strategie zur Feinabstimmung — Parametereffiziente Methoden können die Kosten im Vergleich zu vollständigen Parameteraktualisierungen reduzieren

Berücksichtigen Sie bei der Berechnung der Investitionsrendite (ROI) die Verbesserung der von Ihnen ausgewählten Kennzahlen (z. B. Genauigkeit), multipliziert mit dem Volumen der Anfragen (wie oft das Modell verwendet wird) und der erwarteten Dauer, bis das Modell von neueren Versionen übertroffen wird.

Berücksichtigen Sie auch die Lebensdauer Ihres Basis-LLMs. Alle 6—12 Monate kommen neue Basismodelle auf den Markt. Wenn die Feinabstimmung und Validierung Ihres Detektors für seltene Krankheiten 8 Monate in Anspruch nimmt, erhalten Sie möglicherweise nur 4 Monate überragende Leistung, bevor neuere Modelle die Lücke schließen.

Durch die Berechnung der Kosten, des ROI und der potenziellen Lebensdauer für Ihren Anwendungsfall können Sie eine datengestützte Entscheidung treffen. Wenn beispielsweise die Feinabstimmung Ihres Modells zur Unterstützung klinischer Entscheidungen zu einer messbaren Reduzierung von Diagnosefehlern bei Tausenden von Fällen pro Jahr führt, kann sich die Investition schnell auszahlen. Umgekehrt kann es ratsam sein, mit der Feinabstimmung zu warten, bis die nächste Generation von Modellen verfügbar ist, wenn Ihr Dokumentenanalyse-Workflow allein schon durch schnelle technische Umsetzung Ihrer Zielgenauigkeit erreicht wird.

Feinabstimmung ist es nicht. one-size-fits-all Wenn Sie sich für eine Feinabstimmung entscheiden, hängt der richtige Ansatz von Ihrem Anwendungsfall, Ihren Daten und Ressourcen ab.

Wahl einer Strategie zur Feinabstimmung

Nachdem Sie festgestellt haben, dass die Feinabstimmung der richtige Ansatz für Ihren Anwendungsfall im Gesundheitswesen ist, besteht der nächste Schritt darin, die am besten geeignete Feinabstimmungsstrategie auszuwählen. Es stehen mehrere Ansätze zur Verfügung. Jeder hat unterschiedliche Vorteile und Kompromisse für Anwendungen im Gesundheitswesen. Die Wahl

zwischen diesen Methoden hängt von Ihren spezifischen Zielen, den verfügbaren Daten und Ihren Ressourcenbeschränkungen ab.

Ziele der Schulung

Beim domänenadaptiven Vortraining (DAPT) handelt es sich um eine unbeaufsichtigte Methode, bei der das Modell anhand einer großen Menge domänenspezifischen, unbeschrifteten Textes (z. B. Millionen von medizinischen Dokumenten) vorab trainiert wird. Dieser Ansatz eignet sich hervorragend zur Verbesserung der Fähigkeit der Modelle, Abkürzungen für medizinische Fachgebiete und die von Radiologen, Neurologen und anderen spezialisierten Anbietern verwendete Terminologie zu verstehen. DAPT erfordert jedoch riesige Datenmengen und ist nicht auf bestimmte Aufgaben zugeschnitten.

Supervised Fine-Tuning (SFT) bringt dem Modell anhand strukturierter Input-Output-Beispiele bei, explizite Anweisungen zu befolgen. Dieser Ansatz eignet sich hervorragend für Workflows zur Analyse medizinischer Dokumente, z. B. für die Zusammenfassung von Dokumenten oder die klinische Kodierung. Die Befehlsoptimierung ist eine gängige Form von SFT, bei der das Modell anhand von Beispielen trainiert wird, die explizite Anweisungen mit den gewünschten Ergebnissen kombinieren. Dies verbessert die Fähigkeit des Modells, verschiedene Benutzeranweisungen zu verstehen und zu befolgen. Diese Technik ist im Gesundheitswesen besonders wertvoll, da sie das Modell anhand spezifischer klinischer Beispiele trainiert. Der Hauptnachteil besteht darin, dass dafür sorgfältig beschriftete Beispiele erforderlich sind. Darüber hinaus könnte das fein abgestimmte Modell Probleme mit Grenzfällen haben, für die es keine Beispiele gibt. Eine Anleitung zur Feinabstimmung mit Amazon SageMaker Jumpstart finden Sie unter Anleitung zur Feinabstimmung für FLAN T5 XL mit Amazon SageMaker Jumpstart (Blogbeitrag).AWS

Reinforcement Learning from Human Feedback (RLHF) optimiert das Modellverhalten auf der Grundlage von Expertenfeedback und Präferenzen. Verwenden Sie ein Belohnungsmodell, das auf menschlichen Präferenzen und Methoden wie Proximal Policy Optimization (PPO) oder Direct Preference Optimization (DPO) trainiert wurde, um das Modell zu optimieren und gleichzeitig zerstörerische Aktualisierungen zu verhindern. RLHF ist ideal, um die Ergebnisse mit den klinischen Leitlinien in Einklang zu bringen und sicherzustellen, dass die Empfehlungen im Rahmen der genehmigten Protokolle bleiben. Dieser Ansatz erfordert viel Zeit für Rückmeldungen durch Ärzte und beinhaltet eine komplexe Trainingspipeline. RLHF ist jedoch im Gesundheitswesen besonders wertvoll, da es medizinischen Experten hilft, die Art und Weise zu gestalten, wie KI-Systeme kommunizieren und Empfehlungen aussprechen. So können Ärzte beispielsweise Feedback geben, um sicherzustellen, dass das Modell eine angemessene Art und Weise am Krankenbett beibehält, weiß, wann Unsicherheiten geäußert werden müssen, und dass es die klinischen Richtlinien einhält.

Techniken wie PPO optimieren das Modellverhalten iterativ auf der Grundlage von Expertenfeedback und schränken gleichzeitig die Aktualisierung der Parameter ein, um medizinisches Kernwissen zu erhalten. Auf diese Weise können Modelle komplexe Diagnosen in einer patientenfreundlichen Sprache vermitteln und gleichzeitig schwerwiegende Erkrankungen für eine sofortige medizinische Behandlung kennzeichnen. Dies ist entscheidend für das Gesundheitswesen, wo es sowohl auf Genauigkeit als auch auf den Kommunikationsstil ankommt. Weitere Informationen zu RLHF finden Sie unter Feinabstimmung umfangreicher Sprachmodelle mit verstärkendem Lernen anhand von menschlichem oder künstlichem Feedback (AWS Blogbeitrag).

Methoden der Implementierung

Ein vollständiges Parameter-Update beinhaltet die Aktualisierung aller Modellparameter während des Trainings. Dieser Ansatz eignet sich am besten für Systeme zur Unterstützung klinischer Entscheidungen, die eine umfassende Integration von Patientenanamnese, Laborergebnissen und sich weiterentwickelnden Richtlinien erfordern. Zu den Nachteilen gehören hohe Rechenkosten und das Risiko einer Überanpassung, wenn Ihr Datensatz nicht umfangreich und vielfältig ist.

Bei Methoden zur parametereffizienten Feinabstimmung (PEFT) wird nur eine Teilmenge von Parametern aktualisiert, um eine Überanpassung oder einen katastrophalen Verlust von Sprachkenntnissen zu verhindern. Zu den Typen gehören Low-Rank Adaptation (LoRa), Adapter und Präfix-Tuning. PEFT-Methoden bieten geringere Rechenkosten, schnellere Schulungen und eignen sich hervorragend für Experimente wie die Anpassung eines Modells zur Unterstützung klinischer Entscheidungen an die Protokolle oder Terminologie eines neuen Krankenhauses. Die größte Einschränkung ist die potenziell verringerte Leistung im Vergleich zu vollständigen Parameteraktualisierungen.

Weitere Informationen zu Feinabstimmungsmethoden finden Sie unter Erweiterte Feinabstimmungsmethoden auf Amazon SageMaker AI (AWS Blogbeitrag).

Einen Datensatz zur Feinabstimmung erstellen

Die Qualität und Vielfalt des Datensatzes zur Feinabstimmung ist entscheidend für die Leistung, Sicherheit und Vermeidung von Verzerrungen von Modellen. Die folgenden drei wichtigen Bereiche sollten bei der Erstellung dieses Datensatzes berücksichtigt werden:

- Das Volumen basiert auf einem Feinabstimmungsansatz
- Datenanmerkung von einem Fachexperten
- Vielfalt des Datensatzes

Wie in der folgenden Tabelle dargestellt, variieren die Anforderungen an die Datensatzgröße für die Feinabstimmung je nach Art der durchgeführten Feinabstimmung.

Strategie für die Feinabstimmung	Größe des Datensatzes
An die Domäne angepasste Vorschulung	Über 100.000 Domain-Texte
Beaufsichtigte Feinabstimmung	Über 10.000 beschriftete Paare
Verstärktes Lernen aus menschlichem Feedback	Präferenzpaare für mehr als 1.000 Experten

Sie können Amazon EMR und Amazon SageMaker Data Wrangler verwenden AWS Glue, um den Datenextraktions - und Transformationsprozess zu automatisieren, um einen Datensatz zu kuratieren, den Sie besitzen. Wenn Sie nicht in der Lage sind, einen ausreichend großen Datensatz zu kuratieren, können Sie Datensätze finden und direkt in Ihren Browser herunterladen. AWS-Konto AWS Data Exchange Konsultieren Sie Ihren Rechtsbeistand, bevor Sie Datensätze von Drittanbietern verwenden.

Erfahrene Annotatoren mit Fachkenntnissen wie Ärzte, Biologen und Chemiker sollten Teil des Datenkurationsprozesses sein, um die Nuancen medizinischer und biologischer Daten in die Modellausgabe einfließen zu lassen. Amazon SageMaker Ground Truth bietet eine Low-Code-Benutzeroberfläche, über die Experten den Datensatz kommentieren können.

Ein Datensatz, der die menschliche Bevölkerung repräsentiert, ist für die Feinabstimmung von Anwendungsfällen im Gesundheitswesen und in den Biowissenschaften unerlässlich, um Verzerrungen zu vermeiden und reale Ergebnisse widerzuspiegeln. AWS Glue interaktive Sitzungen oder SageMaker Amazon-Notebook-Instances bieten eine leistungsstarke Möglichkeit, Datensätze iterativ zu untersuchen und Transformationen mithilfe von Jupyter-kompatiblen Notebooks zu optimieren. Interaktive Sitzungen ermöglichen es Ihnen, mit einer Auswahl beliebter integrierter Entwicklungsumgebungen () in Ihrer lokalen Umgebung zu arbeiten. IDEs Alternativ können Sie mit AWS Glue oder Amazon SageMaker Studio-Notizbüchern über die arbeiten AWS-Managementkonsole.

Feinabstimmung des Modells

AWS bietet Dienste wie Amazon SageMaker AI und Amazon Bedrock, die für eine erfolgreiche Feinabstimmung entscheidend sind.

SageMaker KI ist ein vollständig verwalteter Service für maschinelles Lernen, der Entwicklern und Datenwissenschaftlern hilft, ML-Modelle schnell zu erstellen, zu trainieren und bereitzustellen. Zu den drei nützlichen Funktionen von SageMaker KI für die Feinabstimmung gehören:

- SageMakerSchulung — Eine vollständig verwaltete ML-Funktion, mit der Sie eine Vielzahl von Modellen effizient und in großem Maßstab trainieren können
- SageMaker JumpStart— Eine Funktion, die auf SageMaker Trainingsaufgaben aufbaut und vortrainierte Modelle, integrierte Algorithmen und Lösungsvorlagen für ML-Aufgaben bereitstellt
- SageMaker HyperPod— Eine speziell entwickelte Infrastrukturlösung für die verteilte Schulung von Basismodellen und LLMs

Amazon Bedrock ist ein vollständig verwalteter Service, der über eine API Zugriff auf leistungsstarke Fundamentmodelle mit integrierten Sicherheits-, Datenschutz- und Skalierbarkeitsfunktionen bietet. Der Service bietet die Möglichkeit, mehrere verfügbare Basismodelle zu verfeinern. Weitere Informationen finden Sie in der Amazon Bedrock-Dokumentation unter Unterstützte Modelle und Regionen zur Feinabstimmung und weiteren Vorschulung.

Bei der Feinabstimmung mit einem der beiden Services sollten Sie das Basismodell, die Feinabstimmungsstrategie und die Infrastruktur berücksichtigen.

Wahl des Basismodells

Closed-Source-Modelle wie Anthropic Claude, Meta Llama und Amazon Nova bieten eine starke out-of-the-box Leistung mit verwalteter Compliance, beschränken aber die Flexibilität bei der Feinabstimmung auf vom Anbieter unterstützte Optionen wie Amazon Bedrock. APIs Dies schränkt die Anpassungsfähigkeit ein, insbesondere für regulierte Anwendungsfälle im Gesundheitswesen. Im Gegensatz dazu bieten Open-Source-Modelle wie Meta Llama die volle Kontrolle und Flexibilität über die Amazon SageMaker AI-Services hinweg und eignen sich daher ideal, wenn Sie ein Modell an Ihre spezifischen Daten- oder Workflow-Anforderungen anpassen, prüfen oder tiefgreifend anpassen müssen.

Feinabstimmung der Strategie

Die einfache Anpassung der Anweisungen kann über Amazon Bedrock Model Customization oder Amazon SageMaker JumpStart vorgenommen werden. Komplexe PEFT-Ansätze, wie LoRa oder Adapter, erfordern SageMaker Schulungsaufgaben oder benutzerdefinierte Feinabstimmungsfunktionen in Amazon Bedrock. Verteilte Schulungen für sehr große Modelle werden unterstützt von. SageMaker HyperPod

Skalierung und Kontrolle der Infrastruktur

Vollständig verwaltete Services wie Amazon Bedrock minimieren das Infrastrukturmanagement und eignen sich ideal für Unternehmen, die Wert auf Benutzerfreundlichkeit und Compliance legen. Teilweise verwaltete Optionen, wie z. B. SageMaker JumpStart, bieten eine gewisse Flexibilität bei geringerer Komplexität. Diese Optionen eignen sich für schnelles Prototyping oder für die Verwendung vorgefertigter Workflows. Vollständige Kontrolle und Anpassung sind mit SageMaker Schulungsaufträgen verbunden. Diese erfordern HyperPod jedoch mehr Fachwissen und eignen sich am besten, wenn Sie für große Datenmengen skalieren müssen oder benutzerdefinierte Pipelines benötigen.

Überwachung fein abgestimmter Modelle

Im Gesundheitswesen und in den Biowissenschaften erfordert die Überwachung der LLM-Feinabstimmung die Überwachung mehrerer wichtiger Leistungsindikatoren. Genauigkeit ist eine Basismessung, die jedoch gegen Präzision und Wiederauffindbarkeit abgewogen werden muss, insbesondere bei Anwendungen, bei denen Fehlklassifizierungen erhebliche Folgen haben. Der F1-Score hilft dabei, Probleme des Klassenungleichgewichts zu lösen, die in medizinischen Datensätzen häufig vorkommen können. Weitere Informationen finden Sie unter Evaluierung LLMs für Anwendungen im Gesundheitswesen und in den Biowissenschaften in diesem Handbuch.

Mithilfe von Kalibrierungsmetriken können Sie sicherstellen, dass die Konfidenzniveaus des Modells den realen Wahrscheinlichkeiten entsprechen. Fairness-Metriken können Ihnen dabei helfen, potenzielle Verzerrungen bei verschiedenen demografischen Merkmalen der Patienten zu erkennen.

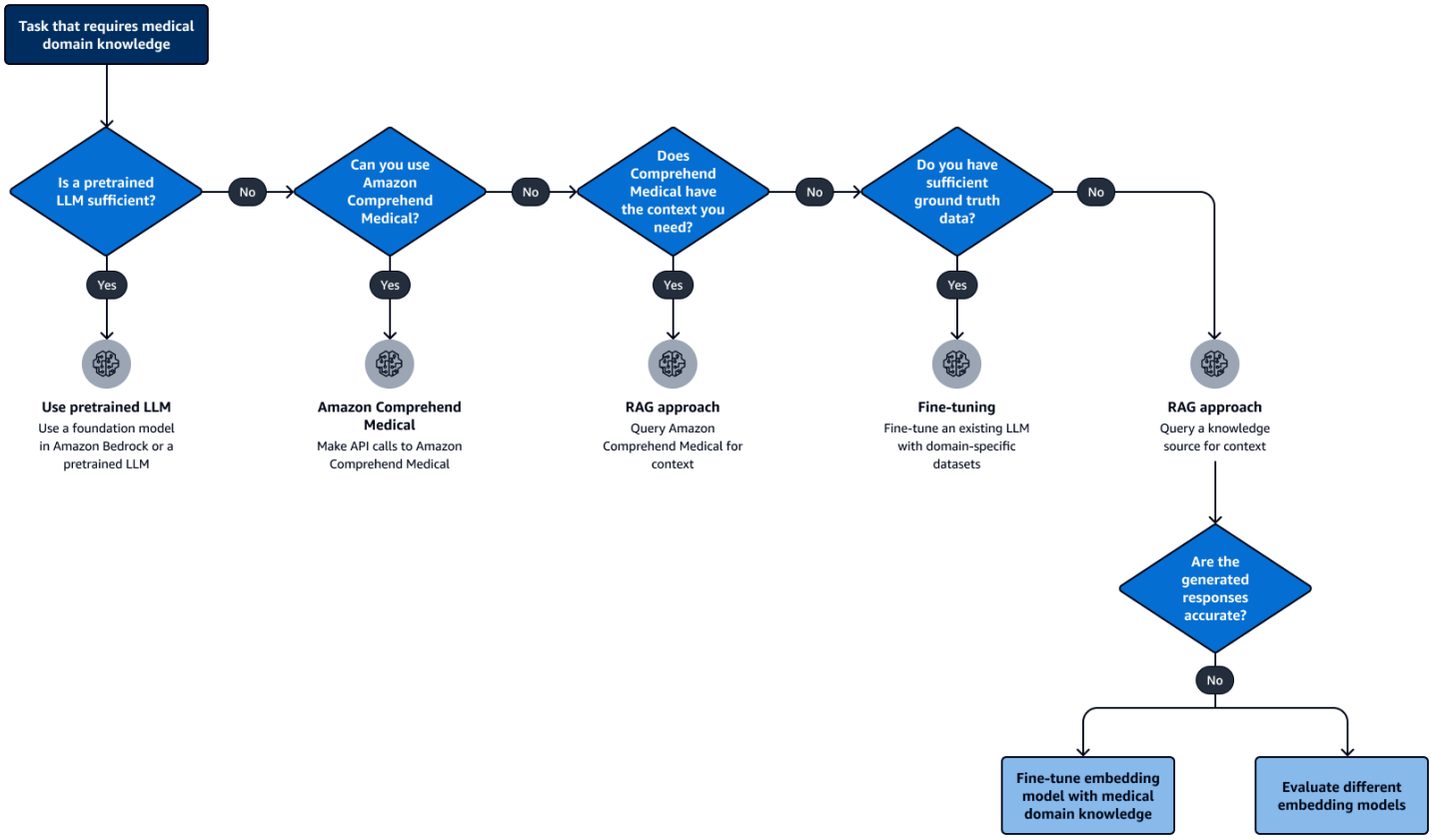
MLflowist eine Open-Source-Lösung, mit der Sie Feinabstimmungsexperimente nachverfolgen können. MLflow wird von Amazon SageMaker AI nativ unterstützt, sodass Sie Metriken von Trainingsläufen visuell vergleichen können. Für die Feinabstimmung von Jobs auf Amazon Bedrock werden Metriken an Amazon gestreamt, CloudWatch sodass Sie die Metriken in der Konsole visualisieren können. CloudWatch

Wahl eines NLP-Ansatzes für das Gesundheitswesen und die Biowissenschaften

In Generative KI- und NLP-Ansätze für das Gesundheitswesen und die Biowissenschaften diesem Abschnitt werden die folgenden Ansätze zur Bewältigung von Aufgaben der Verarbeitung natürlicher Sprache (NLP) für Anwendungen im Gesundheitswesen und in den Biowissenschaften beschrieben:

- Verwenden von Amazon Comprehend Medical
- Kombination von Amazon Comprehend Medical mit einem LLM in einem Retrieval Augment Generation (RAG) -Workflow
- Verwendung eines fein abgestimmten LLM
- Verwendung eines RAG-Workflows

Indem Sie die bekannten Einschränkungen von LLMs Aufgaben im medizinischen Bereich und Ihren Anwendungsfall abwägen, können Sie entscheiden, welcher Ansatz für Ihre Aufgabe am besten geeignet ist. Der folgende Entscheidungsbaum kann Ihnen bei der Auswahl eines LLM-Ansatzes für Ihre medizinische NLP-Aufgabe helfen:



Das Diagramm zeigt den folgenden Workflow:

1. Identifizieren Sie für Anwendungsfälle im Gesundheitswesen und in den Biowissenschaften, ob für die NLP-Aufgabe spezifisches Fachwissen erforderlich ist. Stimmen Sie sich bei Bedarf mit Fachexperten ab (SMEs).
2. Wenn Sie ein allgemeines LLM oder ein Modell verwenden können, das anhand medizinischer Datensätze trainiert wurde, verwenden Sie ein verfügbares Basismodell in Amazon Bedrock oder das vortrainierte LLM. Weitere Informationen finden Sie unter Einen LLM auswählen in diesem Handbuch.
3. Wenn die Funktionen zur Erkennung von Entitäten und zur Verknüpfung von Ontologien von Amazon Comprehend Medical auf Ihren Anwendungsfall zugeschnitten sind, verwenden Sie Amazon Comprehend Medical. APIs Weitere Informationen finden Sie unter Verwenden von Amazon Comprehend Medical in diesem Handbuch.
4. Manchmal verfügt Amazon Comprehend Medical über den erforderlichen Kontext, unterstützt Ihren Anwendungsfall jedoch nicht. Beispielsweise benötigen Sie möglicherweise unterschiedliche Entitätsdefinitionen, erhalten eine überwältigende Anzahl von Ergebnissen, benötigen benutzerdefinierte Entitäten oder benötigen eine benutzerdefinierte NLP-Aufgabe. Wenn dies

der Fall ist, verwenden Sie einen RAG-Ansatz, um Amazon Comprehend Medical nach Kontext abzufragen. Weitere Informationen finden Sie unter Kombination von Amazon Comprehend Medical mit großen Sprachmodellen in diesem Handbuch.

5. Wenn Sie über eine ausreichende Menge an Ground-Truth-Daten verfügen, optimieren Sie ein vorhandenes LLM. Weitere Informationen finden Sie unter Anpassungsansätze in diesem Handbuch.
6. Wenn die anderen Ansätze nicht den medizinischen Zielen Ihrer NLP-Aufgaben entsprechen, implementieren Sie eine RAG-Lösung. Weitere Informationen finden Sie unter Anpassungsansätze in diesem Handbuch.
7. Prüfen Sie nach der Implementierung der RAG-Lösung, ob die generierten Antworten korrekt sind. Weitere Informationen finden Sie unter Evaluierung LLMs für Anwendungen im Gesundheitswesen und in den Biowissenschaften in diesem Handbuch. Es ist üblich, mit einem Amazon Titan Text Embeddings-Modell oder einem allgemeinen Satztransformator-Modell wie All-MiniLM-L6-V2 zu beginnen. Aufgrund des fehlenden Domänenkontextes erfassen diese Modelle jedoch möglicherweise nicht die medizinische Terminologie des Textes. Falls erforderlich, sollten Sie die folgenden Anpassungen in Betracht ziehen:
   a. Evaluieren Sie andere Einbettungsmodelle
   b. Optimieren Sie das Einbettungsmodell mit domänenspezifischen Datensätzen

Überlegungen zur Geschäftsreife

Bei der Anpassung von LLM-Lösungen für Anwendungen im Gesundheitswesen und in den Biowissenschaften ist die Geschäftsreife entscheidend. Diese Unternehmen sind bei der Implementierung LLMs je nach ihren Akzeptanzkriterien mit unterschiedlicher Komplexität konfrontiert. Häufig investieren Unternehmen, denen es an AI/ML Ressourcen mangelt, in die Unterstützung von Auftragnehmern, um LLM-Lösungen zu entwickeln. In diesen Situationen ist es wichtig, die folgenden Kompromisse zu verstehen:

- Hohe Leistung bei hohen Kosten und hohem Wartungsaufwand — Möglicherweise benötigen Sie eine komplexe Lösung, die fein abgestimmt oder kundenspezifisch angepasst ist, um strenge Leistungsstandards LLMs zu erfüllen. Dies ist jedoch mit höheren Kosten und Wartungsanforderungen verbunden. Möglicherweise müssen Sie spezialisierte Ressourcen einstellen oder mit Auftragnehmern zusammenarbeiten, um diese ausgeklügelten Lösungen zu warten. Dies kann die Entwicklung potenziell verlangsamen.

- Gute Leistung bei niedrigen Kosten und geringem Wartungsaufwand — Alternativ stellen Sie möglicherweise fest, dass Dienste wie Amazon Bedrock oder Amazon Comprehend Medical eine akzeptable Leistung bieten. Mit diesen LLMs Methoden können zwar perfekte Ergebnisse erzielt werden, mit diesen Lösungen lassen sich jedoch häufig gleichbleibende, qualitativ hochwertige Ergebnisse erzielen. Diese Lösungen sind kostengünstiger und reduzieren den Wartungsaufwand. Dies kann die Entwicklung beschleunigen.

Wenn ein einfacherer, kostengünstigerer Ansatz durchweg zu qualitativ hochwertigen Ergebnissen führt, die Ihren Akzeptanzkriterien entsprechen, sollten Sie überlegen, ob die Steigerung der Leistung die Kompromisse bei Kosten, Wartung und Zeit wert ist. Wenn die einfachere Lösung jedoch deutlich hinter der angestrebten Leistung zurückbleibt und Ihrem Unternehmen die Investitionskapazität für komplexe Lösungen und deren Wartungsanforderungen fehlt, sollten Sie erwägen, die AI/ML Entwicklung zu verschieben, bis mehr Ressourcen oder alternative Lösungen verfügbar sind.

Darüber hinaus empfehlen wir Ihnen für jede medizinische NLP-Lösung, die auf einem LLM basiert, eine kontinuierliche Überwachung und Bewertung durchzuführen. Beurteilen Sie das Feedback der Benutzer im Laufe der Zeit und führen Sie regelmäßige Bewertungen durch, um sicherzustellen, dass die Lösung auch weiterhin Ihren Geschäftszielen entspricht.

Evaluierung LLMs für Anwendungen im Gesundheitswesen und in den Biowissenschaften

Dieser Abschnitt bietet einen umfassenden Überblick über die Anforderungen und Überlegungen zur Bewertung umfangreicher Sprachmodelle (LLMs) in Anwendungsfällen im Gesundheitswesen und in den Biowissenschaften.

Es ist wichtig, Ground-Truth-Daten und Feedback von KMU zu verwenden, um Verzerrungen zu vermeiden und die Genauigkeit der vom LLM generierten Antworten zu überprüfen. In diesem Abschnitt werden bewährte Verfahren für die Erfassung und Kuratierung von Schulungs- und Testdaten beschrieben. Es hilft Ihnen auch dabei, Leitplanken zu implementieren und Datenverzerrungen und Fairness zu messen. Außerdem werden die häufigsten Aufgaben der medizinischen Verarbeitung natürlicher Sprache (NLP) wie Textklassifizierung, Erkennung benannter Entitäten und Textgenerierung sowie die damit verbundenen Bewertungsmetriken behandelt.

Außerdem werden Arbeitsabläufe für die Durchführung der LLM-Evaluierung während der Trainingsexperimentierphase und der Phase nach der Produktion vorgestellt. Die Modellüberwachung und der LLM-Betrieb sind wichtige Elemente dieses Bewertungsprozesses.

Trainings- und Testdaten für medizinische NLP-Aufgaben

Bei medizinischen NLP-Aufgaben werden häufig medizinische Korpora (z. B. PubMed) oder Patienteninformationen (z. B. Notizen zu Krankenhausbesuchen) verwendet, um Erkenntnisse zu klassifizieren, zusammenzufassen und Erkenntnisse zu gewinnen. Medizinisches Personal, wie Ärzte, Gesundheitsverwalter oder Techniker, unterscheidet sich in Bezug auf Fachwissen und Sichtweisen. Aufgrund der Subjektivität zwischen diesen medizinischen Fachkräften besteht bei kleineren Schulungs- und Testdatensätzen die Gefahr von Verzerrungen. Um dieses Risiko zu minimieren, empfehlen wir die folgenden bewährten Methoden:

- Wenn Sie eine vortrainierte LLM-Lösung verwenden, stellen Sie sicher, dass Sie über eine ausreichende Menge an Testdaten verfügen. Die Testdaten sollten den tatsächlichen medizinischen Daten sehr ähnlich sein. Je nach Aufgabe kann dies zwischen 20 und mehr als 100 Datensätzen liegen.
- Sammeln Sie bei der Feinabstimmung eines LLM eine ausreichende Anzahl von markierten (Ground-Truth-Datensätzen) aus einer Vielzahl SMEs von medizinischen Zielgebieten. Ein allgemeiner Ausgangspunkt sind mindestens 100 qualitativ hochwertige Datensätze. Aufgrund

der Komplexität der Aufgabe und Ihrer Akzeptanzkriterien für die Genauigkeit sind jedoch möglicherweise mehr Datensätze erforderlich.

- Falls es für Ihren medizinischen Anwendungsfall erforderlich ist, sollten Sie Leitplanken implementieren und Datenverzerrungen und Fairness messen. Stellen Sie beispielsweise sicher, dass das LLM Fehldiagnosen aufgrund von Rassenprofilen von Patienten verhindert. Weitere Informationen finden Sie im Sicherheit und Leitplanken Abschnitt dieses Handbuchs.

Viele KI-Forschungs- und Entwicklungsunternehmen, wie Anthropic, haben in ihren Gründungsmodellen bereits Leitplanken implementiert, um Toxizität zu vermeiden. Sie können die Toxizitätserkennung verwenden, um die Eingabeaufforderungen und die ausgegebenen Antworten von zu überprüfen. LLMs Weitere Informationen finden Sie unter Toxizitätserkennung in der Amazon Comprehend Comprehend-Dokumentation und unter Guardrails in der Amazon Bedrock-Dokumentation.

Bei jeder generativen KI-Aufgabe besteht die Gefahr einer Halluzination. Sie können dieses Risiko mindern, indem Sie NLP-Aufgaben wie die Klassifizierung ausführen. Sie können auch fortgeschrittenere Techniken verwenden, z. B. Metriken zur Textähnlichkeit. BertScoreist eine häufig verwendete Metrik zur Textähnlichkeit. Weitere Informationen zu Techniken, mit denen Sie Halluzinationen abmildern können, finden Sie unter Umfassender Überblick über Techniken zur Bekämpfung von Halluzinationen in großen Sprachmodellen.

Metriken für medizinische NLP-Aufgaben

Sie können quantifizierbare Metriken erstellen, nachdem Sie Ground-Truth-Daten und von KMU bereitgestellte Labels für Schulungen und Tests erstellt haben. Die Überprüfung der Qualität durch qualitative Prozesse wie Stresstests und die Überprüfung der LLM-Ergebnisse ist hilfreich für eine schnelle Entwicklung. Metriken dienen jedoch als quantitative Benchmarks, die future LLM-Operationen unterstützen, und dienen als Leistungsmaßstäbe für jede Produktionsversion.

Es ist entscheidend, die medizinische Aufgabe zu verstehen. Metriken werden in der Regel einer der folgenden allgemeinen NLP-Aufgaben zugeordnet:

- Textklassifizierung — Das LLM kategorisiert den Text in eine oder mehrere vordefinierte Kategorien, basierend auf der Eingabeaufforderung und dem bereitgestellten Kontext. Ein Beispiel ist die Klassifizierung einer Schmerzkategorie anhand einer Schmerzskala. Beispiele für Metriken zur Textklassifizierung sind:
  - Genauigkeit

  - Präzision, auch bekannt als Makropräzision
  - Recall, auch bekannt als Macro Recall
  - F1-Score, auch bekannt als Makro-F1-Score
  - Hamming, Verlust
- Erkennung benannter Entitäten (NER) — Bei der Erkennung benannter Entitäten, auch Textextraktion genannt, werden benannte Entitäten, die in unstrukturiertem Text erwähnt werden, lokalisiert und in vordefinierte Kategorien eingeteilt. Ein Beispiel ist das Extrahieren der Namen von Medikamenten aus Patientenakten. Beispiele für NER-Metriken sind:
  - Genauigkeit
  - Präzision
  - Erinnern
  - F1-Ergebnis
  - Hamming-Verlust
- Generierung — Das LLM generiert neuen Text, indem es die Eingabeaufforderung und den bereitgestellten Kontext verarbeitet. Die Generierung umfasst Zusammenfassungsaufgaben oder Aufgaben zur Beantwortung von Fragen. Beispiele für Generierungsmetriken sind:
  - Rückruforientiertes Unterstudium zur Bewertung von Gisting (ROUGE)
  - Metrik zur Bewertung von Übersetzungen mit Explicit (METEOR) ORdering
  - Zweisprachige Evaluierung (BLEU) (für Übersetzungen)
  - Abstand zwischen Zeichenketten, auch bekannt als Kosinusähnlichkeit

Häufig gestellte Fragen zu Anwendungsfällen im Gesundheitswesen und in den Biowissenschaften

Im Folgenden finden Sie häufig gestellte Fragen zur Verwendung von Amazon Comprehend Medical oder zu medizinischen LLMs NLP-Aufgaben.

Wie wähle ich zwischen Amazon Comprehend Medical und einem LLM?

Wenn Ihre Aufgabe darin besteht, medizinische Entitäten in Ihrem medizinischen Text zu erkennen, lesen Sie die Dokumentation von Amazon Comprehend Medical, um zu erfahren, welche medizinischen Entitäten extrahiert werden können und ob eine der Ontologien Ihren Anwendungsfall berücksichtigt. Wenn nicht, ziehen Sie die Verwendung eines LLM in Betracht. Weitere Informationen finden Sie unter Anwendungsfälle für Amazon Comprehend Medical und Anwendungsfälle für ein LLM in diesem Handbuch.

Wie kann ich einem LLM Ergebnisse von Amazon Comprehend Medical zur Verfügung stellen?

Sie können die Ergebnisse von Amazon Comprehend Medical als Kontext in Ihre LLM-Eingabeaufforderungen integrieren. Dies erweitert das LLM um zusätzliches medizinisches Wissen und zusätzliche Terminologie. Der bereitgestellte Kontext kann die Leistung des LLM bei Aufgaben wie der Erkennung von Entitäten, der Zusammenfassung oder der Beantwortung von Fragen verbessern. Das Handbuch enthält mehrere Beispiele für die Strukturierung von Eingabeaufforderungen anhand der Ergebnisse von Amazon Comprehend Medical. Weitere Informationen finden Sie unter Kombination von Amazon Comprehend Medical mit großen Sprachmodellen in diesem Handbuch.

Was sind einige bewährte Methoden bei der Verwendung von Amazon Comprehend Medical? LLMs

Wir empfehlen, die Amazon Comprehend Medical Confidence Scores zu verwenden, um Entitäten in Ihren Eingabeaufforderungen zu filtern oder zu priorisieren. Es ist auch wichtig, die Leistung

anhand Ihrer spezifischen Daten zu bewerten und zu überprüfen, ob die Entitätsdefinitionen Ihren Anforderungen entsprechen. Die Kombination von Amazon Comprehend Medical mit domänenspezifischen Wissensquellen kann die Leistung des LLM weiter verbessern. Weitere Informationen finden Sie unter Bewährte Methoden für die Verwendung von Amazon Comprehend Medical in einem RAG-Workflow in diesem Handbuch.

Sollte ich ein vortrainiertes medizinisches LLM verwenden oder ein allgemeines LLM auf meinen Anwendungsfall im Gesundheitswesen abstimmen?

Die Entscheidung hängt von Ihren spezifischen Anforderungen und der Verfügbarkeit hochwertiger Trainingsdaten ab. Vortrainierte Ärzte LLMs können einen guten Ausgangspunkt bieten. Möglicherweise müssen Sie sie jedoch noch an Ihre domänenspezifischen Daten anpassen. Wenn Sie über ausreichend beschriftete Daten verfügen, kann die Feinabstimmung eines allgemeinen LLM eine praktikable Option sein. Weitere Informationen finden Sie unter Einen LLM auswählen und Wahl eines NLP-Ansatzes für das Gesundheitswesen und die Biowissenschaften in diesem Handbuch.

Wie beurteile ich die Leistung von NLP-Aufgaben LLMs für medizinische Zwecke?

Wir empfehlen die Verwendung quantitativer Kennzahlen wie Genauigkeit, Präzision, Erinnerungsvermögen und F1-Score für Aufgaben zur Textklassifizierung und Erkennung benannter Entitäten. Sie können ROUGE und METEOR für Aufgaben zur Textgenerierung verwenden. Es ist wichtig, zuverlässige Ground-Truth-Daten von Fachexperten kennzeichnen zu lassen und Prozesse zur Überwachung der Modellleistung im Laufe der Zeit zu implementieren. Weitere Informationen finden Sie unter Evaluierung LLMs für Anwendungen im Gesundheitswesen und in den Biowissenschaften in diesem Handbuch.

Was sind die Kompromisse zwischen LLM-Lösungen mit hoher Komplexität und niedriger Komplexität?

Die Feinabstimmung eines LLM oder der Aufbau eines kundenspezifischen LLM sind hochkomplexe Lösungen. Diese Ansätze können die Leistung verbessern, sind jedoch mit höheren Kosten und Wartungsanforderungen verbunden. Einfachere Lösungen, wie die Verwendung von Pretrained LLMs

oder Amazon Comprehend Medical, bieten möglicherweise eine akzeptable Leistung bei niedrigeren Kosten und schnelleren Entwicklungszyklen. In einigen Anwendungsfällen erfüllen diese Ansätze jedoch möglicherweise nicht die strengen Genauigkeitsanforderungen. Weitere Informationen finden Sie unter Überlegungen zur Geschäftsreife in diesem Handbuch.

Nächste Schritte und Ressourcen

Dieser Leitfaden hilft Ihnen dabei AWS-Services , medizinische NLP- und generative KI-Aufgaben für reale Anwendungen in Produktionsumgebungen zu automatisieren. Darin wird beschrieben, wie Sie Amazon Comprehend Medical, unterstützt durch Amazon Bedrock, vortrainierte medizinische Fachkräfte oder fein abgestimmt einsetzen können LLMs, um Ihre Geschäftsziele LLMs im Gesundheitswesen und in den Biowissenschaften LLMs zu erreichen. In diesem Leitfaden werden die Vor- und Nachteile der folgenden Ansätze beschrieben:

- Unabhängige Nutzung von Amazon Comprehend Medical
- Bereitstellung der Ergebnisse von Amazon Comprehend Medical für einen LLM
- Verwendung eines vortrainierten allgemeinen LLM oder eines medizinischen LLM im Rahmen eines Retrieval Augmented Generation (RAG) -Ansatzes
- Feinabstimmung eines allgemeinen LLM oder eines medizinischen LLM

Verwenden Sie den Entscheidungsbaum und die Überlegungen zur Geschäftsreife in diesem Leitfaden, um je nach Reifegrad Ihres Unternehmens AI/ML zwischen diesen Ansätzen zu wählen. Amazon Comprehend Medical und Amazon Bedrock LLMs bieten zwar leistungsstarke Funktionen, sind aber nur erfolgreich, wenn Sie sie ordnungsgemäß implementieren und auswerten. Verwenden Sie die in diesem Leitfaden beschriebenen Bewertungsinformationen und Kennzahlen, um die Leistung Ihrer Lösung zu überprüfen.

Für die nächsten Schritte empfehlen wir, dass IT-Manager, Architekten und technische Leiter des Gesundheitswesens mit AI/ML Ärzten zusammenarbeiten, um ihre medizinische NLP-Aufgabe zu ermitteln. Verwenden Sie diesen Leitfaden, um einen Entwicklungspfad auszuwählen, und verwenden Sie dann die entsprechenden AWS-Services Funktionen, um eine automatisierte Lösung erfolgreich zu implementieren. AWS

AWS Ressourcen

- Dokumentation von Amazon Comprehend Medical:
  - Entwicklerhandbuch
  - API Reference
- Dokumentation zu Amazon Bedrock

  - Bewertung des Amazon Bedrock-Modells
  - Feinabstimmung in Amazon Bedrock
- Feinabstimmung eines Modells in Amazon AI SageMaker
- Amazon SageMaker Ground Truth
- Amazon Comprehend Toxizitätserkennung
- AWS Kompetenzpartner im Gesundheitswesen

Sonstige Ressourcen

- Öffnen Sie die Medical-LLM-Bestenliste
- Eine Übersicht über umfangreiche Sprachmodelle für das Gesundheitswesen: von Daten, Technologie und Anwendungen bis hin zu Rechenschaftspflicht und Ethik
- Große Sprachmodelle sind schlechte Programmierer im medizinischen Bereich — Benchmarking bei der Abfrage von medizinischem Code
- Vom Anfänger zum Experten: Medizinisches Wissen allgemein modellieren LLMs

Mitwirkende

Verfassen

- Joe King, AWS leitender Datenwissenschaftler
- Ankith Ede, Lösungsarchitekt AWS
- Clement Perrot, leitender Stratege für generative AWS KI
- Jillian Forde, leitende Lösungsarchitektin AWS
- Rajesh Sitaraman, leitender Berater für Bereitstellung AWS
- Ross Claytor, Leitender angewandter Wissenschaftler AWS
- Shivesh Ummat, Lösungsarchitekt AWS

Überprüft

- Dilshad Raihan Akkam Veettil, leitender Datenwissenschaftler AWS
- Joseph Cottingham AWS , Architekt für Deep Learning

Technisches Schreiben

- Lilly AbouHarb, AWS leitende technische Redakteurin

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
Neue Abschnitte	Wir haben den Bereich Feinabstimmung umfangreicher Sprachmodelle im Gesundheitswesen und den Bereich Prompt Engineering hinzugefügt.	5. Dezember 2025
Erste Veröffentlichung	—	16. Dezember 2024

AWS Glossar zu präskriptiven Leitlinien

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

Zahlen

7 Rs

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

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

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

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

A

ABAC

Siehe attributbasierte Zugriffskontrolle.

abstrahierte Dienste

Siehe Managed Services.

ACID

Siehe Atomarität, Konsistenz, Isolierung und Haltbarkeit.

Aktiv-Aktiv-Migration

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

Aktiv-Passiv-Migration

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

Aggregatfunktion

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

AI

Siehe künstliche Intelligenz.

AIOps

Siehe Operationen im Bereich künstliche Intelligenz.

Anonymisierung

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

Anti-Muster

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

Anwendungssteuerung

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

Anwendungsportfolio

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

künstliche Intelligenz (KI)

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

Operationen mit künstlicher Intelligenz (AIOps)

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

Asymmetrische Verschlüsselung

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

Atomizität, Konsistenz, Isolierung, Haltbarkeit (ACID)

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

Attributbasierte Zugriffskontrolle (ABAC)

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

maßgebliche Datenquelle

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

Availability Zone

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

AWS Framework für die Einführung der Cloud (AWS CAF)

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

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

AWS Workload-Qualifizierungsrahmen (AWS WQF)

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

B

schlechter Bot

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

BCP

Siehe Planung der Geschäftskontinuität.

Verhaltensdiagramm

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

Big-Endian-System

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

Binäre Klassifikation

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

Bloom-Filter

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

Blau/Grün-Bereitstellung

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

Bot

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

Botnetz

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

branch

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

Zugang durch Glasbruch

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

Brownfield-Strategie

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

Puffer-Cache

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

Geschäftsfähigkeit

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

Planung der Geschäftskontinuität (BCP)

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

C

CAF

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

Bereitstellung auf Kanaren

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

CCoE

Weitere Informationen finden Sie im Cloud Center of Excellence.

CDC

Siehe Erfassung von Änderungsdaten.

Erfassung von Datenänderungen (CDC)

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

Chaos-Technik

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

CI/CD

Siehe Continuous Integration und Continuous Delivery.

Klassifizierung

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

clientseitige Verschlüsselung

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

Cloud-Exzellenzzentrum (CCoE)

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

Cloud Computing

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

Cloud-Betriebsmodell

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

Phasen der Einführung der Cloud

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

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

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

CMDB

Siehe Datenbank für das Konfigurationsmanagement.

Code-Repository

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

Kalter Cache

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

Kalte Daten

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

Computer Vision (CV)

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

Drift in der Konfiguration

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

Verwaltung der Datenbankkonfiguration (CMDB)

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

Konformitätspaket

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

Kontinuierliche Bereitstellung und kontinuierliche Integration (CI/CD)

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

CV

Siehe Computer Vision.

D

Daten im Ruhezustand

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

Datenklassifizierung

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

Datendrift

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

Daten während der Übertragung

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

Datennetz

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

Datenminimierung

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

Datenperimeter

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

Vorverarbeitung der Daten

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

Herkunft der Daten

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

betroffene Person

Eine Person, deren Daten gesammelt und verarbeitet werden.

Data Warehouse

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

Datenbankdefinitionssprache (DDL)

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

Datenbankmanipulationssprache (DML)

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

DDL

Siehe Datenbankdefinitionssprache.

Deep-Ensemble

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

Deep Learning

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

defense-in-depth

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

delegierter Administrator

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

Einsatz

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

Entwicklungsumgebung

Siehe Umgebung.

Detektivische Kontrolle

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

Abbildung des Wertstroms in der Entwicklung (DVSM)

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

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

digitaler Zwilling

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

Maßtabelle

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

Katastrophe

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

Notfallwiederherstellung (DR)

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

DML

Siehe Sprache zur Datenbankmanipulation.

Domainorientiertes Design

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

DR

Siehe Disaster Recovery.

Erkennung von Driften

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

DVSM

Siehe Abbildung des Wertstroms in der Entwicklung.

E

EDA

Siehe explorative Datenanalyse.

EDI

Siehe elektronischer Datenaustausch.

Edge-Computing

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

elektronischer Datenaustausch (EDI)

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

Verschlüsselung

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

Verschlüsselungsschlüssel

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

Endianismus

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

Endpunkt

Siehe Service-Endpunkt.

Endpunkt-Services

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

Unternehmensressourcenplanung (ERP)

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

Envelope-Verschlüsselung

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

Umgebung

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

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

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

Epics

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

ERP

Siehe Enterprise Resource Planning.

Explorative Datenanalyse (EDA)

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

F

Faktentabelle

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

schnell scheitern

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

Grenze zur Fehlerisolierung

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

Feature-Zweig

Siehe Zweig.

Features

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

Bedeutung der Feature

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

Featuretransformation

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

Eingabeaufforderung mit wenigen Klicks

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

FGAC

Siehe detaillierte Zugriffskontrolle.

Feinkörnige Zugriffskontrolle (FGAC)

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

Flash-Cut-Migration

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

FM

Siehe Fundamentmodell.

Fundamentmodell (FM)

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

G

Generative KI

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

Geoblocking

Siehe geografische Einschränkungen.

Geografische Einschränkungen (Geoblocking)

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

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

Gitflow-Workflow

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

goldenes Bild

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

Greenfield-Strategie

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

Integritätsschutz

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

H

HEKTAR

Siehe Hochverfügbarkeit.

Heterogene Datenbankmigration

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

hohe Verfügbarkeit (HA)

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

historische Modernisierung

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

Daten zurückhalten

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

Homogene Datenbankmigration

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

heiße Daten

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

Hotfix

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

Hypercare-Phase

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

I

IaC

Sehen Sie Infrastruktur als Code.

Identitätsbasierte Richtlinie

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

Leerlaufanwendung

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

IIoT

Siehe Industrielles Internet der Dinge.

unveränderliche Infrastruktur

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

Eingehende (ingress) VPC

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

Inkrementelle Migration

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

Industrie 4.0

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

Infrastruktur

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

Infrastructure as Code (IaC)

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

industrielles Internet der Dinge (T) IIo

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

Inspektions-VPC

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

Internet of Things (IoT)

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

Interpretierbarkeit

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

IoT

Siehe Internet der Dinge.

IT information library (ITIL, IT-Informationsbibliothek)

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

T service management (ITSM, IT-Servicemanagement)

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

BIS

Siehe IT-Informationsbibliothek.

ITSM

Siehe IT-Servicemanagement.

L

Labelbasierte Zugangskontrolle (LBAC)

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

Landing Zone

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

großes Sprachmodell (LLM)

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

Große Migration

: Eine Migration von 300 oder mehr Servern.

SCHWARZ

: Siehe Labelbasierte Zugriffskontrolle.

Geringste Berechtigung

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

Lift and Shift

: Siehe 7 Rs.

Little-Endian-System

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

LLM

Siehe großes Sprachmodell.

Niedrigere Umgebungen

Siehe Umgebung.

M

Machine Learning (ML)

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

Hauptzweig

Siehe Filiale.

Malware

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

verwaltete Dienste

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

Manufacturing Execution System (MES)

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

MAP

Siehe Migration Acceleration Program.

Mechanismus

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

Mitgliedskonto

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

MES

Siehe Manufacturing Execution System.

Message Queuing-Telemetrietransport (MQTT)

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

Microservice

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

Microservices-Architekturen

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

Migration Acceleration Program (MAP)

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

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

Migration in großem Maßstab

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

Migrationsfabrik

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

Migrationsmetadaten

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

Migrationsmuster

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

Migration Portfolio Assessment (MPA)

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

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

Migration Readiness Assessment (MRA)

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

Migrationsstrategie

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

ML

Siehe maschinelles Lernen.

Modernisierung

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

Bewertung der Modernisierungsfähigkeit

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

Monolithische Anwendungen (Monolithen)

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

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

MPA

Siehe Bewertung des Migrationsportfolios.

MQTT

Siehe Message Queuing-Telemetrietransport.

Mehrklassen-Klassifizierung

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

veränderbare Infrastruktur

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

O

OAC

Siehe Origin Access Control.

EICHE

Siehe Zugriffsidentität von Origin.

COM

Siehe organisatorisches Change-Management.

Offline-Migration

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

OI

Siehe Betriebsintegration.

OLA

Siehe Vereinbarung auf operativer Ebene.

Online-Migration

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

OPC-UA

Siehe Open Process Communications — Unified Architecture.

Offene Prozesskommunikation — Einheitliche Architektur (OPC-UA)

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

Vereinbarung auf Betriebsebene (OLA)

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

Überprüfung der Betriebsbereitschaft (ORR)

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

Betriebstechnologie (OT)

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

Betriebsintegration (OI)

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

Organisationspfad

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

Organisatorisches Veränderungsmanagement (OCM)

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

Ursprungszugriffskontrolle (OAC)

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

Ursprungszugriffsidentität (OAI)

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

ORR

Weitere Informationen finden Sie unter Überprüfung der Betriebsbereitschaft.

NICHT

Siehe Betriebstechnologie.

Ausgehende (egress) VPC

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

P

Berechtigungsgrenze

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

persönlich identifizierbare Informationen (PII)

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

Personenbezogene Daten

Siehe persönlich identifizierbare Informationen.

Playbook

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

PLC

Siehe programmierbare Logiksteuerung.

PLM

Siehe Produktlebenszyklusmanagement.

policy

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

Polyglotte Beharrlichkeit

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

Portfoliobewertung

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

predicate

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

Prädikat Pushdown

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

Präventive Kontrolle

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

Prinzipal

Eine Entität AWS , die Aktionen ausführen und auf Ressourcen zugreifen kann. Diese Entität ist in der Regel ein Root-Benutzer für eine AWS-Konto, eine IAM-Rolle oder einen Benutzer. Weitere Informationen finden Sie unter Prinzipal in Rollenbegriffe und -konzepte in der IAM-Dokumentation.

Datenschutz von Natur aus

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

Privat gehostete Zonen

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

proaktive Steuerung

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

Produktlebenszyklusmanagement (PLM)

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

Produktionsumgebung

Siehe Umgebung.

Speicherprogrammierbare Steuerung (SPS)

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

schnelle Verkettung

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

Pseudonymisierung

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

publish/subscribe (pub/sub)

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

Q

Abfrageplan

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

Abfrageplanregression

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

R

RACI-Matrix

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

RAG

Siehe Erweiterte Generierung beim Abrufen.

Ransomware

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

RASCI-Matrix

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

RCAC

Siehe Zugriffskontrolle für Zeilen und Spalten.

Read Replica

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

neu strukturieren

Siehe 7 Rs.

Recovery Point Objective (RPO)

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

Wiederherstellungszeitziel (RTO)

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

Refaktorierung

Siehe 7 Rs.

Region

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

Regression

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

rehosten

Siehe 7 Rs.

Veröffentlichung

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

umziehen

Siehe 7 Rs.

neue Plattform

Siehe 7 Rs.

Rückkauf

Siehe 7 Rs.

Ausfallsicherheit

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

Ressourcenbasierte Richtlinie

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

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

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

Reaktive Kontrolle

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

Beibehaltung

Siehe 7 Rs.

zurückziehen

Siehe 7 Rs.

Retrieval Augmented Generation (RAG)

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

Drehung

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

Zugriffskontrolle für Zeilen und Spalten (RCAC)

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

RPO

Siehe Recovery Point Objective.

RTO

Siehe Ziel für die Erholungszeit.

Runbook

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

S

SAML 2.0

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

SCADA

Siehe Aufsichtskontrolle und Datenerfassung.

SCP

Siehe Richtlinie zur Dienstkontrolle.

Secret

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

Sicherheit durch Design

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

Sicherheitskontrolle

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

Härtung der Sicherheit

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

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

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

Automatisierung von Sicherheitsreaktionen

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

Serverseitige Verschlüsselung

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

Service-Kontrollrichtlinie (SCP)

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

Service-Endpunkt

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

Service Level Agreement (SLA)

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

Service-Level-Indikator (SLI)

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

Service-Level-Ziel (SLO)

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

Modell der geteilten Verantwortung

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

SIEM

Siehe Sicherheitsinformations- und Event-Management-System.

Single Point of Failure (SPOF)

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

SLA

Siehe Service Level Agreement.

SLI

Siehe Service-Level-Indikator.

ALSO

Siehe Service-Level-Ziel.

split-and-seed Modell

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

SPOTTEN

Siehe Single Point of Failure.

Sternschema

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

Strangler-Fig-Muster

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

Subnetz

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

Aufsichtskontrolle und Datenerfassung (SCADA)

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

Symmetrische Verschlüsselung

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

synthetisches Testen

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

Systemaufforderung

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

T

tags

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

Zielvariable

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

Aufgabenliste

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

Testumgebungen

Siehe Umgebung.

Training

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

Transit-Gateway

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

Stammbasierter Workflow

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

Vertrauenswürdiger Zugriff

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

Optimieren

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

Zwei-Pizzen-Team

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

U

Unsicherheit

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

undifferenzierte Aufgaben

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

höhere Umgebungen

Siehe Umgebung.

V

Vacuuming

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

Versionskontrolle

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

VPC-Peering

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

Schwachstelle

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

W

Warmer Cache

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

warme Daten

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

Fensterfunktion

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

Workload

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

Workstream

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

WURM

Sehen Sie einmal schreiben, viele lesen.

WQF

Siehe AWS Workload-Qualifizierungsrahmen.

einmal schreiben, viele lesen (WORM)

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

Z

Zero-Day-Exploit

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

Zero-Day-Sicherheitslücke

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

Eingabeaufforderung ohne Zwischenfälle

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

Zombie-Anwendung

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

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