ヘルスケア AWS 向けの での取得拡張生成ソリューションの作成

AWS 規範ガイダンス

AWS 規範ガイダンス: ヘルスケア AWS 向けの での取得拡張生成ソリューションの作成

Table of Contents

ヘルスケア向けの で AWS の取得拡張生成ソリューションの作成

Amazon Web Services、Accenture、および Cadiem (寄稿者)

2025 年 3 月 (ドキュメント履歴)

大規模言語モデル (LLMs) と生成 AI 以前は、ヘルスケア業界で自動化および高精度アプリケーションを開発する作業は困難でした。従来の方法は、手動によるデータ入力と分析に大きく依存していました。医療画像と患者記録の分析は複雑であるため、人間による広範な介入が必要であり、多くの場合、ワークフローが断片化され非効率になります。AI テクノロジーの進歩は、ハイパーパーソナライズされたアプリケーションを大規模に構築するのに役立ちます。ヘルスケアアプリケーションは、医療ナレッジベースと統合し、診断イメージをより正確に解釈し、予測モデルを使用して患者の成果を予測できるようになりました。

このガイドでは、LLMs が、 で構築できる取得拡張生成アプリケーションを通じて医療に変革をもたらしている方法について説明します AWS のサービス。Retrieval Augmented Generation (RAG) は、LLM がレスポンスを生成する前にトレーニングデータソースの外部にある信頼できるデータソースを参照する生成 AI テクノロジーです。RAG アプリケーションは、モデルの出力を実際の知識に基づいて構築し、ハルシネーションを減らし、レスポンスの関連性を高めます。医療部門では、RAG を使用して正確でup-to-date医療情報を提供し、医療プロバイダーが最新の研究および臨床ガイドラインにアクセスできるようにします。これらのテクノロジーは、データを実用的なインサイトに変換し、複雑なプロセスを自動化することで、患者ケアの強化、運用の合理化、医療専門家の生産性の向上に役立ちます。

Amazon Bedrock では、LLMs微調整し、インテリジェントエージェントと統合して、高度な医療ソリューションを作成できます。このガイドでは、Amazon OpenSearch Service と Amazon Neptune のシナジーを強調し、これらのサービスが検索の関連性を高め、高度なマルチソースデータの取得を通じて RAG ソリューションをどのように向上させるかを示します。Amazon Bedrock エージェントを使用する包括的な Amazon Bedrock ソリューションをオーケストレーションし、さまざまなデータリポジトリ間のインタラクションLangChainをシームレスに調整できます。この統合は、特殊なサービスを組み合わせて、より効果的で効率的な AI 駆動型システムを作成する能力を示しています。

患者のケアと生産性

このガイドでは、患者ケアと生産性に関する 2 つの実際のユースケースを紹介します。患者データの拡張と再入院リスクの予測です。これらのソリューションを大規模に実装するための戦略的設計図を提供し、医療組織に AI 主導のプロセスの産業化への明確な道を提供します。これらのインサイトを通じて、医療施設は高度な AI テクノロジーを使用して、より効率的でインテリジェントなワークフローを作成できます。

人材管理

このガイドでは、医療従事者が生成 AI を日常業務にシームレスに統合するための再スキルと権限を与える戦略についても概説します。これにより、生産性と患者ケアの品質の両方を向上させることができます。高度な AI ツールを効果的に使用するスキルをワークフォースに提供することで、医療組織は投資収益率を最大化し、患者ケアのイノベーションを推進できます。

この AI を活用した人材管理ソリューションには、以下の主要な機能が含まれています。

- インテリジェントな人材の再開パーサー – Amazon Bedrock で利用可能な高度な LLMs を使用することで、このツールは重要な人材スキルと属性を再開から効率的に抽出して分析します。このツールを使用すると、採用プロセスを合理化できます。
- 人材ナレッジベース – Amazon Neptune を搭載したこの動的データベースは、人員配置レベル、スキル分布、業界の傾向に関するリアルタイムのインサイトを提供します。これにより、ワークフォース管理に関するデータ主導型の意思決定を行うことができます。
- 学習レコメンデーションエンジン – この AI 駆動型ツールは、組織内のスキルギャップを特定し、医療スタッフ向けにパーソナライズされたトレーニングプログラムを推奨します。このツールは、継続的なプロフェッショナル開発を促進し、進化する医療テクノロジーにワークフォースが適応するのに役立ちます。

これらの AI 主導の機能を組み合わせることで、ワークフォースのパフォーマンスを最適化し、インテリジェンスと効率を向上させて人材管理に変革をもたらします。

機会と課題

Amazon Bedrock は、生産性、スケーラビリティ、コスト効率、データ駆動型のインサイトを強化できます。Amazon Bedrock を使用すると、医療組織はコンテンツの作成やデータ分析から自動化された意思決定まで、さまざまなユースケースで LLMs を効果的に使用できます。このガイドでは、概念実証から本番稼働への移行中に、データ品質の問題、インフラストラクチャのスケーラビリティ、モデルパフォーマンスのメンテナンス、継続的な改善要件など、一般的な生成 AI の課題を克服するためのアプローチを提供します。

ヘルスケアにおける生成 AI アプリケーションの機会

ヘルスケア業界は、生成 AI アプリケーションがもたらす機会に支えられ、変革的な変化に備えています。生成 AI は、患者ケアを強化し、運用を合理化し、医療研究を加速させる可能性があります。高度な AI モデルを使用することで、医療プロバイダーは医療記録の拡張を自動化できます。包括的でup-to-date患者履歴により、より正確な診断と治療計画が容易になります。ソノグラムやその他の医療画像などの AI 駆動型画像分析は、迅速かつ正確なインサイトを提供し、医療専門家のワークロードを減らし、人為的ミスのリスクを最小限に抑えることができます。

診断や治療以外にも、生成 AI は予測分析において重要な役割を果たします。予測分析は、医療組織が患者の成果を予測し、それに応じてケアプランをパーソナライズするのに役立ちます。このテクノロジーは、患者データの管理からプロバイダーと患者間の通信の合理化まで、管理プロセスを最適化することもできます。生成 AI ソリューションを既存の医療システムと統合することで、医療施設はより高い効率を実現し、コストを削減し、最終的にはより質の高いケアを提供できます。AI と医療の統合は単なる強化ではなく、よりインテリジェントで応答性があり、患者中心のケアへの根本的な移行です。

高度な画像分析

Amazon Bedrock を Amazon Neptune や Amazon OpenSearch Service などのデータストアと組み合わせることで、医療における高度な画像分析の複雑さに対処できます。情報取得ソリューションは、診断画像を評価し、ソノグラムを解釈することで、疾患検出プロセスを強化し、解釈精度を高めることができます。このソリューションは、視覚評価データとテキスト評価データを、医師による手動による患者評価レビューと統合できます。

ソリューションの産業化に伴う課題

ヘルスケアで AI ソリューションを産業化する際に取り組むべき主な障害は、データ品質と可用性です。ヘルスケアデータは、多くの場合、断片化され、一貫性のない形式で存在します。AI モデルがクリーンで構造化された代表的なデータにアクセスできることを確認することは、実際のシナリオでパフォーマンスを維持する上で不可欠です。本番環境のため、インフラストラクチャのスケーラビリティが課題になる可能性があります。これらの環境では、医療保険の相互運用性と説明責任に関する法律 (HIPAA) などのデータプライバシー規制への準拠を維持しながら、大量のリアルタイム患者データを処理する必要があります。さらに、時間の経過とともに進化する新しい医療情報と患者データでは、AI モデルを再トレーニングして更新し、関連性を維持し、正確なレコメンデーションを提供する必要があります。最後に、これらの AI ソリューションを既存の医療システムに統合することは、相互運用性の問題と現在の臨床ワークフローとの整合性の必要性により、複雑になる可能性があります。この統合には、技術的変更と運用上の変更の両方が必要です。

ユースケース: 拡張された患者データを使用した医療インテリジェンスアプリケーションの構築

生成 AI は、臨床機能と管理機能の両方を強化することで、患者のケアとスタッフの生産性を高めるのに役立ちます。ソノグラムの解釈などの AI 駆動型画像分析は、診断プロセスを高速化し、精度を向上させます。タイムリーな医療介入をサポートする重要なインサイトを提供できます。

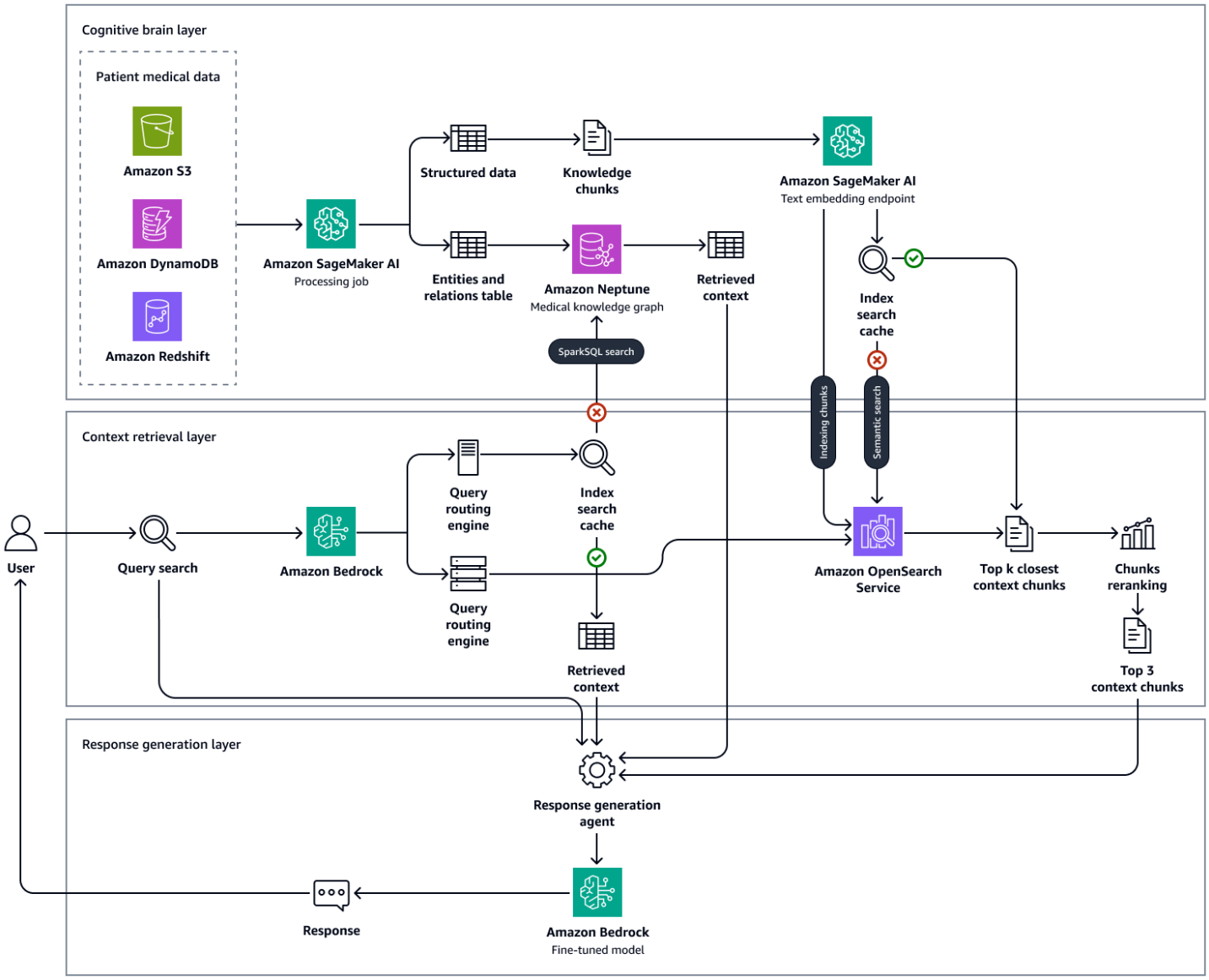
生成 AI モデルをナレッジグラフと組み合わせると、電子患者レコードの時系列整理を自動化できます。これにより、臨床医と患者のやり取り、症状、診断、検査結果、画像分析からのリアルタイムデータを統合できます。これにより、医師には包括的な患者データが提供されます。このデータは、医師がより正確でタイムリーな医療上の意思決定を行い、患者の成果と医療提供者の生産性の両方を向上させるのに役立ちます。

ソリューションの概要

AI は、患者データと医療知識を合成して貴重なインサイトを提供することで、医師と臨床医を強化できます。この検索拡張生成 (RAG) ソリューションは、何百万もの臨床インタラクションから包括的な患者データと知識のセットを消費する医療インテリジェンスエンジンです。生成 AI の能力を活用して、患者ケアを改善するための証拠ベースのインサイトを作成します。これは、臨床ワークフローを強化し、エラーを減らし、患者の成果を向上させるように設計されています。

このソリューションには、LLMs。この機能により、医療担当者が同様の診断画像を手動で検索し、診断結果を分析しなければならない時間が短縮されます。

次の図は、このソリューションのend-to-end-workflowを示しています。Amazon Neptune、Amazon SageMaker AI、Amazon OpenSearch Service、Amazon Bedrock の基盤モデルを使用します。Neptune の医療知識グラフを操作するコンテキスト検索エージェントでは、Amazon Bedrock エージェントとLangChainエージェントのいずれかを選択できます。



サンプル医療質問を用いた実験では、Neptune、臨床ナレッジベースを格納する OpenSearch ベクトルデータベース、Amazon Bedrock LLMs で維持されているナレッジグラフを使用してアプローチによって生成された最終的なレスポンスが事実に基づいており、誤検出を減らし、真陽性を増やすことではるかに正確であることがわかりました。このソリューションは、患者の健康状態に関する証拠ベースのインサイトを生成し、臨床ワークフローを強化し、エラーを減らし、患者の成果を向上させることを目的としています。

このソリューションの構築は、次のステップで構成されます。

- ステップ 1: データを検出する
- ステップ 2: 医療知識グラフを構築する
- ステップ 3: 医療知識グラフをクエリするためのコンテキスト検索エージェントを構築する

ステップ 1: データを検出する

ヘルスケア AI 駆動型ソリューションの開発をサポートするために使用できるオープンソースの医療データセットは多数あります。このようなデータセットの 1 つは、MIMIC-IV データセットです。MIMIC-IV データセットは、医療研究コミュニティで広く使用されている公開されている電子医療記録 (EHR) データセットです。MIMIC-IV には、患者記録からのフリーテキストの退床メモなど、詳細な臨床情報が含まれています。これらのレコードを使用して、テキスト集計とエンティティ抽出の手法を試すことができます。これらの手法は、非構造化テキストから医療情報 (患者の症状、薬剤、処方された治療など) を抽出するのに役立ちます。

また、研究目的で特別にキュレートされた、注釈付きで識別されていない患者の退床サマリーを提供するデータセットを使用することもできます。放出サマリーデータセットは、エンティティ抽出を試すのに役立ちます。これにより、テキストから主要な医療エンティティ (状態、処置、薬剤など) を識別できます。ステップ 2: 医療知識グラフを構築するこのガイドでは、MIMIC-IV および放出サマリーデータセットから抽出された構造化データを使用して医療知識グラフを作成する方法について説明します。この医療知識グラフは、医療専門家向けの高度なクエリおよび意思決定支援システムのバックボーンとして機能します。

テキストベースのデータセットに加えて、画像データセットを使用できます。例えば、筋骨格X線画像 (MURA) データセットは、骨のマルチビューX線画像の包括的なデータベースです。このような画像データセットを使用して、医療画像デコード技術による診断評価を試します。これらのデコード手法は、筋骨格疾患、循環器疾患、骨粗鬆病などの疾患の早期診断に不可欠です。医療画像データセットのビジョンモデルと言語基盤モデルを微調整することで、診断画像の異常を検出できます。これにより、システムは臨床医に早期かつ正確な診断インサイトを提供できます。画像データセットとテキストデータセットを使用することで、テキストデータと画像データの両方を処理できる AI 駆動型の医療アプリケーションを作成して、患者のケアを向上させることができます。

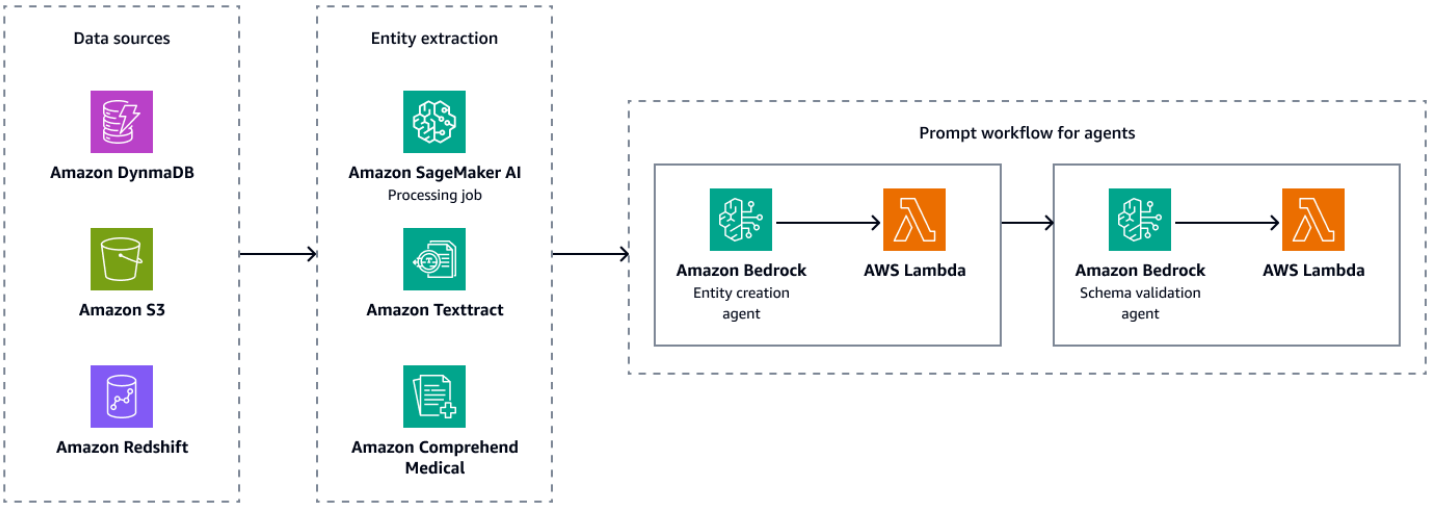
ステップ 2: 医療知識グラフを構築する

大規模なナレッジベースに基づいて意思決定支援システムを構築したい医療組織にとって、重要な課題は、臨床ノート、医学ジャーナル、解雇概要、その他のデータソースに存在する医療エンティティを見つけて抽出することです。また、抽出されたエンティティ、属性、関係を効果的に使用するには、これらの医療記録から時間的関係、件名、確実性の評価をキャプチャする必要があります。

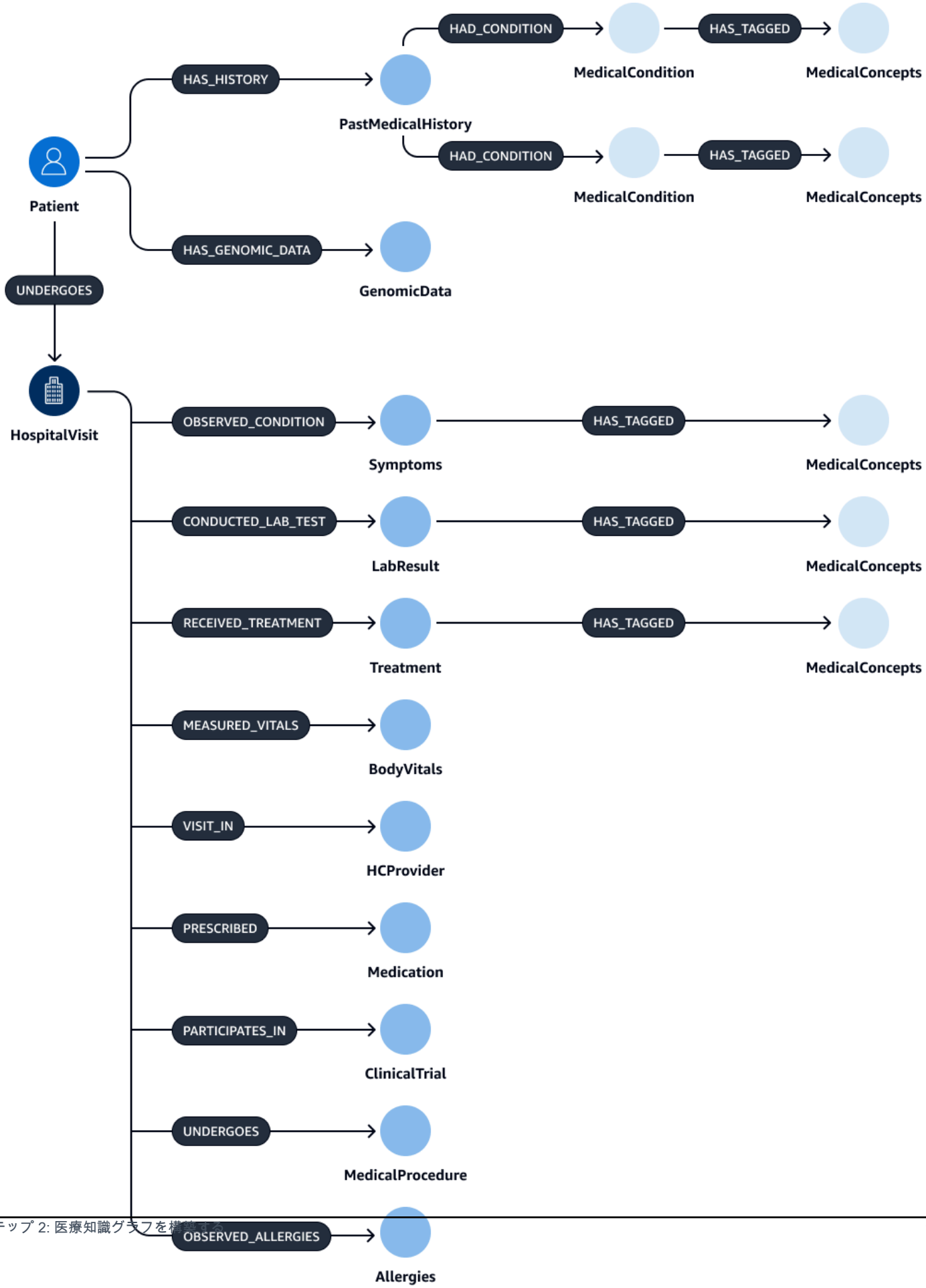
最初のステップは、Amazon Bedrock の Llama 3 などの基盤モデルに数ショットのプロンプトを使用して、非構造化医療テキストから医療概念を抽出することです。少数のショットプロンプトは、LLM に同様のタスクの実行を求める前に、タスクと必要な出力を示す少数の例を提供する場合です。LLM ベースの医療エンティティエクストラクタを使用すると、構造化されていない医療テキストを解析し、医療知識エンティティの構造化データ表現を生成できます。ダウンストリーム分析と自動化のために患者属性を保存することもできます。エンティティ抽出プロセスには、次のアクションが含まれます。

- 疾患、薬剤、医療デバイス、投与量、薬剤頻度、薬剤期間、症状、医療処置、およびそれらの臨床的な関連属性などの医療概念に関する情報を抽出します。
- 抽出されたエンティティ、サブジェクト、確実性評価間の時間的関係などの機能機能をキャプチャします。
- 次のような標準医療語彙を拡張します。
  - RxNorm データベースからの概念識別子 (RxCUI) RxNorm
  - 国際疾病分類、第 10 回リビジョン、臨床変更 (ICD-10-CM) からのコード
  - Medical Subject Headings (MeSH) の用語
  - Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) の概念
  - Unified Medical Language System (UMLS) からのコード
- 退床メモを要約し、トランスクリプトから医療インサイトを取得します。

次の図は、エンティティ、属性、および関係の有効なペアの組み合わせを作成するためのエンティティ抽出とスキーマ検証の手順を示しています。退床サマリーや患者メモなどの非構造化データを Amazon Simple Storage Service (Amazon S3) に保存できます。エンタープライズリソースプランニング (ERP) データ、電子患者レコード、ラボ情報システムなどの構造化データを Amazon Redshift と Amazon DynamoDB に保存できます。Amazon Bedrock エンティティ作成エージェントを構築できます。このエージェントは、Amazon SageMaker AI データ抽出パイプライン、Amazon Textract、Amazon Comprehend Medical などのサービスを統合して、構造化データソースと非構造化データソースからエンティティ、関係、属性を抽出できます。最後に、Amazon Bedrock スキーマ検証エージェントを使用して、抽出されたエンティティと関係が事前定義されたグラフスキーマに準拠していること、ノードエッジ接続と関連プロパティの整合性が維持されていることを確認します。



エンティティ、リレーション、属性を抽出して検証した後、それらをリンクして subject-object-predicate トリプレットを作成できます。次の図に示すように、このデータを Amazon Neptune グラフデータベースに取り込みます。グラフデータベースは、データ項目間の関係を保存してクエリするように最適化されています。


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

このデータを使用して包括的なナレッジグラフを作成できます。ナレッジグラフは、あらゆる種類の接続情報を整理およびクエリするのに役立ちます。たとえば、、、HospitalVisit、、PastMedicalHistorySymptomsMedicationMedicalProcedures、および のメジャーノードを持つナレッジグラフを作成できますTreatment。

次の表に、排出ノートから抽出できるエンティティとその属性を示します。

エンティティ	属性
Patient	PatientID , Name, Age, Gender, Address, ContactInformation
HospitalVisit	VisitDate , Reason, Notes
HealthcareProvider	ProviderID , Name, Specialty , ContactInformation , Address, AffiliatedInstitution
Symptoms	Description , RiskFactors
Allergies	AllergyType , Duration
Medication	MedicationID , Name, Description , Dosage, SideEffects , Manufacturer
PastMedicalHistory	ContinuingMedicines
MedicalCondition	ConditionName , Severity, Treatment Received , DoctorinCharge , HospitalName , MedicinesFollowed
BodyVitals	HeartRate , BloodPressure , RespiratoryRate , BodyTemperature , BMI
LabResult	LabResultID , PatientID , TestName, Result, Date
ClinicalTrial	TrialID, Name, Description , Phase, Status, StartDate , EndDate

エンティティ	属性
GenomicData	GenomicDataID , PatientID , SequenceData , VariantInformation
Treatment	TreatmentID , Name, Description , Type, SideEffects
MedicalProcedure	ProcedureID , Name, Description , Risks, Outcomes
MedicalConcepts	UMLSCodes , MedicalVocabularies

次の表に、エンティティが持つ可能性のある関係とその対応する属性を示します。たとえば、Patientエンティティは [UNDERGOES]関係を持つHospitalVisitエンティティに接続できます。この関係の属性は ですVisitDate。

サブジェクトエンティティ	関係	オブジェクトエンティティ	属性
Patient	[UNDERGOES]	HospitalVisit	VisitDate
HospitalVisit	[VISIT_IN]	HealthcareProvider	ProviderName , Location, ProviderID , VisitDate
HospitalVisit	[OBSERVED_CONDITION]	Symptoms	Severity, CurrentStatus , VisitDate
HospitalVisit	[RECEIVED_TREATMENT]	Treatment	Duration, Dosage, VisitDate
HospitalVisit	[PRESCRIBED]	Medication	Duration, Dosage, Adherence , VisitDate

サブジェクトエンティティ	関係	オブジェクトエンティティ	属性
Patient	[HAS_HISTORY]	PastMedicalHistory	なし
PastMedicalHistory	[HAD_CONDITION]	MedicalCondition	DiagnosisDate , CurrentStatus
HospitalVisit	[PARTICIPATES_IN]	ClinicalTrial	VisitDate , Status, Outcomes
Patient	[HAS_GENOMIC_DATA]	GenomicData	CollectionDate
HospitalVisit	[OBSERVED_ALLERGIES]	Allergies	VisitDate
HospitalVisit	[CONDUCTED_LAB_TEST]	LabResult	VisitDate , AnalysisDate , Interpretation
HospitalVisit	[UNDERGOES]	MedicalProcedure	VisitDate , Outcome
MedicalCondition	[HAS_TAGGED]	MedicalConcepts	なし
LabResult	[HAS_TAGGED]	MedicalConcepts	なし
Treatment	[HAS_TAGGED]	MedicalConcepts	なし
Symptoms	[HAS_TAGGED]	MedicalConcepts	なし

ステップ 3: 医療知識グラフをクエリするためのコンテキスト検索エージェントを構築する

医療グラフデータベースを構築した後、次のステップはグラフ操作用のエージェントを構築することです。エージェントは、臨床医または臨床医が入力するクエリの適切で必要なコンテキストを取得します。ナレッジグラフからコンテキストを取得するこれらのエージェントを設定するには、いくつかのオプションがあります。

- Amazon Bedrock エージェント
- LangChain エージェント

グラフインタラクション用の Amazon Bedrock エージェント

Amazon Bedrock エージェントは、Amazon Neptune グラフデータベースとシームレスに連携します。Amazon Bedrock アクショングループを使用して高度なインタラクションを実行できます。アクショングループは、Neptune openCypher クエリを実行する AWS Lambda 関数を呼び出してプロセスを開始します。

ナレッジグラフのクエリには、直接クエリの実行とコンテキスト埋め込みを使用したクエリの 2 つの異なるアプローチを使用できます。これらのアプローチは、特定のユースケースとランク付け基準に応じて、個別に適用することも組み合わせることもできます。両方のアプローチを組み合わせることで、より包括的なコンテキストを LLM に提供できるため、結果を向上させることができます。以下は、2 つのクエリ実行アプローチです。

- 埋め込みなしで直接 Cypher クエリを実行する – Lambda 関数は、埋め込みベースの検索なしで Neptune に対して直接クエリを実行します。このアプローチの例を次に示します。

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

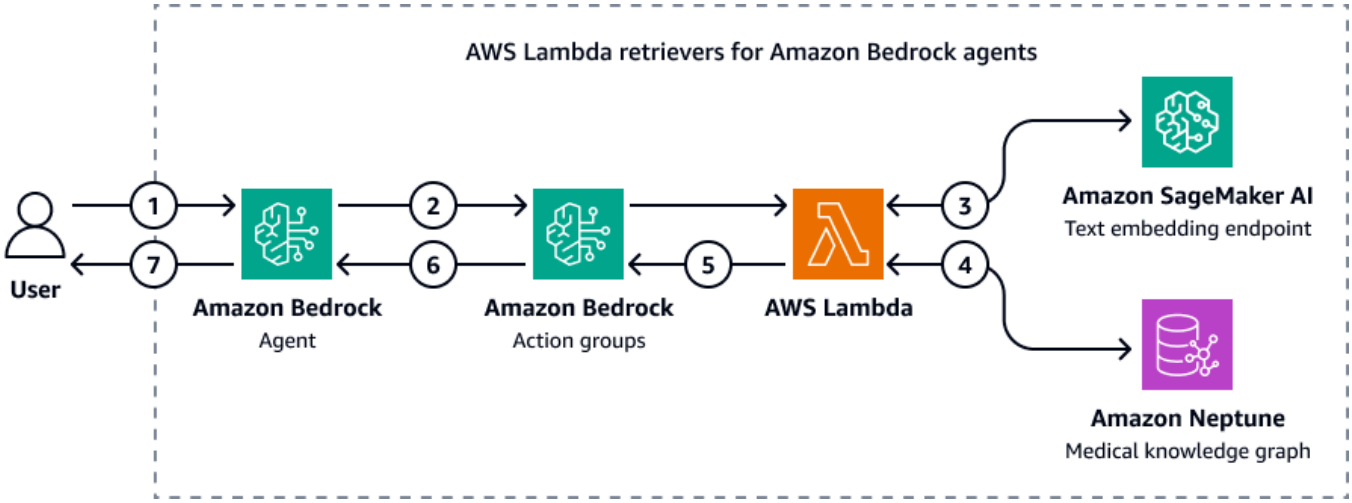
- 埋め込み検索を使用した直接 Cypher クエリの実行 – Lambda 関数は埋め込み検索を使用してクエリ結果を強化します。このアプローチは、データの高密度ベクトル表現である埋め込みを組み込むことで、クエリの実行を強化します。埋め込みは、クエリにセマンティック類似性や完全一致を超える広範な理解が必要な場合に特に役立ちます。事前トレーニング済みモデルまたはカスタムトレーニング済みモデルを使用して、病状ごとに埋め込みを生成できます。このアプローチの例を次に示します。

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

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

この例では、search_embedding("Acute Diabetes")関数は意味的に「Acute" に近い条件を取得します。これにより、クエリは、前骨茎や異化症候群などの状態を持つ患者も検索できます。

次の図は、医療知識グラフの Cypher クエリを実行するために Amazon Bedrock エージェントが Amazon Neptune とやり取りする方法を示しています。



この図表は、次のワークフローを示しています：

1. ユーザーは Amazon Bedrock エージェントに質問を送信します。
2. Amazon Bedrock エージェントは、質問および入力フィルター変数を Amazon Bedrock アクショングループに渡します。これらのアクショングループには、Amazon SageMaker AI テキスト埋め込みエンドポイントと Amazon Neptune 医療知識グラフを操作する AWS Lambda 関数が含まれています。
3. Lambda 関数は SageMaker AI テキスト埋め込みエンドポイントと統合され、openCypher クエリ内でセマンティック検索を実行します。基になるLangChainエージェントを使用して、自然言語クエリを openCypher クエリに変換します。
4. Lambda 関数は、Neptune 医療ナレッジグラフに正しいデータセットをクエリし、Neptune 医療ナレッジグラフから出力を受け取ります。

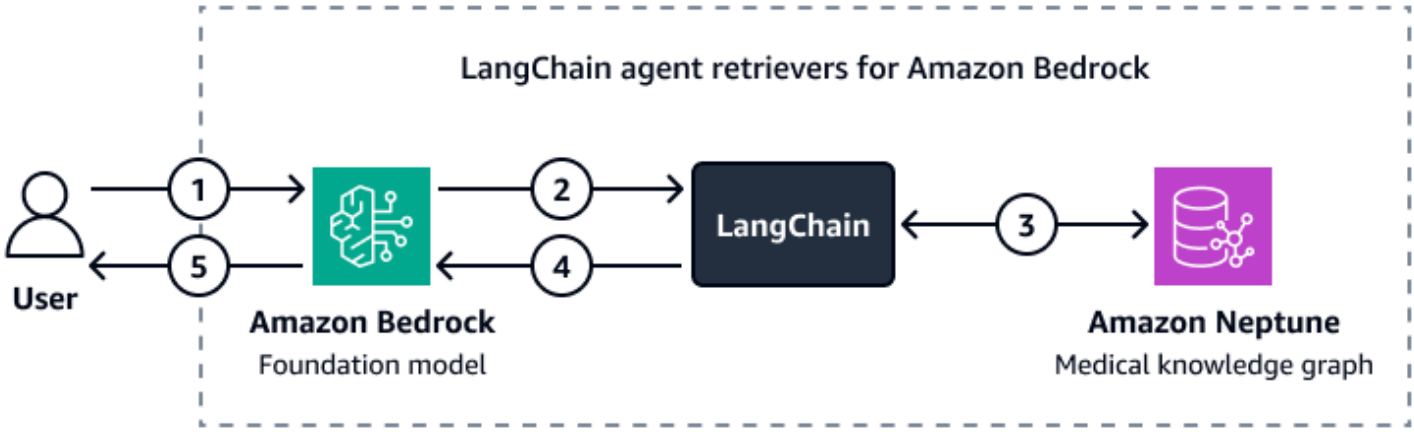
5. Lambda 関数は、Neptune から Amazon Bedrock アクショングループに結果を返します。
6. Amazon Bedrock アクショングループは、取得したコンテキストを Amazon Bedrock エージェントに送信します。
7. Amazon Bedrock エージェントは、元のユーザークエリとナレッジグラフから取得したコンテキストを使用してレスポンスを生成します。

LangChain グラフインタラクションのエージェント

Neptune LangChainと統合して、グラフベースのクエリと取得を有効にすることができます。このアプローチでは、Neptune のグラフデータベース機能を使用することで、AI 主導のワークフローを強化できます。カスタムリLangChainトリーバーは仲介として機能します。Amazon Bedrock の基盤モデルは、直接 Cypher クエリとより複雑なグラフアルゴリズムの両方を使用して Neptune とやり取りできます。

カスタムリトリーバーを使用して、LangChainエージェントが Neptune グラフアルゴリズムとやり取りする方法を絞り込むことができます。たとえば、数ショットプロンプトを使用できます。これにより、特定のパターンや例に基づいて基盤モデルのレスポンスを調整できます。LLM で識別されたフィルターを適用してコンテキストを絞り込み、レスポンスの精度を向上させることもできます。これにより、複雑なグラフデータを操作する際の全体的な取得プロセスの効率と精度を向上させることができます。

次の図は、カスタムLangChainエージェントが Amazon Bedrock 基盤モデルと Amazon Neptune 医療知識グラフ間のインタラクションをオーケストレーションする方法を示しています。



この図表は、次のワークフローを示しています：

1. ユーザーは Amazon Bedrock とLangChainエージェントに質問を送信します。

2. Amazon Bedrock 基盤モデルは、LangChainエージェントによって提供される Neptune スキーマを使用して、ユーザーの質問に対するクエリを生成します。
3. LangChain エージェントは、Amazon Neptune 医療知識グラフに対してクエリを実行します。
4. LangChain エージェントは、取得したコンテキストを Amazon Bedrock 基盤モデルに送信します。
5. Amazon Bedrock 基盤モデルは、取得したコンテキストを使用して、ユーザーの質問に対する回答を生成します。

ステップ 4: リアルタイムの記述的データのナレッジベースを作成する

次に、リアルタイムの記述的な臨床医と患者のやり取りに関するメモ、診断画像評価、およびラボ分析レポートのナレッジベースを作成します。このナレッジベースはベクトルデータベースです。ベクトルデータベースを使用すると、記述的な医療知識をインデックス化されたベクトル化された形式で保存できるため、医療プロバイダーは膨大なリポジトリから関連情報を効率的にクエリしてアクセスできます。これらのベクトル化された表現は、意味的に類似したデータを取得するのに役立ちます。医療提供者は、臨床記録、医療画像、および検査結果をすばやくナビゲートできます。これにより、コンテキストに関連する情報に瞬時にアクセスできるため、情報に基づいた意思決定が迅速になり、診断や治療計画の精度とスピードが向上します。

OpenSearch Service 医療ナレッジベースの使用

Amazon OpenSearch Service は、大量の高次元医療データを管理できます。これは、高性能な検索とリアルタイム分析を容易にするマネージドサービスです。RAG アプリケーションのベクトルデータベースとして最適です。OpenSearch Service は、医療記録、研究記事、臨床ノートなど、大量の非構造化または半構造化データを管理するバックエンドツールとして機能します。高度なセマンティック検索機能は、コンテキストに関連する情報を取得するのに役立ちます。これにより、臨床意思決定支援システム、患者クエリ解決ツール、医療知識管理システムなどのアプリケーションで特に役立ちます。例えば、臨床医は特定の症状や治療プロトコルに一致する関連する患者データや研究研究をすばやく見つけることができます。これにより、臨床医はup-to-date関連情報に基づいて意思決定を行うことができます。

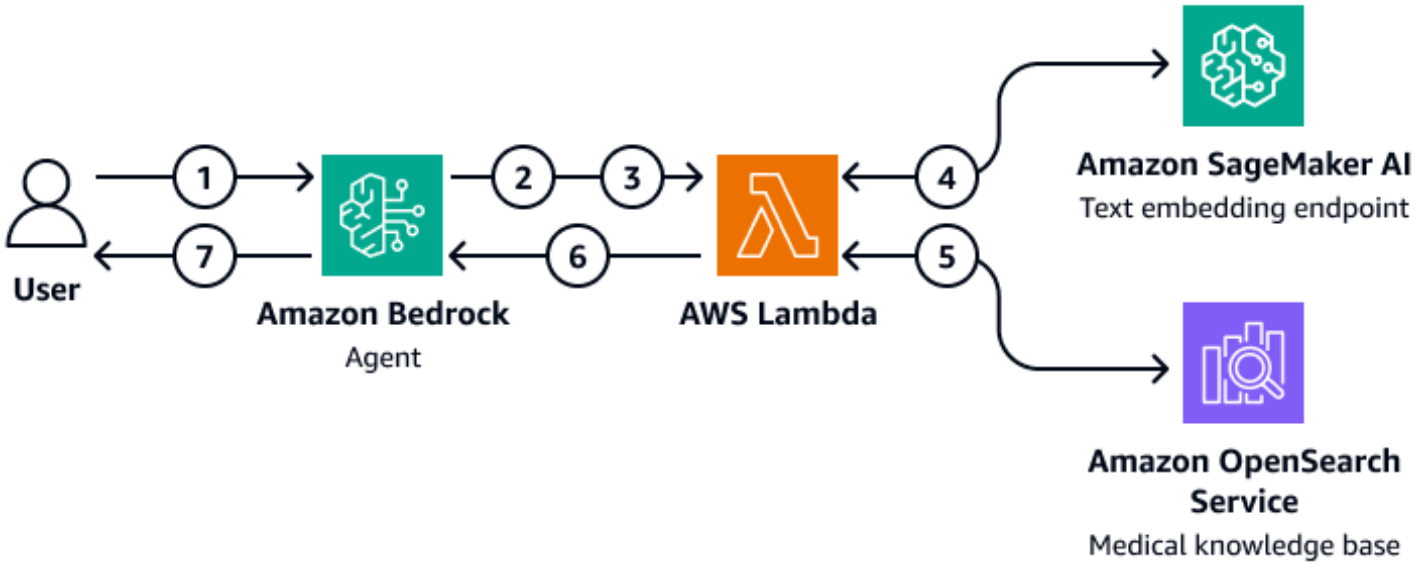
OpenSearch Service は、リアルタイムのデータインデックス作成とクエリをスケーリングして処理できます。これにより、正確な情報へのタイムリーなアクセスが重要な動的な医療環境に最適です。さらに、医療画像や医師のメモなど、複数の入力を必要とする検索に最適なマルチモーダル検索機能を備えています。ヘルスケアアプリケーションに OpenSearch Service を実装する場合、データの

インデックス作成と取得を最適化するために、正確なフィールドとマッピングを定義することが重要です。フィールドは、患者記録、医療履歴、診断コードなど、個々のデータを表します。マッピングは、これらのフィールドの保存方法 (埋め込み形式または元の形式) とクエリの方法を定義します。ヘルスケアアプリケーションでは、構造化データ (数値テスト結果など) 、半構造化データ (患者メモなど) 、非構造化データ (医療画像など) など、さまざまなデータ型に対応するマッピングを確立することが不可欠です。

OpenSearch Service では、厳選されたプロンプトを通じてフルテキストのニューラル検索クエリを実行して、医療記録、臨床ノート、または研究論文を検索し、特定の症状、治療、または患者の履歴に関する関連情報をすばやく見つけることができます。ニューラル検索クエリは、組み込みのニューラルネットワークモデルを使用して、入力プロンプトとイメージの埋め込みを自動的に処理します。これにより、マルチモーダルデータのより深いセマンティック関係を理解してキャプチャし、k-Nearest Neighbor (k-NN) 検索などの他の検索クエリアルゴリズムと比較して、コンテキスト認識が高く正確な検索結果が得られます。

RAG アーキテクチャの作成

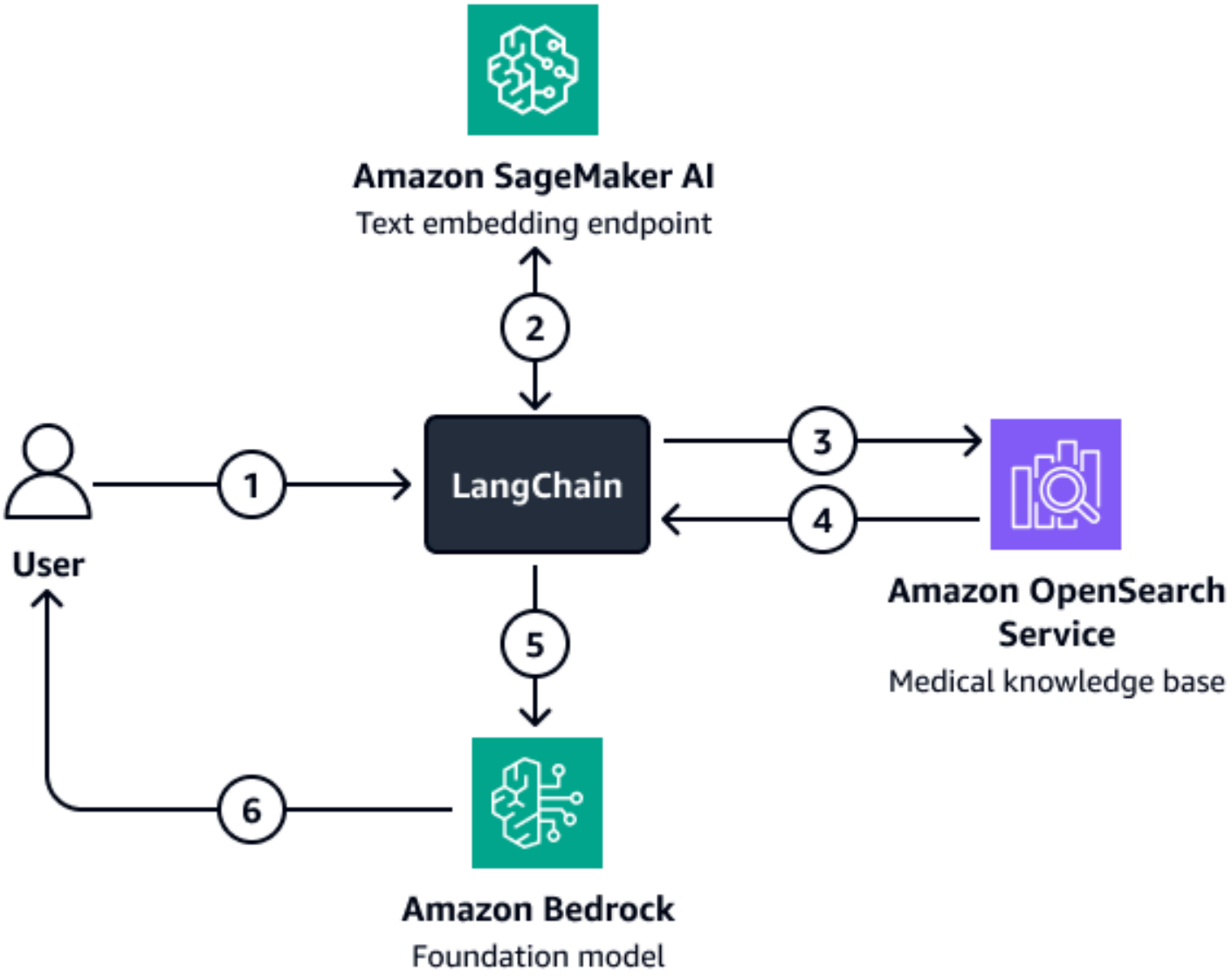
Amazon Bedrock エージェントを使用して OpenSearch Service の医療ナレッジベースをクエリするカスタマイズされた RAG ソリューションをデプロイできます。これを行うには、OpenSearch Service とやり取りしてクエリできる AWS Lambda 関数を作成します。Lambda 関数は、SageMaker AI テキスト埋め込みエンドポイントにアクセスして、ユーザーの入力質問を埋め込みます。Amazon Bedrock エージェントは、追加のクエリパラメータを入力として Lambda 関数に渡します。関数は OpenSearch Service の医療ナレッジベースをクエリし、関連する医療コンテンツを返します。Lambda 関数を設定したら、Amazon Bedrock エージェント内のアクショングループとして追加します。Amazon Bedrock エージェントは、ユーザーの入力を受け取り、必要な変数を識別し、変数と質問を Lambda 関数に渡してから、関数を開始します。関数は、基盤モデルがユーザーの質問に対してより正確な回答を提供するのに役立つコンテキストを返します。



この図表は、次のワークフローを示しています：

1. ユーザーが Amazon Bedrock エージェントに質問を送信します。
2. Amazon Bedrock エージェントは、開始するアクショングループを選択します。
3. Amazon Bedrock エージェントは AWS Lambda 関数を開始し、パラメータを渡します。
4. Lambda 関数は、Amazon SageMaker AI テキスト埋め込みモデルを開始してユーザー質問を埋め込みます。
5. Lambda 関数は、埋め込みテキストと追加のパラメータとフィルターを Amazon OpenSearch Service に渡します。Amazon OpenSearch Service は医療ナレッジベースをクエリし、結果を Lambda 関数に返します。
6. Lambda 関数は、結果を Amazon Bedrock エージェントに返します。
7. Amazon Bedrock エージェントの基盤モデルは、結果に基づいてレスポンスを生成し、ユーザーにレスポンスを返します。

より複雑なフィルタリングが関係する状況では、カスタムリLangChainトリーバーを使用できます。に直接ロードされる OpenSearch Service ベクトル検索クライアントを設定して、このリトリーバーを作成しますLangChain。このアーキテクチャでは、フィルターパラメータを作成するためにより多くの変数を渡すことができます。リトリーバーを設定したら、Amazon Bedrock モデルとリトリーバーを使用して、取得に関する質問への回答チェーンを設定します。このチェーンは、ユーザー入力と潜在的なフィルターをリトリーバーに渡すことで、モデルとリトリーバー間のインタラクションを調整します。リトリーバーは、基盤モデルがユーザーの質問に答えるのに役立つ関連コンテキストを返します。



この図表は、次のワークフローを示しています：

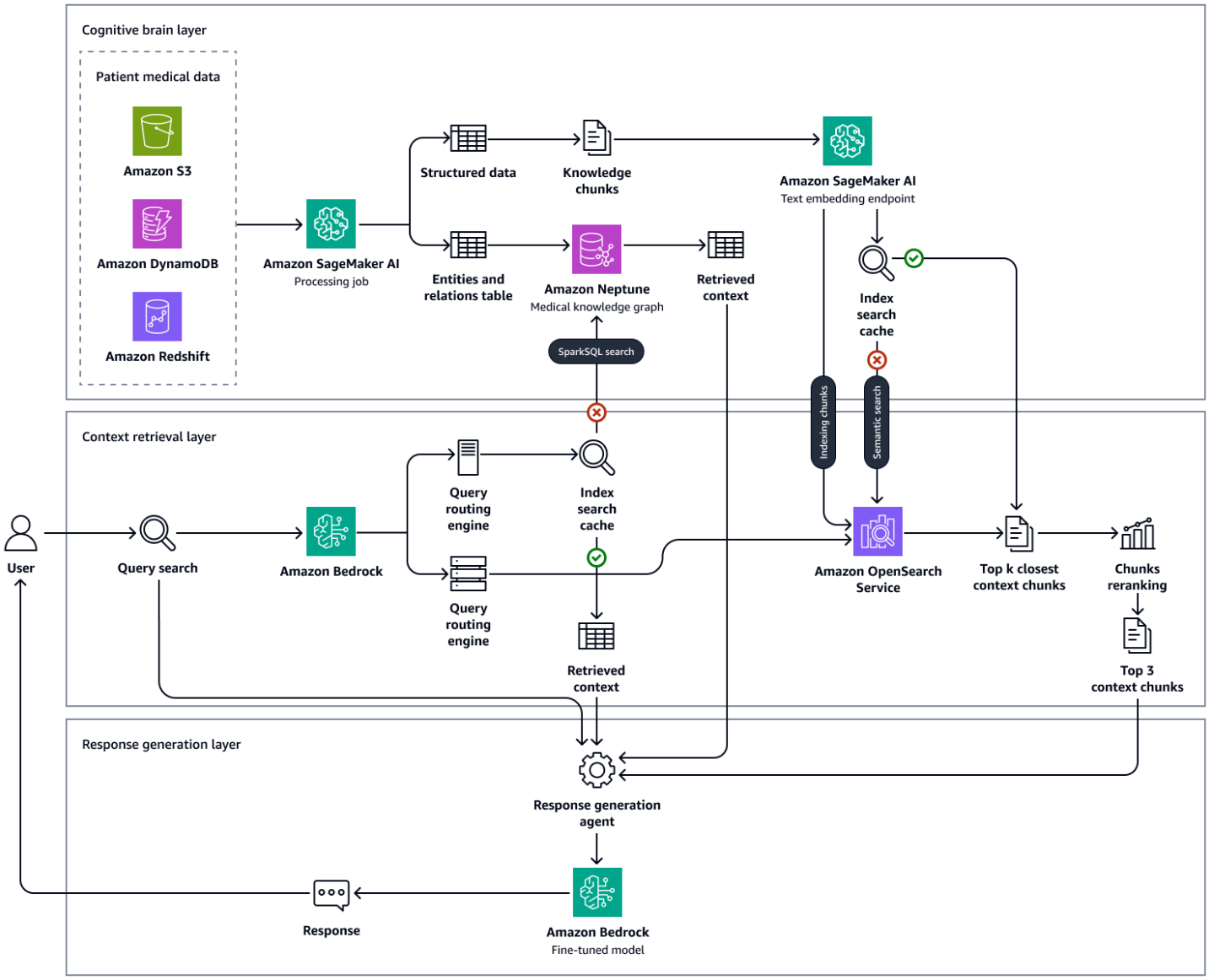
1. ユーザーがリLangChainトリーバーエージェントに質問を送信します。
2. リLangChainトリーバーエージェントは、質問を埋め込むために Amazon SageMaker AI テキスト埋め込みエンドポイントに質問を送信します。
3. リLangChainトリーバーエージェントは埋め込みテキストを Amazon OpenSearch Service に渡します。
4. Amazon OpenSearch Service は、取得したドキュメントをLangChainリトリーバーエージェントに返します。
5. リLangChainトリーバーエージェントは、ユーザーの質問と取得したコンテキストを Amazon Bedrock 基盤モデルに渡します。

6. 基盤モデルはレスポンスを生成し、ユーザーに送信します。

ステップ 5: LLMsを使用して医療上の質問に回答する

前のステップは、患者の医療記録を取得し、関連する薬剤や潜在的な診断を要約できる医療インテリジェンスアプリケーションを構築するのに役立ちます。次に、生成レイヤーを構築します。このレイヤーは、Llama 3 などの Amazon Bedrock の LLM の生成機能を使用して、アプリケーションの出力を補強します。

臨床医がクエリを入力すると、アプリケーションのコンテキスト取得レイヤーはナレッジグラフから取得プロセスを実行し、患者の履歴、人口統計、症状、診断、および結果に関連する上位レコードを返します。また、ベクトルデータベースから、リアルタイムの記述的な医師と患者のやり取りに関するメモ、診断画像評価のインサイト、ラボ分析レポートの概要、医学研究や学術書の膨大なコーパスからのインサイトも取得します。これらの取得された上位の結果、臨床医のクエリ、プロンプト (クエリの性質に基づいて回答をキュレートするように調整) は、Amazon Bedrock の基盤モデルに渡されます。これはレスポンス生成レイヤーです。LLM は、取得したコンテキストを使用して、臨床医のクエリに対するレスポンスを生成します。次の図は、このソリューションのステップのend-to-endのワークフローを示しています。



Amazon Bedrock では、Llama 3 などの事前トレーニング済みの基盤モデルを、医療インテリジェンスアプリケーションが処理する必要があるさまざまなユースケースに使用できます。特定のタスクに最も効果的な LLM は、ユースケースによって異なります。例えば、事前トレーニング済みのモデルは、患者と医師の会話を要約し、薬剤や患者の履歴を検索し、社内の医療データセットや科学的知識からインサイトを取得するのに十分です。ただし、リアルタイムの検査評価、医療処置の推奨事項、患者の成果の予測など、他の複雑なユースケースでは、微調整された LLM が必要になる場合があります。LLM は、医療ドメインデータセットでトレーニングすることで微調整できます。特定の、または複雑なヘルスケアやライフサイエンスの要件により、これらの微調整されたモデルの開発が促進されます。

LLM の微調整または医療ドメインデータでトレーニングされた既存の LLM の選択の詳細については、「Using large language models for healthcare and life science use cases」を参照してください。

AWS Well-Architected フレームワークへの調整

このソリューションは、次のように AWS Well-Architected フレームワークの 6 つの柱すべてと一致します。

- 運用上の優秀性 – アーキテクチャは、効率的なモニタリングと更新のために切り離されます。Amazon Bedrock エージェントと は、ツールを迅速にデプロイおよびロールバックする AWS Lambda のに役立ちます。
- セキュリティ – このソリューションは、HIPAA などの医療規制に準拠するように設計されています。暗号化、きめ細かなアクセスコントロール、Amazon Bedrock ガードレールを実装して、患者データを保護することもできます。
- Amazon OpenSearch Service や Amazon Bedrock などの信頼性 AWS マネージドサービスは、継続的なモデルインタラクションのためのインフラストラクチャを提供します。
- パフォーマンス効率 – RAG ソリューションは、最適化されたセマンティック検索と Cypher クエリを使用して関連データをすばやく取得し、エージェントルーターはユーザークエリに最適なルートを特定します。
- コストの最適化 – Amazon Bedrock および RAG アーキテクチャのpay-per-tokenモデルにより、推論とトレーニング前のコストを削減できます。
- 持続可能性 — サーバーレスインフラストラクチャとpay-per-tokenコンピューティングを使用すると、リソースの使用量が最小限に抑えられ、持続可能性が向上します。

ユースケース: 患者の成果と再入院率の予測

AI を活用した予測分析は、患者の成果を予測し、パーソナライズされた治療計画を可能にすることで、さらにメリットをもたらします。これにより、患者の満足度と健康上の成果を向上させることができます。これらの AI 機能を Amazon Bedrock やその他のテクノロジーと統合することで、医療プロバイダーは大幅な生産性の向上、コストの削減、患者ケアの全体的な品質の向上を実現できます。

患者の履歴、臨床メモ、薬剤、治療などの医療データをナレッジグラフに保存できます。LLMs の深いコンテキスト理解と医療知識グラフの構造化された時間データを組み合わせることで、医療提供者は個々の患者のパターンに関する追加のインサイトを得ることができます。予測分析を使用すると、潜在的な非準拠または治療の複雑さを早期に特定し、パーソナライズされた再アドミッション傾向スコアを生成できます。

このソリューションは、再入院の可能性を予測するのに役立ちます。これらの予測により、患者の成果が向上し、医療コストを削減できます。このソリューションは、病院の臨床医や管理者が再入院のリスクが高い患者に注意を向けるのにも役立ちます。また、アラート、セルフサービス、データ駆動型のアクションを通じて、これらの患者との積極的な介入を開始するのにも役立ちます。

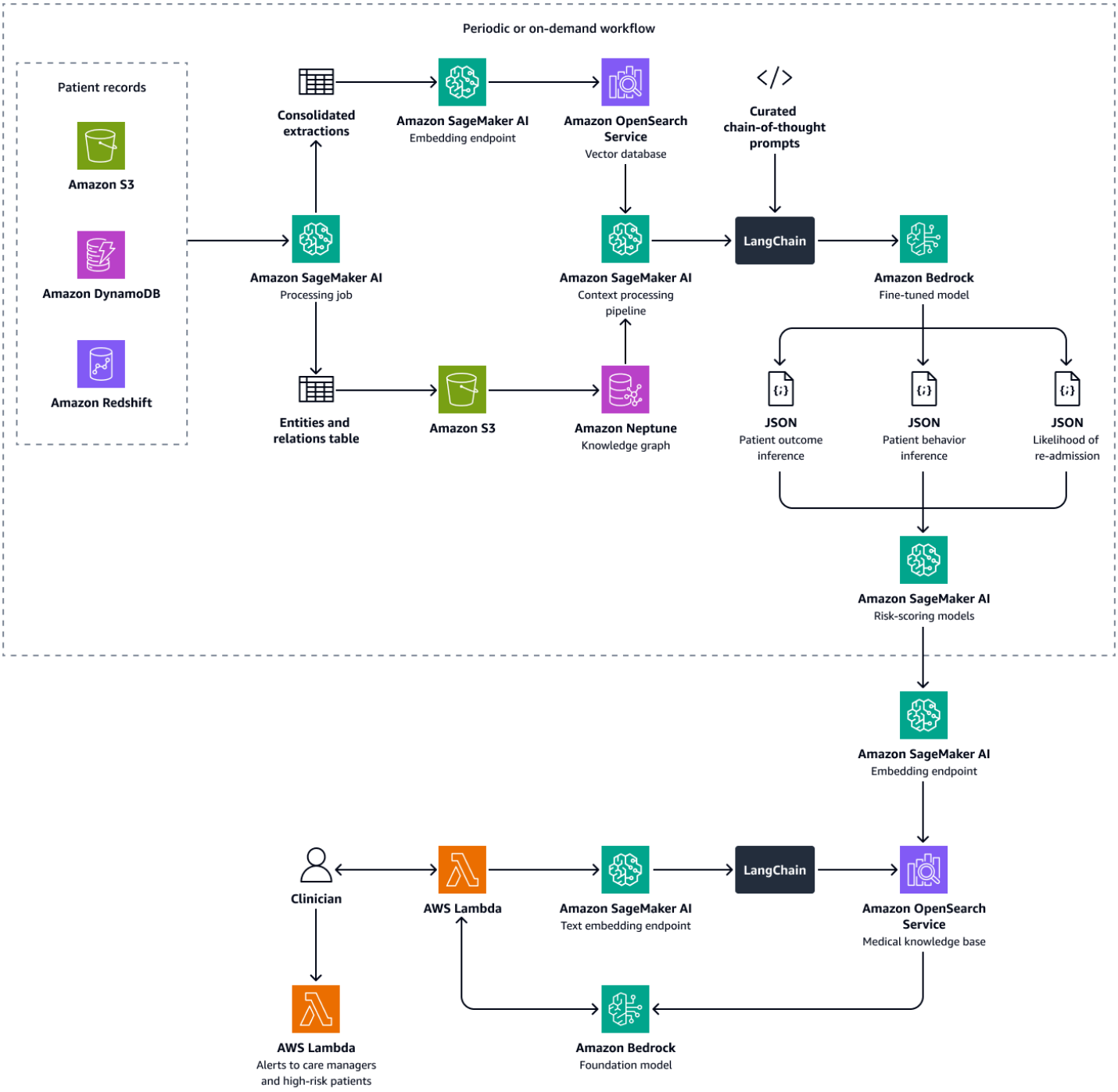
ソリューションの概要

このソリューションは、マルチリトリーバー取得拡張生成 (RAG) フレームワークを使用して、患者データを分析します。個々の患者の再入院の可能性を予測し、病院レベルの再入院傾向スコアを計算するのに役立ちます。このソリューションには、次の機能が統合されています。

- ナレッジグラフ – 入院、以前の再入院、症状、検査結果、処方された治療、薬剤遵守履歴などの構造化された時系列患者データを保存します。
- ベクトルデータベース – 退床サマリー、医師のメモ、予約ミスや報告された薬剤副作用の記録など、構造化されていない臨床データを保存します。
- 微調整された LLM – 患者の行動、治療の遵守状況、再入院の可能性に関する推論を生成するために、ナレッジグラフの構造化データとベクトルデータベースの非構造化データの両方を消費します。

リスクスコアモデルは、LLM からの推論を数値スコアに定量化します。スコアを病院レベルの再入院傾向スコアに集計できます。このスコアは、各患者のリスクエクスポージャーを定義し、定期的に、または必要に応じて計算できます。すべての推論とリスクスコアはインデックス化され、Amazon OpenSearch Service に保存されるため、ケアマネージャーと臨床医はそれを取得できます。会話型 AI エージェントをこのベクトルデータベースと統合することで、臨床医とケアマネー

ジャーは個々の患者レベル、施設全体、または医療専門分野別にシームレスにインサイトを抽出できます。また、リスクスコアに基づいて自動アラートを設定して、プロアクティブな介入を促すこともできます。



このソリューションの構築は、次のステップで構成されます。

- ステップ 1: 医療知識グラフを使用して患者の成果を予測する

ステップ 1: 医療知識グラフを使用して患者の成果を予測する

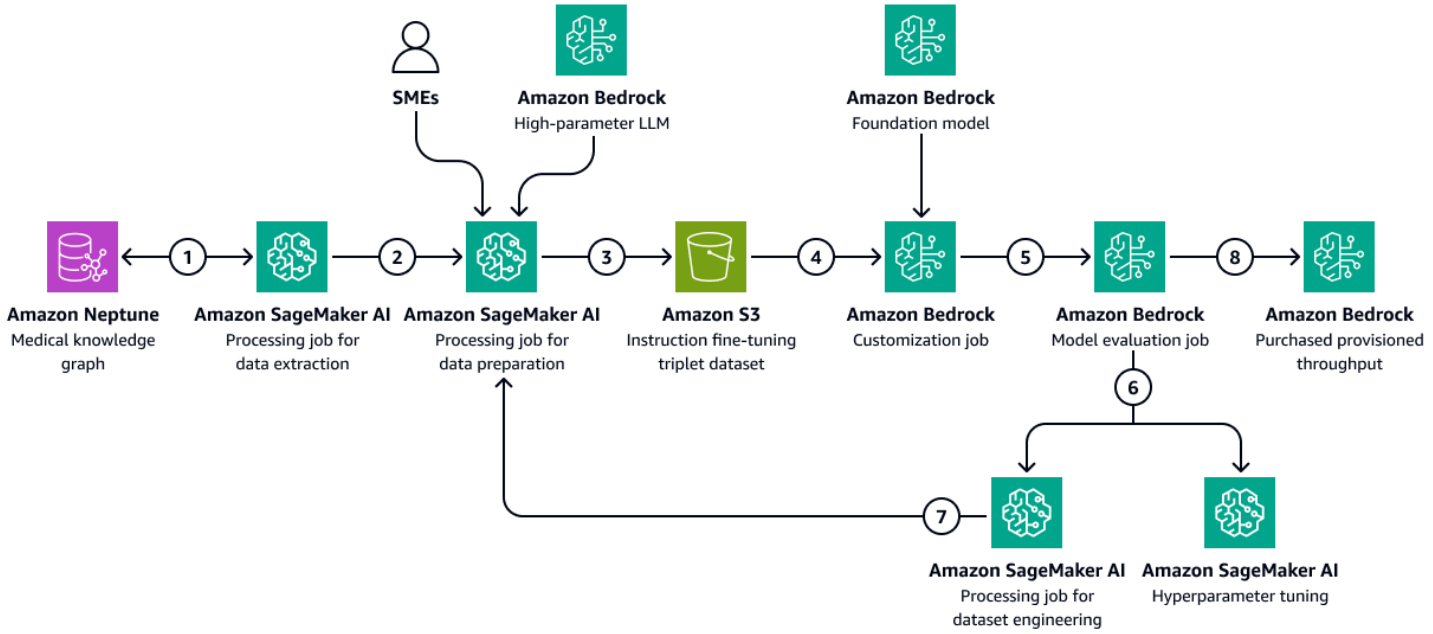
Amazon Neptune では、ナレッジグラフを使用して、患者の診察と結果に関する一時的な知識を経時的に保存できます。ナレッジグラフを構築して保存する最も効果的な方法は、グラフモデルとグラフデータベースを使用することです。グラフデータベースは、関係を保存およびナビゲートする目的で構築されています。グラフデータベースを使用すると、高度に接続されたデータのモデル化と管理が容易になり、柔軟なスキーマを持つことができます。

ナレッジグラフは、時系列分析の実行に役立ちます。以下は、患者の結果の一時的な予測に使用されるグラフデータベースの主要な要素です。

- 履歴データ - 患者の以前の診断、継続的な薬剤、以前に使用した薬剤、および検査結果
- 患者の訪問 (時系列) – 訪問日、症状、観察されたアレルギー、臨床記録、診断、処置、治療、処方された薬剤、および検査結果
- 症状と臨床パラメータ – 重大度、進行パターン、薬剤に対する患者の反応など、臨床および症状ベースの情報

医療知識グラフからのインサイトを使用して、Llama 3 などの Amazon Bedrock の LLM を微調整できます。LLM は、一連の薬剤や治療に対する患者の反応に関するシーケンシャルな患者データで微調整します。の一連の薬剤または治療と患者と臨床の相互作用データを、患者の健康状態を示す事前定義されたカテゴリに分類するラベル付きデータセットを使用します。これらのカテゴリの例としては、健康状態の悪化、改善、または安定した進行などがあります。臨床医が患者とその症状に関する新しいコンテキストを入力すると、微調整された LLM はトレーニングデータセットのパターンを使用して、潜在的な患者の結果を予測できます。

次の図は、医療固有のトレーニングデータセットを使用して Amazon Bedrock で LLM をファインチューニングする手順を示しています。このデータには、患者の病状や時間の経過に伴う治療への反応が含まれる場合があります。このトレーニングデータセットは、モデルが患者の結果について一般化された予測を行うのに役立ちます。



この図表は、次のワークフローを示しています：

1. Amazon SageMaker AI データ抽出ジョブは、ナレッジグラフをクエリして、一連の薬剤または治療に対するさまざまな患者の応答に関する時系列データを取得します。
2. SageMaker AI データ準備ジョブは、Amazon Bedrock LLM と対象分野のエキスパート (SMEs。ジョブは、ナレッジグラフから取得したデータを、各患者のヘルスステータスを示す事前定義されたカテゴリ (ヘルスの悪化、改善、安定した進行状況など) に分類します。
3. ジョブは、ナレッジグラフから抽出された情報、chain-of-thoughtプロンプト、および患者結果カテゴリを含むファインチューニングデータセットを作成します。トレーニングデータセットを Amazon S3 バケットにアップロードします。
4. Amazon Bedrock カスタマイズジョブは、このトレーニングデータセットを使用して LLM を微調整します。
5. Amazon Bedrock カスタマイズジョブは、選択した Amazon Bedrock 基盤モデルをトレーニング環境に統合します。微調整ジョブが開始され、設定したトレーニングデータセットとトレーニングハイパーパラメータが使用されます。
6. Amazon Bedrock 評価ジョブは、事前に設計されたモデル評価フレームワークを使用して、微調整されたモデルを評価します。
7. モデルを改善する必要がある場合、トレーニングデータセットを慎重に検討した後、トレーニングジョブはより多くのデータで再度実行されます。モデルが増分パフォーマンスの向上を示していない場合は、トレーニングハイパーパラメータの変更も検討してください。

8. モデル評価がビジネスステークホルダーによって定義された基準を満たしたら、微調整されたモデルを Amazon Bedrock のプロビジョニングされたスループットにホストします。

ステップ 2: 処方された薬剤または治療に対する患者の行動を予測する

微調整された LLMs は、一時的な医療知識グラフから臨床メモ、退床概要、およびその他の患者固有のドキュメントを処理できます。患者は、処方された薬剤や治療に従う可能性が高いかどうかを評価できます。

このステップでは、 で作成されたナレッジグラフを使用しますステップ 1: 医療知識グラフを使用して患者の成果を予測する。ナレッジグラフには、ノードとしての患者の過去の準拠状況など、患者のプロファイルからのデータが含まれます。また、医薬品や治療への非準拠、医薬品への副作用、医薬品へのアクセス不足やコスト障壁、このようなノードの属性としての複雑な治療計画も含まれます。

微調整された LLMs は、医療知識グラフの過去の処方フルフィルメントデータと、Amazon OpenSearch Service ベクトルデータベースの臨床メモの記述的な概要を消費できます。これらの臨床メモには、頻繁に予約を見逃したり、治療へのコンプライアンス違反があったりすることが記載されている場合があります。LLM は、これらのメモを使用して、将来の非準拠の可能性を予測できます。

1. 次のように入力データを準備します。
   - 構造化データ – 過去 3 回の訪問や検査結果などの最近の患者データを医療知識グラフから抽出します。
   - 非構造化データ – Amazon OpenSearch Service ベクトルデータベースから最新の臨床メモを取得します。
2. 患者の履歴と現在のコンテキストを含む入力プロンプトを作成します。プロンプトの例を次に示します。

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

### **Patient Details**
- **Patient ID:** {patient_id}
- **Age:** {age}
- **Gender:** {gender}
```

```
- **Medical Conditions:** {medical_conditions}
- **Current Medications:** {current_medications}
- **Prescribed Treatments:** {prescribed_treatments}

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

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

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

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

3. プロンプトを微調整された LLM に渡します。LLM はプロンプトを処理し、結果を予測します。LLM からのレスポンスの例を次に示します。

```
{
  "patient_id": "P12345",
  "likelihood_of_non_adherence": "high",
```
```

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

4. モデルのレスポンスを解析して、予測された結果カテゴリを抽出します。たとえば、前のステップのレスポンス例のカテゴリは、非準拠である可能性が高い可能性があります。
5. (オプション) モデルログまたは追加の方法を使用して、信頼スコアを割り当てます。ログは、特定のクラスまたはカテゴリに属する項目の正規化されていない確率です。

ステップ 3: 患者の再入院の可能性を予測する

医療管理のコストが高く、患者の健康に影響するため、病院の再入室は大きな懸念事項です。再入院率の計算は、患者のケアの質と医療提供者のパフォーマンスを測定する方法の 1 つです。

再入院率を計算するために、7 日間の再入院率などのインジケータを定義しました。このインジケータは、入院した患者のうち、7 日以内に予定外の訪問のために再入院した患者の割合を示します。患者の再入院の可能性を予測するために、微調整された LLM は、 で作成した医療知識グラフの時間データを消費できますステップ 1: 医療知識グラフを使用して患者の成果を予測する。このナレッジグラフは、患者の診察、処置、薬剤、症状の時系列記録を維持します。これらのデータレコードには以下が含まれます。

- 患者の最後の退床からの期間
- 過去の治療や薬剤に対する患者の反応
- 時間の経過に伴う症状または状態の進行

これらの時系列イベントを処理して、キュレートされたシステムプロンプトを通じて患者の再入院の可能性を予測できます。プロンプトは、微調整された LLM に予測ロジックを付与します。

1. 次のように入力データを準備します。
   - 準拠履歴 – 医療知識グラフから、薬剤の集荷日、薬剤補充の頻度、診断と薬剤の詳細、時系列の医療履歴、その他の情報を抽出します。
   - 行動指標 - 予約ミスや患者報告の副作用に関する臨床記録を取得して含めます。
2. 準拠履歴と動作インジケータを含む入力プロンプトを作成します。プロンプトの例を次に示します。

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

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

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

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

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

### **Output Format (JSON)**
Return the prediction in the following structured format:
```json
```
```

```
{
  "patient_id": "{patient_id}",
  "likelihood_of_readmission": "{low | moderate | high}",
  "reasoning": "{detailed_explanation_based_on_patient_history}"
}
```

3. ファインチューニングされた LLM にプロンプトを渡します。LLM はプロンプトを処理し、再入院の可能性と理由を予測します。LLM からのレスポンスの例を次に示します。

```
{
  "patient_id": "P67890",
  "likelihood_of_readmission": "high",
  "reasoning": "The patient was discharged only 5 days ago, has a history of more
 than two readmissions to hospitals where the patient received treatment. Recent
 lab results indicate abnormal kidney function and high liver enzymes. These factors
 suggest a medium risk of readmission."
}
```

4. 予測を低、中、高などの標準化されたスケールに分類します。
5. LLM が提供する推論を確認し、予測に寄与する主要な要因を特定します。
6. 定性的出力を量的スコアにマッピングします。例えば、非常に高い場合、確率は 0.9 になります。
7. 検証データセットを使用して、実際の再アドミッション率に対してモデル出力をキャリブレーションします。

ステップ 4: 再入院傾向スコアを計算する

次に、患者ごとに再入院傾向スコアを計算します。このスコアは、前のステップで実行された 3 つの分析の正味の影響を反映しています。考えられる患者の成果、薬剤や治療に対する患者の行動、患者の再入院の可能性です。患者レベルの再入院傾向スコアを専門分野レベルに集約し、次に病院レベルで集計することで、臨床医、ケアマネージャー、管理者に関するインサイトを得ることができます。再入院傾向スコアは、施設、専門分野、または状態別に全体的なパフォーマンスを評価するのに役立ちます。次に、このスコアを使用してプロアクティブな介入を実装できます。

1. さまざまな要因 (結果予測、準拠可能性、再入院) に重みを割り当てます。重みの例を次に示します。
   - 結果予測の重み: 0.4
   - 準拠予測の重み: 0.3

- 再入院の可能性の重み: 0.3

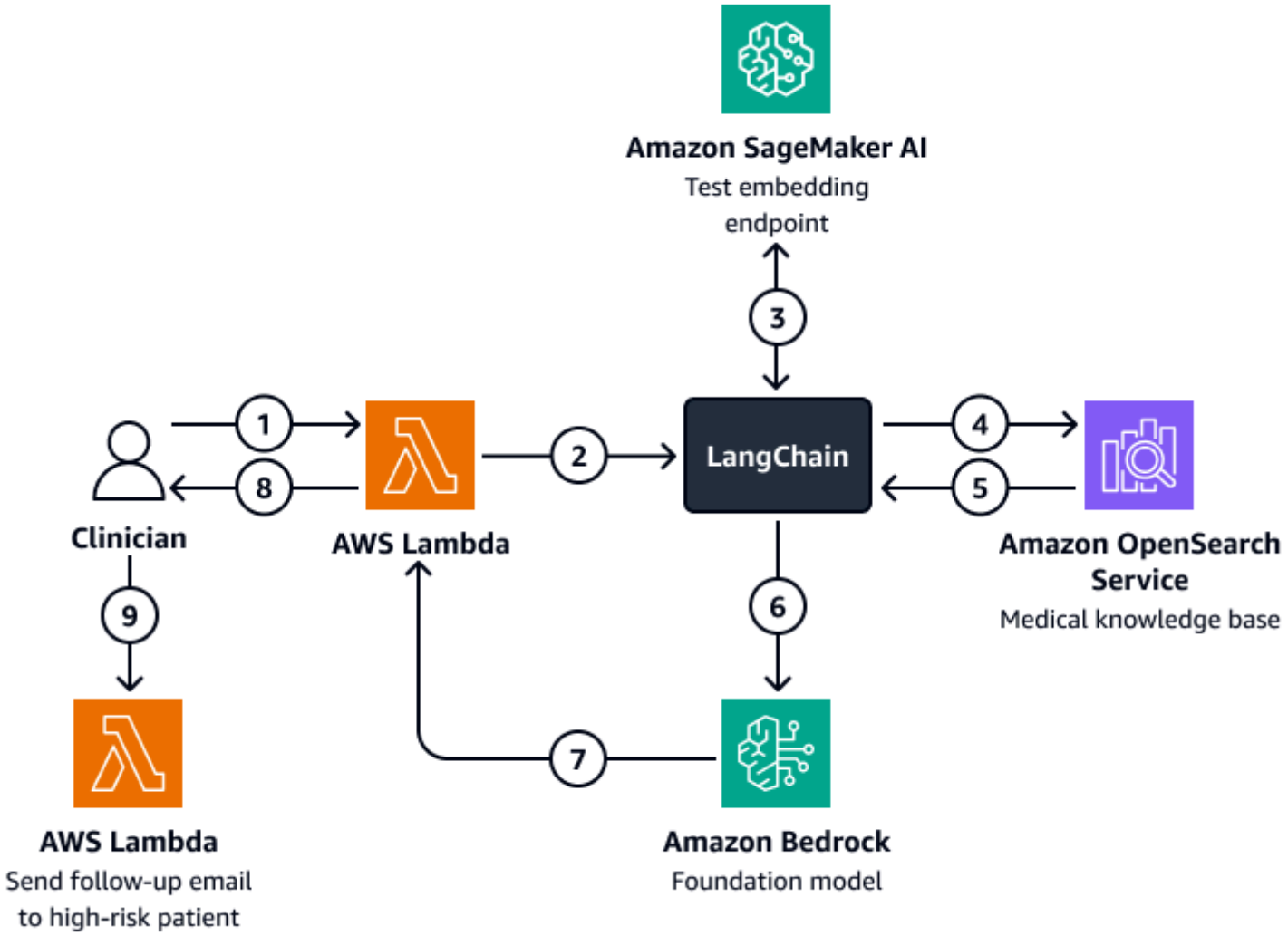
2. 複合スコアを計算するには、次の計算を使用します。

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

3. 0 ~ 1 など、すべての個々のスコアが同じスケールであることを確認します。
4. アクションのしきい値を定義します。たとえば、0.7 を超えるスコアはアラートを開始します。

上記の分析と患者の再入院傾向スコアに基づいて、臨床医またはケアマネージャーは、計算されたスコアに基づいて個々の患者を監視するアラートを設定できます。事前定義されたしきい値を超えると、そのしきい値に達したときに通知されます。これにより、ケアマネージャーは患者の退床ケアプランを作成する際に、事後対応的ではなく事前対応的になることができます。患者の結果、動作、再入院傾向スコアをインデックス付き形式で Amazon OpenSearch Service ベクトルデータベースに保存して、ケアマネージャーが会話 AI エージェントを使用してシームレスに取得できるようにします。

次の図は、臨床医またはケアマネージャーが患者の成果、予想される動作、再入院傾向に関するインサイトを取得するために使用できる会話型 AI エージェントのワークフローを示しています。ユーザーは、患者レベル、部門レベル、または病院レベルでインサイトを取得できます。AI エージェントはこれらのインサイトを取得し、Amazon OpenSearch Service ベクトルデータベースのインデックス付き形式で保存されます。エージェントはクエリを使用して関連データを取得し、再入院のリスクが高い患者に対して推奨されるアクションなど、カスタマイズされたレスポンスを提供します。リスクのレベルに基づいて、エージェントは患者とケア提供者のリマインダーを設定することもできます。



この図表は、次のワークフローを示しています：

1. 臨床医は、 関数を格納する会話型 AI エージェントに質問をします AWS Lambda 。
2. Lambda 関数はLangChainエージェントを開始します。
3. LangChain エージェントは、ユーザーの質問を Amazon SageMaker AI テキスト埋め込みエンドポイントに送信します。エンドポイントは質問を埋め込みます。
4. LangChain エージェントは埋め込み質問を Amazon OpenSearch Service の医療ナレッジベースに渡します。
5. Amazon OpenSearch Service は、ユーザークエリに最も関連性の高い特定のインサイトを LangChainエージェントに返します。
6. LangChain エージェントは、ナレッジベースから取得したクエリとコンテキストを Amazon Bedrock 基盤モデルに送信します。
7. Amazon Bedrock 基盤モデルはレスポンスを生成し、Lambda 関数に送信します。

8. Lambda 関数は、臨床医にレスポンスを返します。
9. 臨床医は、再入院のリスクが高い患者にフォローメールを送信する Lambda 関数を開始します。

AWS Well-Architected フレームワークへの調整

患者の行動を追跡し、再入院率を予測するアーキテクチャは AWS のサービス、AWS Well-Architected フレームワークの 6 つの柱に沿って、医療成果を向上させるために、、、医療知識グラフ、LLMs を統合します。

- 運用上の優秀性 – このソリューションは、Amazon Bedrock と を使用してリアルタイムのアラート AWS Lambda を行う、分離された自動化されたシステムです。
- セキュリティ – このソリューションは、HIPAA などの医療規制に準拠するように設計されています。暗号化、きめ細かなアクセスコントロール、Amazon Bedrock ガードレールを実装して、患者データを保護することもできます。
- 信頼性 – このアーキテクチャでは、耐障害性のあるサーバーレスを使用します AWS のサービス。
- パフォーマンス効率 – Amazon OpenSearch Service と微調整された LLMs は、迅速かつ正確な予測を提供できます。
- コストの最適化 – サーバーレステクノロジーとpay-per-inferenceモデルは、コストを最小限に抑えるのに役立ちます。微調整された LLM を使用すると追加料金が発生する可能性がありますが、モデルは微調整プロセスに必要なデータと計算時間を短縮する RAG アプローチを使用します。
- 持続可能性 – このアーキテクチャは、サーバーレスインフラストラクチャを使用してリソースの消費を最小限に抑えます。また、効率的でスケーラブルな医療オペレーションもサポートしています。

ユースケース: 医療スタッフの管理とスキル向上

人材トランスフォーメーション戦略とスキル向上戦略を実装することで、ワークフォースは医療および医療サービスで新しいテクノロジーとプラクティスの使用に精通し続けることができます。プロアクティブなスキル向上イニシアチブにより、医療専門家は高品質の患者ケアを提供し、運用効率を最適化し、規制基準に準拠し続けることができます。さらに、人材トランスフォーメーションは継続的な学習の文化を育みます。これは、変化する医療状況に適応し、新たなパブリックヘルスの課題に対処するために不可欠です。クラスルームベースのトレーニングや静的学習モジュールなどの従来のトレーニングアプローチは、幅広い視聴者に統一されたコンテンツを提供します。多くの場合、個々の実務者の特定のニーズと習熟度に対処するために不可欠な、パーソナライズされた学習パスがありません。この one-size-fits-all戦略は、エンゲージメントを失い、知識保持が最適でなくなる可能性があります。

したがって、医療組織は、現在の状態と潜在的な将来の状態における各従業員のギャップを決定できる、画期的でスケーラブルなテクノロジー主導のソリューションを採用する必要があります。これらのソリューションは、ハイパーパーソナライズされた学習パスと適切な学習コンテンツのセットを推奨する必要があります。これにより、医療の未来に向けてワークフォースが効果的に準備されます。

ヘルスケア業界では、生成 AI を適用して、ワークフォースの理解とスキル向上に役立てることができます。大規模言語モデル (LLMs) と高度なリトリーバーを接続することで、組織は現在どのようなスキルを持っているかを理解し、将来必要になる可能性のある主要なスキルを特定できます。この情報は、新しいワーカーを雇用し、現在のワークフォースのスキルを向上させることで、ギャップを埋めるのに役立ちます。Amazon Bedrock とナレッジグラフを使用すると、医療組織は継続的な学習とスキル開発を容易にするドメイン固有のアプリケーションを開発できます。

このソリューションが提供する知識は、人材の効果的な管理、従業員のパフォーマンスの最適化、組織の成功の促進、既存のスキルの特定、人材戦略の策定に役立ちます。このソリューションは、数か月ではなく数週間でこれらのタスクを実行するのに役立ちます。

ソリューションの概要

このソリューションは、以下のコンポーネントで構成される医療人材変換フレームワークです。

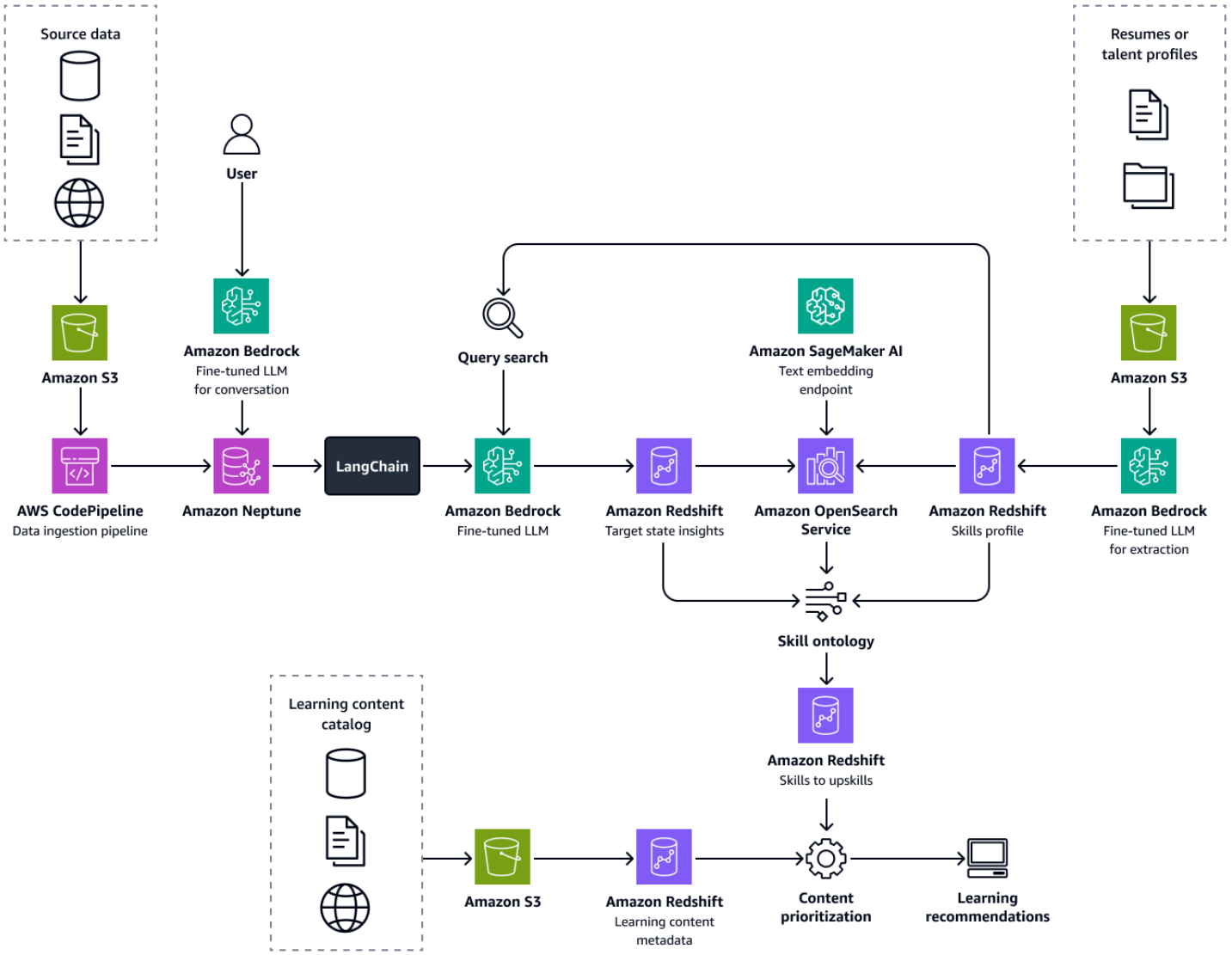
- インテリジェント再開パーサー – このコンポーネントは、候補の再開を読み取って、スキルを含む候補情報を正確に抽出できます。Amazon Bedrock で微調整された Llama 2 モデルを使用して構築されたインテリジェントな情報抽出ソリューションは、19 を超える業界の再開と人材プロファイルをカバーする独自のトレーニングデータセット上に構築されています。この LLM ベースのプ

ロセスは、再開の手動レビュープロセスを自動化し、トップ候補をマッチングしてロールを開くことで、数百時間を節約します。

- ナレッジグラフ – Amazon Neptune 上に構築されたナレッジグラフ。Amazon Neptune は、組織および業界の役割とスキルの分類を含む人材情報の統一されたリポジトリであり、スキル、役割とその特性、関係、論理的制約の定義を使用して医療人材のセマンティクスをキャプチャします。
- スキルオントロジー – 候補スキルと理想的な現在の状態または将来の状態スキル (ナレッジグラフを使用して取得) の間のスキル近接性の発見は、候補スキルとターゲット状態スキルの間の意味論的類似性を測定するオントロジーアルゴリズムを通じて達成されます。
- 学習経路とコンテンツ – このコンポーネントは、特定されたスキルギャップに基づいて、任意のベンダーからの学習マテリアルのカタログから適切な学習コンテンツを推奨する学習レコメンデーションエンジンです。スキルギャップを分析し、優先順位付けされた学習コンテンツを推奨することで、各候補に最適なスキルアップパスを特定し、新しいロールへの移行中に各候補のシームレスで継続的な専門的な開発を可能にします。

このクラウドベースの自動化されたソリューションは、機械学習サービス、LLMs、ナレッジグラフ、検索拡張生成 (RAG) を利用しています。数十または数千の再開を最小限の時間で処理し、即時の候補プロファイルを作成し、現在または将来の状態のギャップを特定し、これらのギャップを埋めるために適切な学習コンテンツを効率的に推奨できます。

次の図は、フレームワークのend-to-endフローを示しています。このソリューションは、Amazon Bedrock の微調整された LLMs 上に構築されています。これらの LLMs、Amazon Neptune の医療人材ナレッジベースからデータを取得します。データ駆動型アルゴリズムは、候補ごとに最適な学習パスのレコメンデーションを行います。



このソリューションの構築は、次のステップで構成されます。

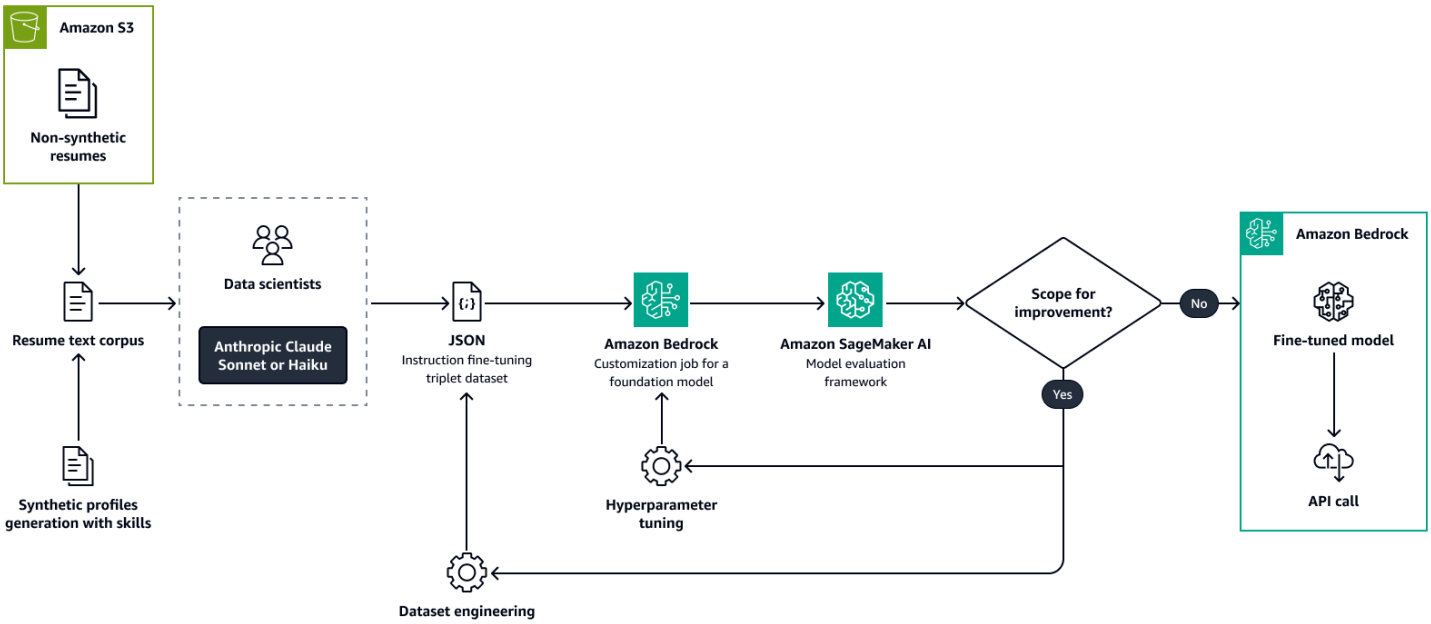
- ステップ 1: 人材情報を抽出し、スキルプロファイルを構築する
- ステップ 2: ナレッジグラフからrole-to-skillの関連性を検出する
- ステップ 3: スキルギャップの特定とトレーニングの推奨

ステップ 1: 人材情報を抽出し、スキルプロファイルを構築する

まず、カスタムデータセットを使用して Amazon Bedrock で Llama 2 などの大規模言語モデルを微調整します。これにより、LLM がユースケースに適応します。トレーニング中、候補者の職務再開や同様の人材プロファイルから主要な人材属性を正確かつ一貫して抽出します。これらの人材属性には、スキル、現在のロールタイトル、日付スパンを含む経験タイトル、教育、認定が含まれます。

詳細については、Amazon Bedrock ドキュメントの「モデルをカスタマイズしてユースケースのパフォーマンスを向上させる」を参照してください。

次の図は、Amazon Bedrock を使用して再開解析モデルを微調整するプロセスを示しています。キー情報を抽出するために、実際の再開と合成で作成された再開の両方が LLM に渡されます。データサイエンティストのグループは、抽出された情報を元の未加工テキストと照合します。次に、chain-of-thoughtプロンプトと元のテキストを使用して抽出された情報を連結し、ファインチューニング用のトレーニングデータセットを取得します。このデータセットは、モデルを微調整する Amazon Bedrock カスタマイズジョブに渡されます。Amazon SageMaker AI バッチジョブは、微調整されたモデルを評価するモデル評価フレームワークを実行します。モデルを改善する必要がある場合、ジョブはより多くのデータまたは異なるハイパーパラメータで再度実行されます。評価が基準を満たしたら、Amazon Bedrock のプロビジョニングされたスループットを通じてカスタムモデルをホストします。



ステップ 2: ナレッジグラフからrole-to-skillの関連性を検出する

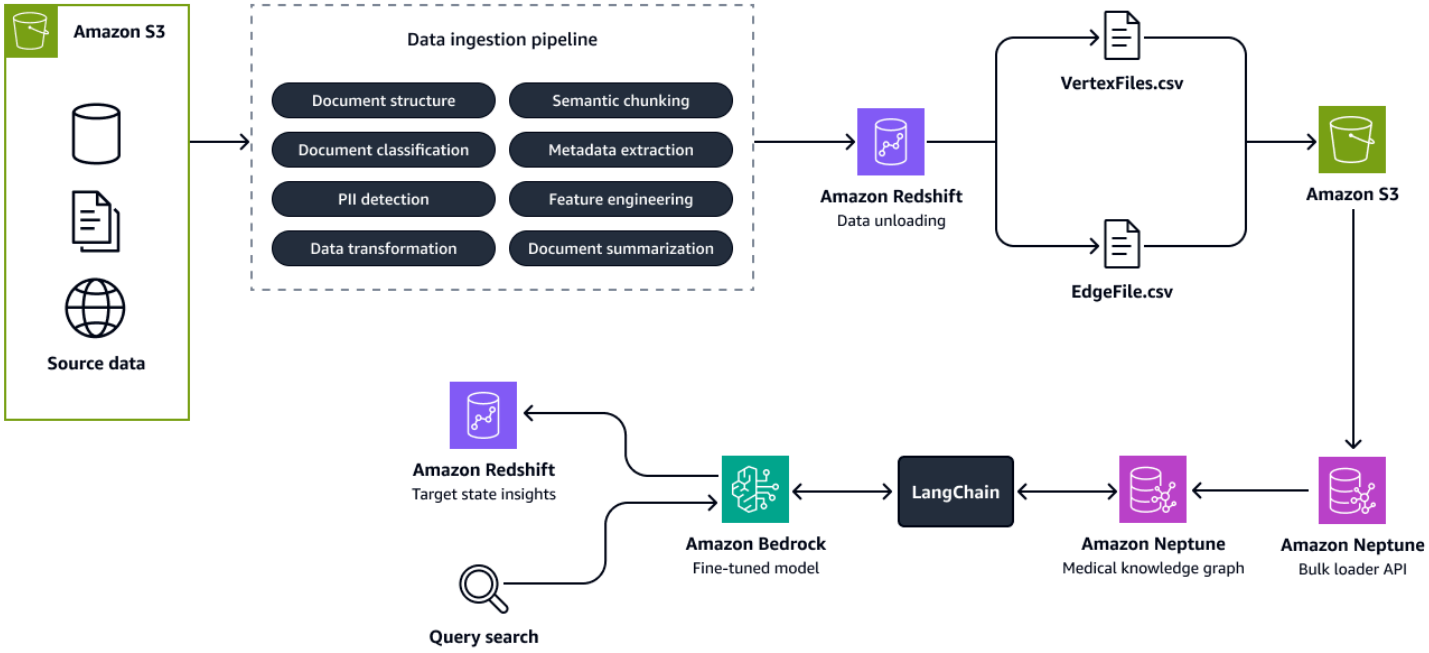
次に、組織およびヘルスケア業界の他の組織のスキルと役割の分類をカプセル化するナレッジグラフを作成します。この強化されたナレッジベースは、Amazon Redshift の集計された人材と組織データから取得されます。さまざまな労働市場データプロバイダーや、エンタープライズリソースプランニング (ERP) システム、人事情報システム (HRIS)、従業員の職務再開、職務記述書、人材アーキテクチャドキュメントなど、組織固有の構造化および非構造化データソースから人材データを収集できます。

Amazon Neptune でナレッジグラフを構築します。ノードはスキルとロールを表し、エッジはそれらの関係を表します。このグラフにメタデータを追加して、組織名、業界、ジョブファミリー、スキルタイプ、ロールタイプ、業界タグなどの詳細を含めます。

次に、Graph Retrieval Augmented Generation (Graph RAG) アプリケーションを開発します。Graph RAG は、グラフデータベースからデータを取得する RAG アプローチです。Graph RAG アプリケーションのコンポーネントは次のとおりです。

- Amazon Bedrock での LLM との統合 – アプリケーションは、自然言語の理解とクエリ生成のために Amazon Bedrock の LLM を使用します。ユーザーは自然言語を使用してシステムとやり取りできます。これにより、技術系以外の利害関係者がアクセスできるようになります。
- オーケストレーションと情報の取得 – LlamaIndex またはLangChainオーケストレーターを使用して、LLM と Neptune ナレッジグラフの統合を容易にします。自然言語クエリを openCypher クエリに変換するプロセスを管理します。次に、ナレッジグラフでクエリを実行します。プロンプトエンジニアリングを使用して、openCypher クエリを構築するためのベストプラクティスについて LLM に指示します。これにより、クエリを最適化して関連するサブグラフを取得します。サブグラフには、クエリされたロールとスキルに関連するすべてのエンティティと関係が含まれます。
- インサイトの生成 – Amazon Bedrock の LLM は、取得したグラフデータを処理します。現在の状態に関する詳細なインサイトを生成し、クエリされたロールと関連するスキルの将来の状態を予測します。

次の図は、ソースデータからナレッジグラフを構築するステップを示しています。構造化ソースデータと非構造化ソースデータをデータ取り込みパイプラインに渡します。パイプラインは、Amazon Neptune と互換性のある CSV 一括ロードフォーメーションに情報を抽出して変換します。バルクローダー API は、Amazon S3 バケットに保存されている CSV ファイルを Neptune ナレッジグラフにアップロードします。人材の将来の状態、関連するロール、スキルに関連するユーザークエリの場合、Amazon Bedrock の微調整された LLM はLangChainオーケストレーターを介してナレッジグラフとやり取りします。オーケストレーターは、ナレッジグラフから関連するコンテキストを取得し、Amazon Redshift のインサイトテーブルにレスポンスをプッシュします。LangChain オーケストレーターは、GraphQAChain と同様に、自然言語クエリをユーザーから openCypher クエリに変換して、ナレッジグラフをクエリします。Amazon Bedrock の微調整モデルは、取得したコンテキストに基づいてレスポンスを生成します。



ステップ 3: スキルギャップの特定とトレーニングの推奨

このステップでは、医療専門家の現在の状態と将来の潜在的な状態の役割の間の近接性を正確に計算します。これを行うには、個人のスキルセットをジョブロールと比較することで、スキルアフィニティ分析を実行します。Amazon OpenSearch Service ベクトルデータベースには、スキルの説明、スキルタイプ、スキルクラスターなどのスキル分類情報とスキルメタデータを保存します。Amazon Titan Text Embeddings モデルなどの Amazon Bedrock 埋め込みモデルを使用して、識別されたキースキルをベクトルに埋め込みます。ベクトル検索を使用して、現在の状態スキルとターゲット状態スキルの説明を取得し、オントロジー分析を実行します。この分析は、現在の状態とターゲット状態のスキルペア間の近接スコアを提供します。ペアごとに、計算されたオントロジースコアを使用して、スキルアフィニティのギャップを特定します。次に、スキル向上に最適な方法をお勧めします。これは、ロールの移行中に候補が検討できます。

各ロールについて、スキルアップや再スキルのために正しい学習コンテンツを推奨するには、学習コンテンツの包括的なカタログの作成から始まる体系的なアプローチが必要です。このカタログは Amazon Redshift データベースに保存され、さまざまなプロバイダーのコンテンツを集約し、コンテンツ期間、難易度、学習モードなどのメタデータが含まれます。次のステップでは、各コンテンツが提供する主要なスキルを抽出し、ターゲットロールに必要な個々のスキルにマッピングします。このマッピングを実現するには、スキル近接性分析を通じてコンテンツによって提供されるカバレッジを分析します。この分析では、コンテンツによって教えられたスキルが、ロールに必要なスキルとどの程度一致しているかを評価します。メタデータは、各スキルに最適なコンテンツを選択する上で

重要な役割を果たし、学習者が学習ニーズに合ったカスタマイズされたレコメンデーションを確実に受け取るようにします。Amazon Bedrock LLMs を使用して、コンテンツメタデータからスキルを抽出し、特徴量エンジニアリングを実行し、コンテンツのレコメンデーションを検証します。これにより、スキルアップまたは再スキルプロセスの精度と関連性が向上します。

AWS Well-Architected フレームワークへの調整

このソリューションは、AWS Well-Architected フレームワークの 6 つの柱すべてと一致しています。

- 運用上の優秀性 – モジュール式の自動化されたパイプラインにより、運用上の優秀性が向上します。パイプラインの主要なコンポーネントは分離および自動化されるため、モデルの更新が迅速になり、モニタリングが容易になります。さらに、自動トレーニングパイプラインは、微調整されたモデルのより迅速なリリースをサポートします。
- セキュリティ – このソリューションは、再開時のデータや人材プロファイルなど、機密性の高い個人を特定できる情報 (PII) を処理します。AWS Identity and Access Management (IAM) で、きめ細かなアクセスコントロールポリシーを実装し、承認された担当者のみがこのデータにアクセスできることを確認します。
- 信頼性 – このソリューションは、Neptune AWS のサービス、Amazon Bedrock、OpenSearch Service などの を使用して、高需要時でも耐障害性、高可用性、インサイトへの中断のないアクセスを提供します。
- パフォーマンス効率 – Amazon Bedrock および OpenSearch Service ベクトルデータベースの微調整された LLMs は、大規模データセットを迅速かつ正確に処理して、パーソナライズされた学習に関する推奨事項をタイムリーに提供できるように設計されています。
- コスト最適化 – このソリューションは RAG アプローチを使用するため、モデルの継続的な事前トレーニングの必要性が軽減されます。モデル全体を繰り返しファインチューニングする代わりに、システムは再開からの情報の抽出や出力の構造化など、特定のプロセスのみをファインチューニングします。これにより、大幅なコスト削減につながります。リソースを大量に消費するモデルトレーニングの頻度と規模を最小限に抑え、pay-per-useクラウドサービスを使用することで、医療組織は高いパフォーマンスを維持しながら運用コストを最適化できます。
- 持続可能性 – このソリューションは、コンピューティングリソースを動的に割り当てるスケーラブルなクラウドネイティブサービスを使用します。これにより、大規模なデータ集約型の人材変革イニシアチブをサポートしながら、エネルギー消費と環境への影響を軽減できます。

ヘルスケア向けの生成 AI ソリューションの開発とオーケストレーション

このガイドのソリューションを構築するには、微調整された LLMs を使用して拡張された患者データ、臨床および診断上のインサイト、予測される患者成果を医療提供者に提供する RAG アーキテクチャを構築する必要があります。これには、まとまりのある効率的なワークフローを作成するために、複数の ツールと AWS のサービス ツールの統合が必要です。このセクションでは、以下について説明します。

- Amazon Q Developer – Amazon Q Developer を使用して、開発プロセス中のエンジニアリングに関する質問やコードエラーに対処します。
- マルチリトリーバー RAG 設計 – 複数のリトリーバーを使用してユーザーの質問に適した医療コンテキストを取得する RAG ソリューションを設計および実装します。
- ReAct エージェント – 推論と動的アクションを組み合わせたエージェントを実装します。

Amazon Q Developer

生成 AI ソリューションを構築する場合、AI エージェントと接続キーサービスを作成するのは難しい場合があります。ただし、Amazon Q Developer は、高度な生成 AI アシスタントへのアクセスを提供することで、データサイエンティストと AI エンジニアを支援します。Amazon Q は、ユーザーの質問やコードエラーに迅速かつ正確に対処できるため、LLM 開発プロセスの最適化に役立ちます。Amazon Q は、Amazon Bedrock 基盤モデルを使用するアプリケーションを作成するデベロッパーに大きな利点を提供します。ワークフローを合理化し、コード品質を向上させることができます。Python スクリプトとInfrastructure as Code (IaC) 設定の生成を自動化し、開発時間と労力を大幅に削減します。高度なリファクタリング機能を通じて、Amazon Q はコードのパフォーマンスを向上させ、セキュリティの脆弱性を特定し、開発者がベストプラクティスを確実に遵守できるようにします。さらに、コンテキストに応じた提案や説明を提供することで、初心者の学習と導入を容易にし、複雑なコーディングタスクをよりアクセスしやすく効率的にします。

マルチリトリーバー RAG 設計

生成 AI アプリケーションでは、マルチリトリーバー RAG パイプラインが複数のデータソースから情報を効率的に取得できるため、医療提供者や臨床医が医療上の質問に答えるのに役立ちます。このパイプラインは、さまざまなタイプのリトリーバーを使用して、さまざまなナレッジベースから関連

データを取得します。各リトリーバーは、患者の履歴、診断インサイト、臨床ノート、医学研究や学術テキストからのコンテンツなど、特定のタイプの情報の取得に特化しています。

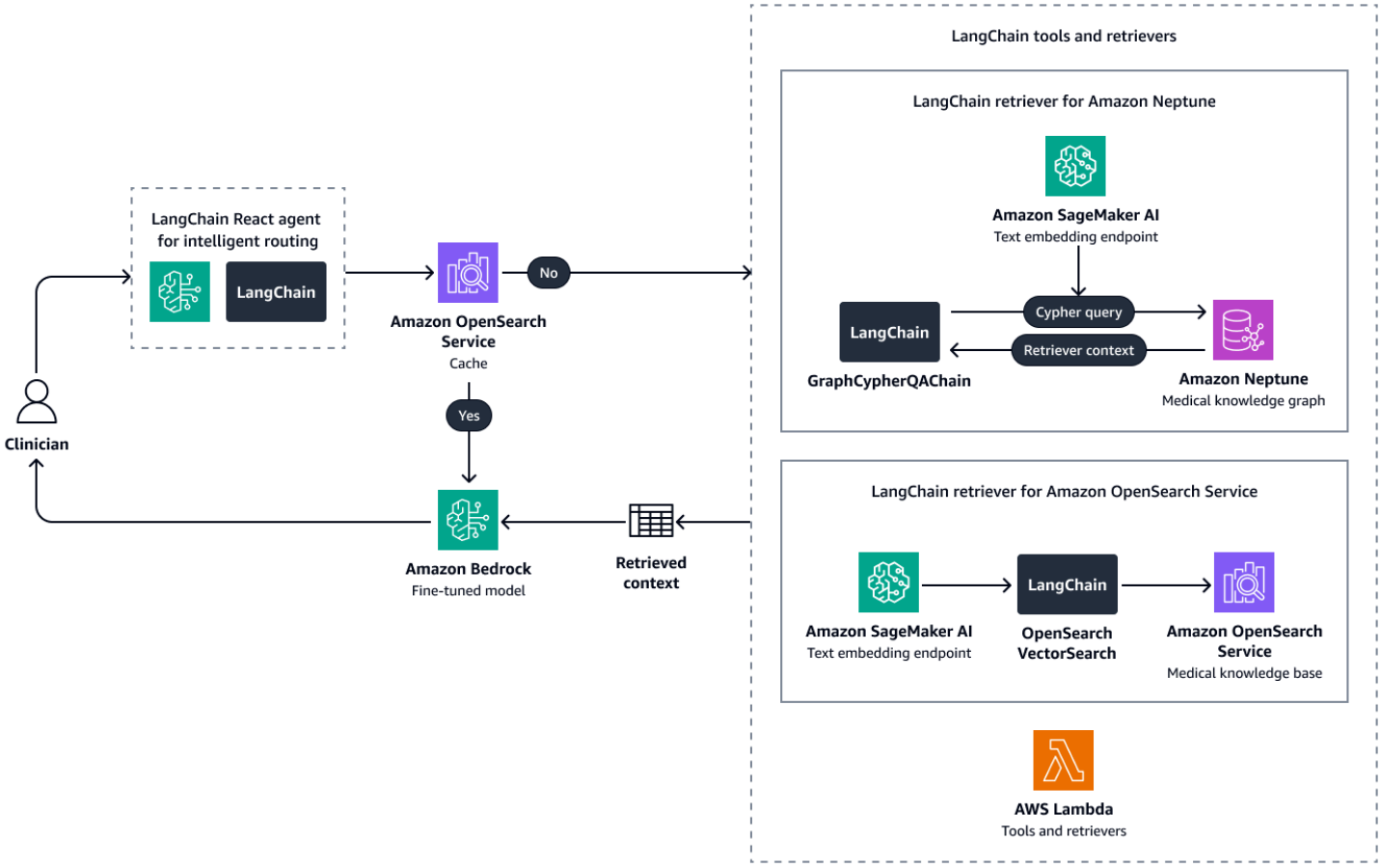
データの性質と特定のアプリケーション要件を使用して、ユースケースに適したバックエンドナレッジベースを決定します。Amazon OpenSearch Service ベクトルデータベースは、画像診断評価概要、退床概要、臨床レポート、医学研究、学術テキストコンテンツなど、非構造化または半構造化の医療データに大量に適しています。一方、Amazon Neptune などのグラフデータベースサービスは、患者、患者の履歴、医療提供者、医薬品、症状、治療などのエンティティ間の時間的関係を深く探索する必要がある医療ユースケースに最適です。

このパイプラインの重要なコンポーネントは、ユーザークエリインテント予測です。これにより、システムはクエリを正しいリトリーバーチェーンにルーティングします。たとえば、臨床医が患者の治療履歴、症状、病院とのやり取り、再入院の可能性、または潜在的な患者の結果について質問した場合、クエリインテント予測モジュールはこのインテントを識別します。これは、医療ナレッジグラフから患者レコードまたは時系列治療データを取得できるリトリーバーチェーンにリクエストを送信します。または、疾患検出、特定の診断評価、または学術教科書からの特定の臨床手順の詳細に関する質問がある場合、クエリは OpenSearch Service ベクトルデータベースからこの情報を取得できるリトリーバーチェーンにルーティングされます。のツール呼び出し機能を使用してLangChain、ユーザー質問を事前定義されたインテントに分類できる Amazon Bedrock LLM にカスタムツールをバインドできます。

このマルチトリーバー RAG システムには、特定のナレッジベースへのアクセスを管理するように設計されたLangChainエージェントが含まれています。LangChain を使用して、Amazon Bedrock LLM、さまざまなリトリーバー、ツール間のインタラクションをオーケストレーションできます。LangChainには、インテント分類子、Neptune のリトリーバー、OpenSearch Service のリトリーバー、または特定のナレッジベースから構造化された形式でユーザーのインテントを分類してデータにアクセスするために開発できるその他のツールなどのカスタムツールを作成するのに役立つツール呼び出しクラスが含まれています。次に、これらのツールを クラスにフィードして、Reasoning and Acting (ReAct) エージェントを作成します。ReAct エージェントは、ユーザーの質問を処理し、質問に回答するためのシーケンシャルステップを計画してから、使用可能なツールを繰り返し実行し、ツールレスポンスを処理してユーザークエリに最後に回答します。

次の図は、効率的なナレッジの取得とインテリジェントなクエリ解決のために設計されたマルチリトリーバー RAG システムの仕組みを示しています。LangChain ReAct エージェントは、ユーザーのインテントを分析し、実行のための構造化計画を策定し、最も関連性の高い取得ツールを選択します。システムは前の質問キャッシュをクエリし、患者 ID、病状、訪問日などの主要な属性に基づいて同様のクエリをチェックします。非常に類似した質問が見つかった場合、対応する回答が直接取得されます。それ以外の場合、エージェントは適切なリトリーバーを実行します。治療履歴、症状、病院

とのやり取り、再入院の可能性など、患者中心の情報を取得するために、システムはグラフリトリーバーを使用します。診断評価、臨床処置、構造化された医学的検出結果のために、エージェントはベクトルデータベースリトリーバーを使用します。両方のデータストアからのコンテキストナレッジの組み合わせを必要とするシナリオでは、包括的なレスポンスを生成するために、システムはナレッジグラフとベクトルデータベースの両方の結果を統合するハイブリッド取得戦略を使用します。



ReAct エージェント

Reasoning and Acting (ReAct) エージェントは、多面的な RAG アプリケーション向けに設計されています。これらのエージェントは、推論と動的アクションの強力な組み合わせを提供します。特に、step-by-stepの論理情報取得ワークフローを含む複雑なアプリケーションに適しています。詳細については、ReAct: Synergizing Reasoning and Acting in Language Models」を参照してください。

医療および医療の文脈では、臨床医または臨床医からのクエリは多面的であることがよくあります。例えば、臨床医が「同様の患者に対してどのような治療が行われたか」と尋ねたとします。ユーザーインテントを特定した後、AI エージェントはこのクエリをサブタスクに分割し、最も効率的な取得戦略を選択する必要があります。この場合、AI エージェントは最も関連性の高いノード (患者の年齢、性別、状態、治療、薬剤など) を特定し、これらのエンティティとその属性と関係についてグラ

フをクエリする必要があります。ReAct エージェントは、LLM の推論 (論理推論) 機能とアクション (外部リソースまたはナレッジベースをクエリまたは操作) を組み合わせるため、非常に役立ちます。

ユーザークエリ「類似する患者に対してどのような治療が行われたか」に回答するために、次の例は ReAct エージェントの仕組みを示しています。

1. エージェントの推論 - ReAct エージェントは、質問には条件 (ダイアベテと高頻度) に関する情報の取得が含まれると推測します。患者の年齢、治療、薬剤、分析する期間を考慮します。
2. エージェントアクション - エージェントは openCypher を使用して、タイプ 2 の「高低」と「高低」に固有の処理についてナレッジグラフをクエリします。また、同様の患者 (性別と年齢が同じ患者など) に対して、投与薬剤、診察日、薬剤の副作用、既知の患者の結果、および相互参照データも取得します。
3. エージェントの観察 – ナレッジグラフから、エージェントは、高低と 2 型の両方の高低の両方を持つ患者に行われた治療に関する最新の 6 か月分の表形式データを取得します。
4. エージェントの推論 – 取得したレコードの結果をランク付けするために、エージェントは、緊急性、薬剤の副作用、既知の患者の結果などの重要な属性を識別します。
5. エージェントアクション – エージェントは、システムプロンプトを介して付与される識別された属性と事前定義されたロジックに基づいてレコードを再ランク付けします。
6. レスポンスの生成 – Amazon Bedrock の LLM は、ReAct エージェントが準備したコンテキストに基づいてレスポンスを生成します。

ヘルスケア向けの生成 AI ソリューションの評価

構築する医療 AI ソリューションを評価することは、実際の医療環境で効果的、信頼性、拡張性を確保するために不可欠です。体系的なアプローチを使用して、ソリューションの各コンポーネントのパフォーマンスを評価します。以下は、ソリューションの評価に使用できる方法論とメトリクスの概要です。

トピック

情報の抽出の評価

インテリジェントな再開パーサーやカスタムエンティティエクストラクタなどの情報抽出ソリューションのパフォーマンスを評価します。テストデータセットを使用して、これらのソリューションのレスポンスのアライメントを測定できます。汎用的な医療人材プロファイルと患者の医療記録をカバーするデータセットがない場合は、LLM の推論機能を使用してカスタムテストデータセットを作成できます。たとえば、モデルなどの大規模なパラメータAnthropic Claudeモデルを使用して、テストデータセットを生成できます。

以下は、情報抽出モデルの評価に使用できる 3 つの主要なメトリクスです。

- 正確性と完全性 – これらのメトリクスは、グラウンドトゥルースデータに存在する正確で完全な情報を出力がキャプチャした範囲を評価します。これには、抽出された情報の正確性と、抽出された情報に関連するすべての詳細の存在の両方を確認することが含まれます。
- 類似度と関連性 – これらのメトリクスは、出力とグラウンドトゥルースデータの間の意味的、構造的、およびコンテキスト的な類似度 (類似度) と、出力がグラウンドトゥルースデータの内容、コンテキスト、インテントとどの程度一致しているか (関連性) を評価します。
- 調整された再現率またはキャプチャ率 – これらの率は、グラウンドトゥルースデータ内の現在の値のうち、モデルによって正しく識別された値の数を経験的に決定します。レートには、モデルが抽出するすべての false 値に対するペナルティを含める必要があります。
- 精度スコア – 精度スコアは、真陽性と比較して、予測に存在する誤検出の数を決定するのに役立ちます。たとえば、精度メトリクスを使用して、抽出されたスキルの習熟度の精度を測定できます。

複数のリトリーバーによる RAG ソリューションの評価

システムが関連情報をどの程度効果的に取得し、その情報を使用して正確でコンテキストに応じた適切なレスポンスを生成するかを評価するには、次のメトリクスを使用できます。

- レスポンスの関連性 – 取得したコンテキストを使用する生成されたレスポンスが元のクエリにどの程度関連しているかを測定します。
- コンテキストの精度 – 取得された結果の合計のうち、クエリに関連する取得されたドキュメントまたはスニペットの割合を評価します。コンテキストの精度が高いほど、取得メカニズムが関連情報の選択に有効であることを示します。
- 忠実度 – 生成されたレスポンスが、取得したコンテキスト内の情報を反映する精度を評価します。つまり、レスポンスがソース情報に当てはまるかどうかを測定します。

LLM を使用したソリューションの評価

LLM-as-a-judge と呼ばれる手法を使用して、生成 AI ソリューションからのテキストレスポンスを評価できます。これには、LLMs を使用してモデル出力のパフォーマンスを評価および評価することが含まれます。この手法では、Amazon Bedrock の機能を使用して、人間の好みやグラウンドトゥルースデータに対する応答品質、一貫性、準拠性、正確性、完全性など、さまざまな属性に関する判断を提供します。包括的な評価にはchain-of-thought (CoT) と数ショットプロンプト手法を使用します。プロンプトは、生成されたレスポンスをスコアリングルーブリックで評価するように LLM に指示し、プロンプト内の数ショットのサンプルは実際の評価プロセスを示しています。プロンプトには、LLM 評価者が従うべきガイドラインも含まれています。たとえば、LLM を使用して生成されたレスポンスを判断する、次の 1 つ以上の評価手法を使用することを検討できます。

- ペア比較 – LLM 評価者に、作成したさまざまな反復バージョンの RAG システムによって生成された医療質問と複数の回答を提供します。LLM 評価者に、レスポンスの品質、一貫性、元の質問への準拠に基づいて最適なレスポンスを決定するよう促します。
- 単一回答評価 – この手法は、患者の結果分類、患者の行動分類、患者の再入院の可能性、リスク分類など、分類の精度を評価する必要があるユースケースに適しています。LLM 評価者を使用して、個別の分類または分類を個別に分析し、グラウンドトゥルースデータに対して提供された推論を評価します。
- リファレンスガイドによる評価 – LLM 評価者に、説明的な回答を必要とする一連の医療質問を提供します。リファレンス回答や理想的な回答など、これらの質問に対するサンプル回答を作成します。LLM エバリュエーターに LLM 生成レスポンスをリファレンス回答または理想的なレスポンスと比較するよう促し、LLM エバリュエーターに生成されたレスポンスの精度、完全性、類似性、

関連性、またはその他の属性を評価するよう促します。この手法は、生成されたレスポンスが明確に定義された標準回答と代表的な回答のどちらと一致するかを評価するのに役立ちます。

リソース

AWS ドキュメント

- Amazon Bedrock ドキュメント
- Amazon Neptune ドキュメント
- 「Amazon OpenSearch Service」
- Amazon Neptune 用の AWS Well-Architected フレームワークの適用 (AWS 規範ガイダンス)
- Amazon OpenSearch Service の運用上のベストプラクティス (OpenSearch Service ドキュメント)
- ヘルスケアとライフサイエンスに Amazon Comprehend Medical と LLMs を使用する (AWS 規範ガイダンス)

AWS ブログ投稿

- Amazon Bedrock で利用可能な新しい Amazon Titan Text Premier モデルを使用して RAG およびエージェントベースの生成 AI アプリケーションを構築する
- Amazon Neptune でデータウェアハウスからナレッジグラフを構築してコマーシャルインテリジェンスを補完する
- ナレッジグラフを使用して Amazon Bedrock と Amazon Neptune で GraphRAG アプリケーションを構築する

その他のリソース

- 検索拡張生成と大規模言語モデルを二ーロジーに統合する: 実践的なアプリケーションを推進する (PubMed Central、国立医学図書館)
- の概要 LangChain (LangChain ドキュメント)

寄稿者

オーサリング

- Nitu Nivedita、マネージングディレクター – 人工知能リード、データ & AI、Accenture
- Manoj Appully、Founder and CTO、Cadiem
- Conor Folan、コンサルタント – Data & AI、Accenture
- Deepak Krishna AR、コンサルタント – Data & AI、Accenture
- Almore Cato、マネージャー – Data & AI、Accenture
- Soonam Kurian、プリンシパルソリューションアーキテクト、 AWS

レビュー

- Sally Lin、データサイエンスシニアマネージャー – Data & AI、Accenture
- Terry Huang、データサイエンスマネージャー – Data & AI、Accenture
- William Lorenz、パートナーソリューションアーキテクト、 AWS

テクニカルライティング

- " AbouHarb、シニアテクニカルライター、 AWS

ドキュメント履歴

以下の表は、本ガイドの重要な変更点について説明したものです。今後の更新に関する通知を受け取る場合は、RSS フィード をサブスクライブできます。

変更	説明	日付
初版発行	—	2025 年 3 月 14 日

AWS 規範ガイダンス用語集

以下は、 AWS 規範ガイダンスが提供する戦略、ガイド、パターンで一般的に使用される用語です。エントリを提案するには、用語集の最後のフィードバックの提供リンクを使用します。

数字

7 Rs

アプリケーションをクラウドに移行するための 7 つの一般的な移行戦略。これらの戦略は、ガートナーが 2011 年に特定した 5 Rs に基づいて構築され、以下で構成されています。

- リファクタリング/アーキテクチャの再設計 — クラウドネイティブ特徴を最大限に活用して、俊敏性、パフォーマンス、スケーラビリティを向上させ、アプリケーションを移動させ、アーキテクチャを変更します。これには、通常、オペレーティングシステムとデータベースの移植が含まれます。例: オンプレミスの Oracle データベースを Amazon Aurora PostgreSQL 互換エディションに移行します。
- リプラットフォーム (リフトアンドリシェイプ) — アプリケーションをクラウドに移行し、クラウド機能を活用するための最適化レベルを導入します。例: オンプレミスの Oracle データベースを の Oracle 用 Amazon Relational Database Service (Amazon RDS) に移行します AWS クラウド。
- 再購入 (ドロップアンドショップ) — 通常、従来のライセンスから SaaS モデルに移行して、別の製品に切り替えます。例: カスタマーリレーションシップ管理 (CRM) システムを Salesforce.com に移行します。
- リホスト (リフトアンドシフト) — クラウド機能を活用するための変更を加えずに、アプリケーションをクラウドに移行します。例: オンプレミスの Oracle データベースを の EC2 インスタンス上の Oracle に移行します AWS クラウド。
- 再配置 (ハイパーバイザーレベルのリフトアンドシフト) – 新しいハードウェアを購入したり、アプリケーションを書き換えたり、既存の運用を変更したりすることなく、インフラストラクチャをクラウドに移行できます。サーバーをオンプレミスプラットフォームから同じプラットフォームのクラウドサービスに移行します。例: Microsoft Hyper-Vアプリケーションを に移行します AWS。
- 保持 (再アクセス) — アプリケーションをお客様のソース環境で保持します。これには、主要なリファクタリングを必要とするアプリケーションや、お客様がその作業を後日まで延期したいアプリケーション、およびそれらを移行するためのビジネス上の正当性がないため、お客様が保持するレガシーアプリケーションなどがあります。

- 使用停止 — お客様のソース環境で不要になったアプリケーションを停止または削除します。

A

ABAC

属性ベースのアクセスコントロールを参照してください。

抽象化されたサービス

「マネージドサービス」を参照してください。

ACID

アトミック性、一貫性、分離性、耐久性を参照してください。

アクティブ - アクティブ移行

(双方向レプリケーションツールまたは二重書き込み操作を使用して) ソースデータベースとターゲットデータベースを同期させ、移行中に両方のデータベースが接続アプリケーションからのトランザクションを処理するデータベース移行方法。この方法では、1 回限りのカットオーバーの必要がなく、管理された小規模なバッチで移行できます。より柔軟ですが、アクティブ/パッシブ移行よりも多くの作業が必要です。

アクティブ - パッシブ移行

ソースデータベースとターゲットデータベースを同期させながら、データがターゲットデータベースにレプリケートされている間、接続しているアプリケーションからのトランザクションをソースデータベースのみで処理するデータベース移行の方法。移行中、ターゲットデータベースはトランザクションを受け付けません。

集計関数

行のグループで動作し、グループの単一の戻り値を計算する SQL 関数。集計関数の例には、SUMおよび が含まれますMAX。

AI

「人工知能」を参照してください。

AIOps

「人工知能オペレーション」を参照してください。

匿名化

データセット内の個人情報を完全に削除するプロセス。匿名化は個人のプライバシー保護に役立ちます。匿名化されたデータは、もはや個人データとは見なされません。

アンチパターン

繰り返し起こる問題に対して頻繁に用いられる解決策で、その解決策が逆効果であったり、効果がなかったり、代替案よりも効果が低かったりするもの。

アプリケーションコントロール

マルウェアからシステムを保護するために、承認されたアプリケーションのみを使用できるようにするセキュリティアプローチ。

アプリケーションポートフォリオ

アプリケーションの構築と維持にかかるコスト、およびそのビジネス価値を含む、組織が使用する各アプリケーションに関する詳細情報の集まり。この情報は、ポートフォリオの検出と分析プロセス の需要要素であり、移行、モダナイズ、最適化するアプリケーションを特定し、優先順位を付けるのに役立ちます。

人工知能 (AI)

コンピューティングテクノロジーを使用し、学習、問題の解決、パターンの認識など、通常は人間に関連づけられる認知機能の実行に特化したコンピュータサイエンスの分野。詳細については、「人工知能 (AI) とは何ですか?」を参照してください。

AI オペレーション (AIOps)

機械学習技術を使用して運用上の問題を解決し、運用上のインシデントと人の介入を減らし、サービス品質を向上させるプロセス。 AWS 移行戦略での AIOps の使用方法については、オペレーション統合ガイド を参照してください。

非対称暗号化

暗号化用のパブリックキーと復号用のプライベートキーから成る 1 組のキーを使用した、暗号化のアルゴリズム。パブリックキーは復号には使用されないため共有しても問題ありませんが、プライベートキーの利用は厳しく制限する必要があります。

原子性、一貫性、分離性、耐久性 (ACID)

エラー、停電、その他の問題が発生した場合でも、データベースのデータ有効性と運用上の信頼性を保証する一連のソフトウェアプロパティ。

属性ベースのアクセス制御 (ABAC)

部署、役職、チーム名など、ユーザーの属性に基づいてアクセス許可をきめ細かく設定する方法。詳細については、 AWS Identity and Access Management (IAM) ドキュメントの「 の ABAC AWS」を参照してください。

信頼できるデータソース

最も信頼性のある情報源とされるデータのプライマリーバージョンを保存する場所。匿名化、編集、仮名化など、データを処理または変更する目的で、信頼できるデータソースから他の場所にデータをコピーすることができます。

アベイラビリティーゾーン

他のアベイラビリティーゾーンの障害から AWS リージョン 隔離され、同じリージョン内の他のアベイラビリティーゾーンへの低コストで低レイテンシーのネットワーク接続を提供する 内の別の場所。

AWS クラウド導入フレームワーク (AWS CAF)

のガイドラインとベストプラクティスのフレームワークは、組織がクラウドへの移行を成功させるための効率的で効果的な計画を立て AWS るのに役立ちます。 AWS CAF は、ビジネス、人材、ガバナンス、プラットフォーム、セキュリティ、運用という 6 つの重点分野にガイダンスを整理します。ビジネス、人材、ガバナンスの観点では、ビジネススキルとプロセスに重点を置き、プラットフォーム、セキュリティ、オペレーションの視点は技術的なスキルとプロセスに焦点を当てています。例えば、人材の観点では、人事 (HR)、人材派遣機能、および人材管理を扱うステークホルダーを対象としています。この観点から、 AWS CAF は、クラウド導入を成功させるための組織の準備に役立つ人材開発、トレーニング、コミュニケーションのガイダンスを提供します。詳細については、AWS CAF ウェブサイト と AWS CAF のホワイトペーパー を参照してください。

AWS ワークロード認定フレームワーク (AWS WQF)

データベース移行ワークロードを評価し、移行戦略を推奨し、作業見積もりを提供するツール。AWS WQF は AWS Schema Conversion Tool （ AWS SCT) に含まれています。データベーススキーマとコードオブジェクト、アプリケーションコード、依存関係、およびパフォーマンス特性を分析し、評価レポートを提供します。

B

不正なボット

個人や組織を混乱させたり、損害を与えたりすることを意図したボット。

BCP

「事業継続計画」を参照してください。

動作グラフ

リソースの動作とインタラクションを経時的に示した、一元的なインタラクティブビュー。Amazon Detective の動作グラフを使用すると、失敗したログオンの試行、不審な API 呼び出し、その他同様のアクションを調べることができます。詳細については、Detective ドキュメントのData in a behavior graphを参照してください。

ビッグエンディアンシステム

最上位バイトを最初に格納するシステム。エンディアン性も参照してください。

二項分類

バイナリ結果 (2 つの可能なクラスのうちの 1 つ) を予測するプロセス。例えば、お客様の機械学習モデルで「この E メールはスパムですか、それともスパムではありませんか」などの問題を予測する必要があるかもしれません。または「この製品は書籍ですか、車ですか」などの問題を予測する必要があるかもしれません。

ブルームフィルター

要素がセットのメンバーであるかどうかをテストするために使用される、確率的でメモリ効率の高いデータ構造。

ブルー/グリーンデプロイ

2 つの異なる同一の環境を作成するデプロイ戦略。現在のアプリケーションバージョンを 1 つの環境 (青) で実行し、新しいアプリケーションバージョンを別の環境 (緑) で実行します。この戦略は、最小限の影響で迅速にロールバックするのに役立ちます。

ボット

インターネット経由で自動タスクを実行し、人間のアクティビティややり取りをシミュレートするソフトウェアアプリケーション。インターネット上の情報のインデックスを作成するウェブクローラーなど、一部のボットは有用または有益です。悪質なボットと呼ばれる他のボットの中には、個人や組織を混乱させたり、損害を与えたりすることを意図したものもあります。

ボットネット

マルウェアに感染し、ボットハーダーまたはボットオペレーターとして知られる、単一の当事者によって制御されているボットのネットワーク。ボットは、ボットとその影響をスケールするための最もよく知られているメカニズムです。

ブランチ

コードリポジトリに含まれる領域。リポジトリに最初に作成するブランチは、メインブランチといいます。既存のブランチから新しいブランチを作成し、その新しいブランチで機能を開発したり、バグを修正したりできます。機能を構築するために作成するブランチは、通常、機能ブランチと呼ばれます。機能をリリースする準備ができたら、機能ブランチをメインブランチに統合します。詳細については、「ブランチの概要」(GitHub ドキュメント) を参照してください。

ブレークグラスアクセス

例外的な状況では、承認されたプロセスを通じて、ユーザーが AWS アカウント 通常アクセス許可を持たない にすばやくアクセスできるようになります。詳細については、 Well-Architected ガイダンスの「ブレークグラス手順の実装」インジケータ AWS を参照してください。

ブラウンフィールド戦略

環境の既存インフラストラクチャ。システムアーキテクチャにブラウンフィールド戦略を導入する場合、現在のシステムとインフラストラクチャの制約に基づいてアーキテクチャを設計します。既存のインフラストラクチャを拡張している場合は、ブラウンフィールド戦略とグリーンフィールド戦略を融合させることもできます。

バッファキャッシュ

アクセス頻度が最も高いデータが保存されるメモリ領域。

ビジネス能力

価値を生み出すためにビジネスが行うこと (営業、カスタマーサービス、マーケティングなど)。マイクロサービスのアーキテクチャと開発の決定は、ビジネス能力によって推進できます。詳細については、ホワイトペーパー AWSでのコンテナ化されたマイクロサービスの実行 の ビジネス機能を中心に組織化 セクションを参照してください。

ビジネス継続性計画 (BCP)

大規模移行など、中断を伴うイベントが運用に与える潜在的な影響に対処し、ビジネスを迅速に再開できるようにする計画。

C

CAF

AWS 「クラウド導入フレームワーク」を参照してください。

Canary デプロイ

エンドユーザーへのバージョンのスローリリースと増分リリース。確信が持てば、新しいバージョンをデプロイし、現在のバージョン全体を置き換えます。

CCoE

「Cloud Center of Excellence」を参照してください。

CDC

「データキャプチャの変更」を参照してください。

変更データキャプチャ (CDC)

データソース (データベーステーブルなど) の変更を追跡し、その変更に関するメタデータを記録するプロセス。CDC は、ターゲットシステムでの変更を監査またはレプリケートして同期を維持するなど、さまざまな目的に使用できます。

カオスエンジニアリング

障害や破壊的なイベントを意図的に導入して、システムの耐障害性をテストします。AWS Fault Injection Service （ AWS FIS ） を使用して、 AWS ワークロードにストレスを与え、その応答を評価する実験を実行できます。

CI/CD

継続的インテグレーションと継続的デリバリーを参照してください。

分類

予測を生成するのに役立つ分類プロセス。分類問題の機械学習モデルは、離散値を予測します。離散値は、常に互いに区別されます。例えば、モデルがイメージ内に車があるかどうかを評価する必要がある場合があります。

クライアント側の暗号化

ターゲットがデータ AWS のサービス を受信する前のローカルでのデータの暗号化。

Cloud Center of Excellence (CCoE)

クラウドのベストプラクティスの作成、リソースの移動、移行のタイムラインの確立、大規模変革を通じて組織をリードするなど、組織全体のクラウド導入の取り組みを推進する学際的なチーム。詳細については、 AWS クラウド エンタープライズ戦略ブログの CCoE 投稿を参照してください。

クラウドコンピューティング

リモートデータストレージと IoT デバイス管理に通常使用されるクラウドテクノロジー。クラウドコンピューティングは、一般的にエッジコンピューティングテクノロジーに接続されています。

クラウド運用モデル

IT 組織において、1 つ以上のクラウド環境を構築、成熟、最適化するために使用される運用モデル。詳細については、「クラウド運用モデルの構築」 を参照してください。

導入のクラウドステージ

組織が に移行するときに通常実行する 4 つのフェーズ AWS クラウド :

- プロジェクト — 概念実証と学習を目的として、クラウド関連のプロジェクトをいくつか実行する
- 基礎固め — お客様のクラウドの導入を拡大するための基礎的な投資 (ランディングゾーンの作成、CCoE の定義、運用モデルの確立など)
- 移行 — 個々のアプリケーションの移行
- 再発明 — 製品とサービスの最適化、クラウドでのイノベーション

これらのステージは、 AWS クラウド エンタープライズ戦略ブログのブログ記事「クラウドファーストへのジャーニー」と「導入のステージ」で Stephen Orban によって定義されました。移行戦略との関連性については、 AWS「移行準備ガイド」を参照してください。

CMDB

「設定管理データベース」を参照してください。

コードリポジトリ

ソースコードやその他の資産 (ドキュメント、サンプル、スクリプトなど) が保存され、バージョン管理プロセスを通じて更新される場所。一般的なクラウドリポジトリには、 GitHubまたは が含まれますBitbucket Cloud。コードの各バージョンはブランチと呼ばれます。マイクロサービスの構造では、各リポジトリは 1 つの機能専用です。1 つの CI/CD パイプラインで複数のリポジトリを使用できます。

コールドキャッシュ

空である、または、かなり空きがある、もしくは、古いデータや無関係なデータが含まれているバッファキャッシュ。データベースインスタンスはメインメモリまたはディスクから読み取る必要があり、バッファキャッシュから読み取るよりも時間がかかるため、パフォーマンスに影響します。

コールドデータ

めったにアクセスされず、通常は過去のデータです。この種類のデータをクエリする場合、通常は低速なクエリでも問題ありません。このデータを低パフォーマンスで安価なストレージ階層またはクラスに移動すると、コストを削減することができます。

コンピュータビジョン (CV)

機械学習を使用してデジタルイメージやビデオなどのビジュアル形式から情報を分析および抽出する AI の分野。例えば、Amazon SageMaker AI は CV 用の画像処理アルゴリズムを提供します。

設定ドリフト

ワークロードの場合、設定は想定状態から変化します。ワークロードが非準拠になる可能性があり、通常は段階的かつ意図的ではありません。

構成管理データベース (CMDB)

データベースとその IT 環境 (ハードウェアとソフトウェアの両方のコンポーネントとその設定を含む) に関する情報を保存、管理するリポジトリ。通常、CMDB のデータは、移行のポートフォリオの検出と分析の段階で使用します。

コンフォーマンスパック

コンプライアンスチェックとセキュリティチェックをカスタマイズするためにアセンブルできる AWS Config ルールと修復アクションのコレクション。YAML テンプレートを使用して、コンフォーマンスパックを AWS アカウント および リージョンの単一のエンティティとしてデプロイすることも、組織全体にデプロイすることもできます。詳細については、 AWS Config ドキュメントの「コンフォーマンスパック」を参照してください。

継続的インテグレーションと継続的デリバリー (CI/CD)

ソフトウェアリリースプロセスのソース、ビルド、テスト、ステージング、本番の各ステージを自動化するプロセス。CI/CD は一般的にパイプラインと呼ばれます。プロセスの自動化、生産性の向上、コード品質の向上、配信の加速化を可能にします。詳細については、「継続的デリバ

リーの利点」を参照してください。CD は継続的デプロイ (Continuous Deployment) の略語でもあります。詳細については「継続的デリバリーと継続的なデプロイ」を参照してください。

CV

「コンピュータビジョン」を参照してください。

D

保管中のデータ

ストレージ内にあるデータなど、常に自社のネットワーク内にあるデータ。

データ分類

ネットワーク内のデータを重要度と機密性に基づいて識別、分類するプロセス。データに適した保護および保持のコントロールを判断する際に役立つため、あらゆるサイバーセキュリティのリスク管理戦略において重要な要素です。データ分類は、 AWS Well-Architected フレームワークのセキュリティの柱のコンポーネントです。詳細については、データ分類を参照してください。

データドリフト

実稼働データと ML モデルのトレーニングに使用されたデータとの間に有意な差異が生じたり、入力データが時間の経過と共に有意に変化したりすることです。データドリフトは、ML モデル予測の全体的な品質、精度、公平性を低下させる可能性があります。

転送中のデータ

ネットワーク内 (ネットワークリソース間など) を活発に移動するデータ。

データメッシュ

一元管理とガバナンスを備えた分散型の分散型データ所有権を提供するアーキテクチャフレームワーク。

データ最小化

厳密に必要なデータのみを収集し、処理するという原則。でデータ最小化を実践 AWS クラウドすることで、プライバシーリスク、コスト、分析のカーボンフットプリントを削減できます。

データ境界

AWS 環境内の一連の予防ガードレール。信頼された ID のみが、期待されるネットワークから信頼されたリソースにアクセスできるようにします。詳細については、「 でのデータ境界の構築 AWS」を参照してください。

データの前処理

raw データをお客様の機械学習モデルで簡単に解析できる形式に変換すること。データの前処理とは、特定の列または行を削除して、欠落している、矛盾している、または重複する値に対処することを意味します。

データ出所

データの生成、送信、保存の方法など、データのライフサイクル全体を通じてデータの出所と履歴を追跡するプロセス。

データ件名

データを収集、処理している個人。

データウェアハウス

分析などのビジネスインテリジェンスをサポートするデータ管理システム。データウェアハウスには通常、大量の履歴データが含まれており、通常はクエリや分析に使用されます。

データベース定義言語 (DDL)

データベース内のテーブルやオブジェクトの構造を作成または変更するためのステートメントまたはコマンド。

データベース操作言語 (DML)

データベース内の情報を変更 (挿入、更新、削除) するためのステートメントまたはコマンド。

DDL

「データベース定義言語」を参照してください。

ディープアンサンブル

予測のために複数の深層学習モデルを組み合わせる。ディープアンサンブルを使用して、より正確な予測を取得したり、予測の不確実性を推定したりできます。

ディープラーニング

人工ニューラルネットワークの複数層を使用して、入力データと対象のターゲット変数の間のマッピングを識別する機械学習サブフィールド。

多層防御

一連のセキュリティメカニズムとコントロールをコンピュータネットワーク全体に層状に重ねて、ネットワークとその内部にあるデータの機密性、整合性、可用性を保護する情報セキュリ

ティの手法。この戦略を採用するときは AWS、 AWS Organizations 構造の異なるレイヤーに複数のコントロールを追加して、リソースの安全性を確保します。たとえば、多層防御アプローチでは、多要素認証、ネットワークセグメンテーション、暗号化を組み合わせることができます。

委任管理者

では AWS Organizations、互換性のあるサービスが AWS メンバーアカウントを登録して組織のアカウントを管理し、そのサービスのアクセス許可を管理できます。このアカウントを、そのサービスの委任管理者と呼びます。詳細、および互換性のあるサービスの一覧は、 AWS Organizations ドキュメントのAWS Organizationsで使用できるサービスを参照してください。

デプロイ

アプリケーション、新機能、コードの修正をターゲットの環境で利用できるようにするプロセス。デプロイでは、コードベースに変更を施した後、アプリケーションの環境でそのコードベースを構築して実行します。

開発環境

「環境」を参照してください。

検出管理

イベントが発生したときに、検出、ログ記録、警告を行うように設計されたセキュリティコントロール。これらのコントロールは副次的な防衛手段であり、実行中の予防的コントロールをすり抜けたセキュリティイベントをユーザーに警告します。詳細については、Implementing security controls on AWSのDetective controlsを参照してください。

開発バリューストリームマッピング (DVSM)

ソフトウェア開発ライフサイクルのスピードと品質に悪影響を及ぼす制約を特定し、優先順位を付けるために使用されるプロセス。DVSM は、もともとリーンマニュファクチャリング・プラクティスのために設計されたバリューストリームマッピング・プロセスを拡張したものです。ソフトウェア開発プロセスを通じて価値を創造し、動かすために必要なステップとチームに焦点を当てています。

デジタルツイン

建物、工場、産業機器、生産ラインなど、現実世界のシステムを仮想的に表現したものです。デジタルツインは、予知保全、リモートモニタリング、生産最適化をサポートします。

ディメンションテーブル

スタースキーマでは、ファクトテーブル内の量的データに関するデータ属性を含む小さなテーブル。ディメンションテーブル属性は通常、テキストフィールドまたはテキストのように動作する

離散数値です。これらの属性は、クエリの制約、フィルタリング、結果セットのラベル付けに一般的に使用されます。

ディザスタ

ワークロードまたはシステムが、導入されている主要な場所でのビジネス目標の達成を妨げるイベント。これらのイベントは、自然災害、技術的障害、または意図しない設定ミスやマルウェア攻撃などの人間の行動の結果である場合があります。

ディザスタリカバリ (DR)

災害によるダウンタイムとデータ損失を最小限に抑えるために使用する戦略とプロセス。詳細については、 AWS Well-Architected フレームワークの「 でのワークロードのディザスタリカバリ AWS: クラウドでのリカバリ」を参照してください。

DML

「データベース操作言語」を参照してください。

ドメイン駆動型設計

各コンポーネントが提供している変化を続けるドメイン、またはコアビジネス目標にコンポーネントを接続して、複雑なソフトウェアシステムを開発するアプローチ。この概念は、エリック・エヴァンスの著書、Domain-Driven Design: Tackling Complexity in the Heart of Software (ドメイン駆動設計:ソフトウェアの中心における複雑さへの取り組み) で紹介されています (ボストン: Addison-Wesley Professional、2003)。strangler fig パターンでドメイン駆動型設計を使用する方法の詳細については、コンテナと Amazon API Gateway を使用して、従来の Microsoft ASP.NET (ASMX) ウェブサービスを段階的にモダナイズ を参照してください。

DR

「ディザスタリカバリ」を参照してください。

ドリフト検出

ベースライン設定からの偏差を追跡します。たとえば、 AWS CloudFormation を使用してシステムリソースのドリフトを検出したり、 を使用して AWS Control Tower 、ガバナンス要件のコンプライアンスに影響を与える可能性のあるランディングゾーンの変更を検出したりできます。

DVSM

「開発値ストリームマッピング」を参照してください。

E

EDA

「探索的データ分析」を参照してください。

EDI

「電子データ交換」を参照してください。

エッジコンピューティング

IoT ネットワークのエッジにあるスマートデバイスの計算能力を高めるテクノロジー。クラウドコンピューティングと比較すると、エッジコンピューティングは通信レイテンシーを短縮し、応答時間を短縮できます。

電子データ交換 (EDI)

組織間のビジネスドキュメントの自動交換。詳細については、「電子データ交換とは」を参照してください。

暗号化

人間が読み取り可能なプレーンテキストデータを暗号文に変換するコンピューティングプロセス。

暗号化キー

暗号化アルゴリズムが生成した、ランダム化されたビットからなる暗号文字列。キーの長さは決まっておらず、各キーは予測できないように、一意になるように設計されています。

エンディアン

コンピュータメモリにバイトが格納される順序。ビッグエンディアンシステムでは、最上位バイトが最初に格納されます。リトルエンディアンシステムでは、最下位バイトが最初に格納されます。

エンドポイント

「サービスエンドポイント」を参照してください。

エンドポイントサービス

仮想プライベートクラウド (VPC) 内でホストして、他のユーザーと共有できるサービス。を使用してエンドポイントサービスを作成し AWS PrivateLink 、他の AWS アカウント または AWS Identity and Access Management (IAM) プリンシパルにアクセス許可を付与できます。これら

のアカウントまたはプリンシパルは、インターフェイス VPC エンドポイントを作成することで、エンドポイントサービスにプライベートに接続できます。詳細については、Amazon Virtual Private Cloud (Amazon VPC) ドキュメントの「エンドポイントサービスを作成する」を参照してください。

エンタープライズリソースプランニング (ERP)

エンタープライズの主要なビジネスプロセス (会計、MES、プロジェクト管理など) を自動化および管理するシステム。

エンベロープ暗号化

暗号化キーを、別の暗号化キーを使用して暗号化するプロセス。詳細については、 AWS Key Management Service （ AWS KMS) ドキュメントの「エンベロープ暗号化」を参照してください。

環境

実行中のアプリケーションのインスタンス。クラウドコンピューティングにおける一般的な環境の種類は以下のとおりです。

- 開発環境 — アプリケーションのメンテナンスを担当するコアチームのみが使用できる、実行中のアプリケーションのインスタンス。開発環境は、上位の環境に昇格させる変更をテストするときに使用します。このタイプの環境は、テスト環境と呼ばれることもあります。
- 下位環境 — 初期ビルドやテストに使用される環境など、アプリケーションのすべての開発環境。
- 本番環境 — エンドユーザーがアクセスできる、実行中のアプリケーションのインスタンス。CI/CD パイプラインでは、本番環境が最後のデプロイ環境になります。
- 上位環境 — コア開発チーム以外のユーザーがアクセスできるすべての環境。これには、本番環境、本番前環境、ユーザー承認テスト環境などが含まれます。

エピック

アジャイル方法論で、お客様の作業の整理と優先順位付けに役立つ機能カテゴリ。エピックでは、要件と実装タスクの概要についてハイレベルな説明を提供します。例えば、 AWS CAF セキュリティエピックには、ID とアクセスの管理、検出コントロール、インフラストラクチャセキュリティ、データ保護、インシデント対応が含まれます。 AWS 移行戦略のエピックの詳細については、プログラム実装ガイド を参照してください。

ERP

「エンタープライズリソース計画」を参照してください。

探索的データ分析 (EDA)

データセットを分析してその主な特性を理解するプロセス。お客様は、データを収集または集計してから、パターンの検出、異常の検出、および前提条件のチェックのための初期調査を実行します。EDA は、統計の概要を計算し、データの可視化を作成することによって実行されます。

F

ファクトテーブル

星スキーマの中央テーブル。事業運営に関する量的データを保存します。通常、ファクトテーブルには、メジャーを含む列とディメンションテーブルへの外部キーを含む列の 2 つのタイプの列が含まれます。

フェイルファスト

開発ライフサイクルを短縮するために頻繁で段階的なテストを使用する哲学。これはアジャイルアプローチの重要な部分です。

障害分離の境界

では AWS クラウド、障害の影響を制限し、ワークロードの耐障害性を高めるのに役立つアベイラビリティーゾーン AWS リージョン、コントロールプレーン、データプレーンなどの境界。詳細については、AWS 「障害分離境界」を参照してください。

機能ブランチ

「ブランチ」を参照してください。

特徴量

お客様が予測に使用する入力データ。例えば、製造コンテキストでは、特徴量は製造ラインから定期的にキャプチャされるイメージの可能性もあります。

特徴量重要度

モデルの予測に対する特徴量の重要性。これは通常、Shapley Additive Deskonations (SHAP) や積分勾配など、さまざまな手法で計算できる数値スコアで表されます。詳細については、「 を使用した機械学習モデルの解釈可能性 AWS」を参照してください。

機能変換

追加のソースによるデータのエンリッチ化、値のスケーリング、単一のデータフィールドからの複数の情報セットの抽出など、機械学習プロセスのデータを最適化すること。これにより、機械

学習モデルはデータの恩恵を受けることができます。例えば、「2021-05-27 00:15:37」の日付を「2021 年」、「5 月」、「木」、「15」に分解すると、学習アルゴリズムがさまざまなデータコンポーネントに関連する微妙に異なるパターンを学習するのに役立ちます。

数ショットプロンプト

同様のタスクの実行を求める前に、タスクと必要な出力を示す少数の例を LLM に提供します。この手法は、プロンプトに埋め込まれた例 (ショット) からモデルが学習するコンテキスト内学習のアプリケーションです。少数ショットプロンプトは、特定のフォーマット、推論、またはドメインの知識を必要とするタスクに効果的です。「ゼロショットプロンプト」も参照してください。

FGAC

「きめ細かなアクセスコントロール」を参照してください。

きめ細かなアクセス制御 (FGAC)

複数の条件を使用してアクセス要求を許可または拒否すること。

フラッシュカット移行

段階的なアプローチを使用する代わりに、変更データキャプチャによる継続的なデータレプリケーションを使用して、可能な限り短時間でデータを移行するデータベース移行方法。目的はダウンタイムを最小限に抑えることです。

FM

「基盤モデル」を参照してください。

基盤モデル (FM)

一般化およびラベル付けされていないデータの大規模なデータセットでトレーニングされている大規模な深層学習ニューラルネットワーク。FMs は、言語の理解、テキストと画像の生成、自然言語の会話など、さまざまな一般的なタスクを実行できます。詳細については、「基盤モデルとは」を参照してください。

G

生成 AI

大量のデータでトレーニングされ、シンプルなテキストプロンプトを使用してイメージ、動画、テキスト、オーディオなどの新しいコンテンツやアーティファクトを作成できる AI モデルのサブセット。詳細については、「生成 AI とは」を参照してください。

ジオブロッキング

地理的制限を参照してください。

地理的制限 (ジオブロッキング)

特定の国のユーザーがコンテンツ配信にアクセスできないようにするための、Amazon CloudFront のオプション。アクセスを許可する国と禁止する国は、許可リストまたは禁止リストを使って指定します。詳細については、CloudFront ドキュメントのコンテンツの地理的ディストリビューションの制限を参照してください。

Gitflow ワークフロー

下位環境と上位環境が、ソースコードリポジトリでそれぞれ異なるブランチを使用する方法。Gitflow ワークフローはレガシーと見なされ、トランクベースのワークフローはモダンで推奨されるアプローチです。

ゴールデンイメージ

そのシステムまたはソフトウェアの新しいインスタンスをデプロイするためのテンプレートとして使用されるシステムまたはソフトウェアのスナップショット。例えば、製造では、ゴールデンイメージを使用して複数のデバイスにソフトウェアをプロビジョニングし、デバイス製造オペレーションの速度、スケーラビリティ、生産性を向上させることができます。

グリーンフィールド戦略

新しい環境に既存のインフラストラクチャが存在しないこと。システムアーキテクチャにグリーンフィールド戦略を導入する場合、既存のインフラストラクチャ (別名ブラウンフィールド) との互換性の制約を受けることなく、あらゆる新しいテクノロジーを選択できます。既存のインフラストラクチャを拡張している場合は、ブラウンフィールド戦略とグリーンフィールド戦略を融合させることもできます。

ガードレール

組織単位 (OU) 全般のリソース、ポリシー、コンプライアンスを管理するのに役立つ概略的なルール。予防ガードレールは、コンプライアンス基準に一致するようにポリシーを実施します。これらは、サービスコントロールポリシーと IAM アクセス許可の境界を使用して実装されます。検出ガードレールは、ポリシー違反やコンプライアンス上の問題を検出し、修復のためのアラートを発信します。これらは AWS Config、、Amazon GuardDuty AWS Security Hub、、 AWS Trusted Advisor Amazon Inspector、およびカスタム AWS Lambda チェックを使用して実装されます。

H

HA

「高可用性」を参照してください。

異種混在データベースの移行

別のデータベースエンジンを使用するターゲットデータベースへお客様の出典データベースの移行 (例えば、Oracle から Amazon Aurora)。異種間移行は通常、アーキテクチャの再設計作業の一部であり、スキーマの変換は複雑なタスクになる可能性があります。AWS は、スキーマの変換に役立つ AWS SCTを提供します。

ハイアベイラビリティ (HA)

課題や災害が発生した場合に、介入なしにワークロードを継続的に運用できること。HA システムは、自動的にフェイルオーバーし、一貫して高品質のパフォーマンスを提供し、パフォーマンスへの影響を最小限に抑えながらさまざまな負荷や障害を処理するように設計されています。

ヒストリアンのモダナイゼーション

製造業のニーズによりよく応えるために、オペレーションテクノロジー (OT) システムをモダナイズし、アップグレードするためのアプローチ。ヒストリアンは、工場内のさまざまなソースからデータを収集して保存するために使用されるデータベースの一種です。

ホールドアウトデータ

機械学習モデルのトレーニングに使用されるデータセットから保留される、ラベル付きの履歴データの一部。モデル予測をホールドアウトデータと比較することで、ホールドアウトデータを使用してモデルのパフォーマンスを評価できます。

同種データベースの移行

お客様の出典データベースを、同じデータベースエンジンを共有するターゲットデータベース (Microsoft SQL Server から Amazon RDS for SQL Server など) に移行する。同種間移行は、通常、リホストまたはリプラットフォーム化の作業の一部です。ネイティブデータベースユーティリティを使用して、スキーマを移行できます。

ホットデータ

リアルタイムデータや最近の翻訳データなど、頻繁にアクセスされるデータ。通常、このデータには高速なクエリ応答を提供する高性能なストレージ階層またはクラスが必要です。

ホットフィックス

本番環境の重大な問題を修正するために緊急で配布されるプログラム。緊急性が高いため、通常の DevOps のリリースワークフローからは外れた形で実施されます。

ハイパーケア期間

カットオーバー直後、移行したアプリケーションを移行チームがクラウドで管理、監視して問題に対処する期間。通常、この期間は 1～4 日です。ハイパーケア期間が終了すると、アプリケーションに対する責任は一般的に移行チームからクラウドオペレーションチームに移ります。

I

IaC

「Infrastructure as Code」を参照してください。

ID ベースのポリシー

AWS クラウド 環境内のアクセス許可を定義する 1 つ以上の IAM プリンシパルにアタッチされたポリシー。

アイドル状態のアプリケーション

90 日間の平均的な CPU およびメモリ使用率が 5～20% のアプリケーション。移行プロジェクトでは、これらのアプリケーションを廃止するか、オンプレミスに保持するのが一般的です。

IIoT

「産業用モノのインターネット」を参照してください。

イミュータブルインフラストラクチャ

既存のインフラストラクチャを更新、パッチ適用、または変更する代わりに、本番環境のワークロード用に新しいインフラストラクチャをデプロイするモデル。イミュータブルインフラストラクチャは、本質的にミュータブルインフラストラクチャよりも一貫性、信頼性、予測性が高くなります。詳細については、 AWS 「 Well-Architected フレームワーク」の「イミュータブルインフラストラクチャを使用したデプロイ」のベストプラクティスを参照してください。

インバウンド (受信) VPC

AWS マルチアカウントアーキテクチャでは、アプリケーションの外部からネットワーク接続を受け入れ、検査し、ルーティングする VPC。AWS Security Reference Architecture では、アプリ

ケーションとより広範なインターネット間の双方向のインターフェイスを保護するために、インバウンド、アウトバウンド、インスペクションの各 VPC を使用してネットワークアカウントを設定することを推奨しています。

増分移行

アプリケーションを 1 回ですべてカットオーバーするのではなく、小さい要素に分けて移行するカットオーバー戦略。例えば、最初は少数のマイクロサービスまたはユーザーのみを新しいシステムに移行する場合があります。すべてが正常に機能することを確認できたら、残りのマイクロサービスやユーザーを段階的に移行し、レガシーシステムを廃止できるようにします。この戦略により、大規模な移行に伴うリスクが軽減されます。

インダストリー 4.0

2016 年に Klaus Schwab によって導入された用語で、接続、リアルタイムデータ、オートメーション、分析、AI/ML の進歩によるビジネスプロセスのモダナイゼーションを指します。

インフラストラクチャ

アプリケーションの環境に含まれるすべてのリソースとアセット。

Infrastructure as Code (IaC)

アプリケーションのインフラストラクチャを一連の設定ファイルを使用してプロビジョニングし、管理するプロセス。IaC は、新しい環境を再現可能で信頼性が高く、一貫性のあるものにするため、インフラストラクチャを一元的に管理し、リソースを標準化し、スケールを迅速に行えるように設計されています。

産業分野における IoT (IIoT)

製造、エネルギー、自動車、ヘルスケア、ライフサイエンス、農業などの産業部門におけるインターネットに接続されたセンサーやデバイスの使用。詳細については、「Building an industrial Internet of Things (IIoT) digital transformation strategy」を参照してください。

インスペクション VPC

AWS マルチアカウントアーキテクチャでは、VPC (同一または異なる AWS リージョン) 、インターネット、オンプレミスネットワーク間のネットワークトラフィックの検査を管理する一元化された VPCs。AWS Security Reference Architecture では、アプリケーションとより広範なインターネット間の双方向のインターフェイスを保護するために、インバウンド、アウトバウンド、インスペクションの各 VPC を使用してネットワークアカウントを設定することを推奨しています。

IoT

インターネットまたはローカル通信ネットワークを介して他のデバイスやシステムと通信する、センサーまたはプロセッサが組み込まれた接続済み物理オブジェクトのネットワーク。詳細については、「IoT とは」を参照してください。

解釈可能性

機械学習モデルの特性で、モデルの予測がその入力にどのように依存するかを人間が理解できる度合いを表します。詳細については、「 を使用した機械学習モデルの解釈可能性 AWS」を参照してください。

IoT

「モノのインターネット」を参照してください。

IT 情報ライブラリ (ITIL)

IT サービスを提供し、これらのサービスをビジネス要件に合わせるための一連のベストプラクティス。ITIL は ITSM の基盤を提供します。

IT サービス管理 (ITSM)

組織の IT サービスの設計、実装、管理、およびサポートに関連する活動。クラウドオペレーションと ITSM ツールの統合については、オペレーション統合ガイド を参照してください。

ITIL

「IT 情報ライブラリ」を参照してください。

ITSM

「IT サービス管理」を参照してください。

L

ラベルベースアクセス制御 (LBAC)

強制アクセス制御 (MAC) の実装で、ユーザーとデータ自体にそれぞれセキュリティラベル値が明示的に割り当てられます。ユーザーセキュリティラベルとデータセキュリティラベルが交差する部分によって、ユーザーに表示される行と列が決まります。

ランディングゾーン

ランディングゾーンは、スケーラブルで安全な、適切に設計されたマルチアカウント AWS 環境です。これは、組織がセキュリティおよびインフラストラクチャ環境に自信を持ってワークロー

ドとアプリケーションを迅速に起動してデプロイできる出発点です。ランディングゾーンの詳細については、安全でスケーラブルなマルチアカウント AWS 環境のセットアップ を参照してください。

大規模言語モデル (LLM)

大量のデータに対して事前トレーニングされた深層学習 AI モデル。LLM は、質問への回答、ドキュメントの要約、テキストの他の言語への翻訳、文の完了など、複数のタスクを実行できます。詳細については、LLMs」を参照してください。

大規模な移行

300 台以上のサーバの移行。

LBAC

「ラベルベースのアクセスコントロール」を参照してください。

最小特権

タスクの実行には必要最低限の権限を付与するという、セキュリティのベストプラクティス。詳細については、IAM ドキュメントの最小特権アクセス許可を適用するを参照してください。

リフトアンドシフト

「7 Rs」を参照してください。

リトルエンディアンシステム

最下位バイトを最初に格納するシステム。エンディアン性も参照してください。

LLM

「大規模言語モデル」を参照してください。

下位環境

「環境」を参照してください。

M

機械学習 (ML)

パターン認識と学習にアルゴリズムと手法を使用する人工知能の一種。ML は、モノのインターネット (IoT) データなどの記録されたデータを分析して学習し、パターンに基づく統計モデルを生成します。詳細については、「機械学習」を参照してください。

メインブランチ

「ブランチ」を参照してください。

マルウェア

コンピュータのセキュリティまたはプライバシーを侵害するように設計されたソフトウェア。マルウェアは、コンピュータシステムの中断、機密情報の漏洩、不正アクセスにつながる可能性があります。マルウェアの例としては、ウイルス、ワーム、ランサムウェア、トロイの木馬、スパイウェア、キーロガーなどがあります。

マネージドサービス

AWS のサービス はインフラストラクチャレイヤー、オペレーティングシステム、プラットフォーム AWS を運用し、エンドポイントにアクセスしてデータを保存および取得します。Amazon Simple Storage Service (Amazon S3) と Amazon DynamoDB は、マネージドサービスの例です。これらは抽象化されたサービスとも呼ばれます。

製造実行システム (MES)

生産プロセスを追跡、モニタリング、文書化、制御するためのソフトウェアシステムで、原材料を工場の完成製品に変換します。

MAP

「移行促進プログラム」を参照してください。

メカニズム

ツールを作成し、ツールの導入を推進し、調整を行うために結果を検査する完全なプロセス。メカニズムは、動作中にそれ自体を強化して改善するサイクルです。詳細については、 AWS 「 Well-Architected フレームワーク」の「メカニズムの構築」を参照してください。

メンバーアカウント

組織の一部である管理アカウント AWS アカウント 以外のすべて AWS Organizations。アカウントが組織のメンバーになることができるのは、一度に 1 つのみです。

MES

「製造実行システム」を参照してください。

メッセージキューイングテレメトリトランスポート (MQTT)

リソースに制約のある IoT デバイス用の、パブリッシュ/サブスクライブパターンに基づく軽量 machine-to-machine (M2M) 通信プロトコル。

マイクロサービス

明確に定義された API を介して通信し、通常は小規模な自己完結型のチームが所有する、小規模で独立したサービスです。例えば、保険システムには、販売やマーケティングなどのビジネス機能、または購買、請求、分析などのサブドメインにマッピングするマイクロサービスが含まれる場合があります。マイクロサービスの利点には、俊敏性、柔軟なスケーリング、容易なデプロイ、再利用可能なコード、回復力などがあります。詳細については、AWS「サーバーレスサービスを使用したマイクロサービスの統合」を参照してください。

マイクロサービスアーキテクチャ

各アプリケーションプロセスをマイクロサービスとして実行する独立したコンポーネントを使用してアプリケーションを構築するアプローチ。これらのマイクロサービスは、軽量 API を使用して、明確に定義されたインターフェイスを介して通信します。このアーキテクチャの各マイクロサービスは、アプリケーションの特定の機能に対する需要を満たすように更新、デプロイ、およびスケーリングできます。詳細については、「 でのマイクロサービスの実装 AWS」を参照してください。

Migration Acceleration Program (MAP)

組織がクラウドに移行するための強力な運用基盤を構築し、移行の初期コストを相殺するのに役立つコンサルティングサポート、トレーニング、サービスを提供する AWS プログラム。MAP には、組織的な方法でレガシー移行を実行するための移行方法論と、一般的な移行シナリオを自動化および高速化する一連のツールが含まれています。

大規模な移行

アプリケーションポートフォリオの大部分を次々にクラウドに移行し、各ウェーブでより多くのアプリケーションを高速に移動させるプロセス。この段階では、以前の段階から学んだベストプラクティスと教訓を使用して、移行ファクトリー チーム、ツール、プロセスのうち、オートメーションとアジャイルデリバリーによってワークロードの移行を合理化します。これは、AWS 移行戦略 の第 3 段階です。

移行ファクトリー

自動化された俊敏性のあるアプローチにより、ワークロードの移行を合理化する部門横断的なチーム。移行ファクトリーチームには、通常、運用、ビジネスアナリストおよび所有者、移行エンジニア、デベロッパー、およびスプリントで作業する DevOps プロフェッショナルが含まれます。エンタープライズアプリケーションポートフォリオの 20～50% は、ファクトリーのアプローチによって最適化できる反復パターンで構成されています。詳細については、このコンテンツセットの移行ファクトリーに関する解説とCloud Migration Factory ガイドを参照してください。

移行メタデータ

移行を完了するために必要なアプリケーションおよびサーバーに関する情報。移行パターンごとに、異なる一連の移行メタデータが必要です。移行メタデータの例としては、ターゲットサブネット、セキュリティグループ、 AWS アカウントなどがあります。

移行パターン

移行戦略、移行先、および使用する移行アプリケーションまたはサービスを詳述する、反復可能な移行タスク。例: AWS Application Migration Service を使用して Amazon EC2 への移行をリホストします。

Migration Portfolio Assessment (MPA)

に移行するためのビジネスケースを検証するための情報を提供するオンラインツール AWS クラウド。MPA は、詳細なポートフォリオ評価 (サーバーの適切なサイジング、価格設定、TCO 比較、移行コスト分析) および移行プラン (アプリケーションデータの分析とデータ収集、アプリケーションのグループ化、移行の優先順位付け、およびウェーブプランニング) を提供します。MPA ツール (ログインが必要) は、すべての AWS コンサルタントと APN パートナーコンサルタントが無料で利用できます。

移行準備状況評価 (MRA)

AWS CAF を使用して、組織のクラウド準備状況に関するインサイトを取得し、長所と短所を特定し、特定されたギャップを埋めるためのアクションプランを構築するプロセス。詳細については、移行準備状況ガイド を参照してください。MRA は、AWS 移行戦略の第一段階です。

移行戦略

ワークロードを に移行するために使用するアプローチ AWS クラウド。詳細については、この用語集の「7 Rs エントリ」と「組織を動員して大規模な移行を加速する」を参照してください。

ML

???「機械学習」を参照してください。

モダナイゼーション

古い (レガシーまたはモノリシック) アプリケーションとそのインフラストラクチャをクラウド内の俊敏で弾力性のある高可用性システムに変換して、コストを削減し、効率を高め、イノベーションを活用します。詳細については、「」の「アプリケーションをモダナイズするための戦略 AWS クラウド」を参照してください。

モダナイゼーション準備状況評価

組織のアプリケーションのモダナイゼーションの準備状況を判断し、利点、リスク、依存関係を特定し、組織がこれらのアプリケーションの将来の状態をどの程度適切にサポートできるかを決定するのに役立つ評価。評価の結果として、ターゲットアーキテクチャのブループリント、モダナイゼーションプロセスの開発段階とマイルストーンを詳述したロードマップ、特定されたギャップに対処するためのアクションプランが得られます。詳細については、『』の「アプリケーションのモダナイゼーション準備状況の評価 AWS クラウド」を参照してください。

モノリシックアプリケーション (モノリス)

緊密に結合されたプロセスを持つ単一のサービスとして実行されるアプリケーション。モノリシックアプリケーションにはいくつかの欠点があります。1 つのアプリケーション機能エクスペリエンスの需要が急増する場合は、アーキテクチャ全体をスケーリングする必要があります。モノリシックアプリケーションの特徴を追加または改善することは、コードベースが大きくなると複雑になります。これらの問題に対処するには、マイクロサービスアーキテクチャを使用できます。詳細については、モノリスをマイクロサービスに分解する を参照してください。

MPA

「移行ポートフォリオ評価」を参照してください。

MQTT

「Message Queuing Telemetry Transport」を参照してください。

多クラス分類

複数のクラスの予測を生成するプロセス (2 つ以上の結果の 1 つを予測します)。例えば、機械学習モデルが、「この製品は書籍、自動車、電話のいずれですか?」 または、「このお客様にとって最も関心のある商品のカテゴリはどれですか?」と聞くかもしれません。

ミュータブルインフラストラクチャ

本番ワークロードの既存のインフラストラクチャを更新および変更するモデル。Well-Architected AWS フレームワークでは、一貫性、信頼性、予測可能性を向上させるために、イミュータブルインフラストラクチャの使用をベストプラクティスとして推奨しています。

O

OAC

「オリジンアクセスコントロール」を参照してください。

OAI

「オリジンアクセスアイデンティティ」を参照してください。

OCM

「組織変更管理」を参照してください。

オフライン移行

移行プロセス中にソースワークロードを停止させる移行方法。この方法はダウンタイムが長くなるため、通常は重要ではない小規模なワークロードに使用されます。

OI

「 オペレーションの統合」を参照してください。

OLA

「運用レベルの契約」を参照してください。

オンライン移行

ソースワークロードをオフラインにせずにターゲットシステムにコピーする移行方法。ワークロードに接続されているアプリケーションは、移行中も動作し続けることができます。この方法はダウンタイムがゼロから最小限で済むため、通常は重要な本番稼働環境のワークロードに使用されます。

OPC-UA

「Open Process Communications - Unified Architecture」を参照してください。

オープンプロセス通信 - 統合アーキテクチャ (OPC-UA)

産業用オートメーション用のmachine-to-machine (M2M) 通信プロトコル。OPC-UA は、データの暗号化、認証、認可スキームを備えた相互運用性標準を提供します。

オペレーショナルレベルアグリーメント (OLA)

サービスレベルアグリーメント (SLA) をサポートするために、どの機能的 IT グループが互いに提供することを約束するかを明確にする契約。

運用準備状況レビュー (ORR)

インシデントや潜在的な障害の理解、評価、防止、または範囲の縮小に役立つ質問とそれに関連するベストプラクティスのチェックリスト。詳細については、 AWS Well-Architected フレームワークの「Operational Readiness Reviews (ORR)」を参照してください。

運用テクノロジー (OT)

産業オペレーション、機器、インフラストラクチャを制御するために物理環境と連携するハードウェアおよびソフトウェアシステム。製造では、OT と情報技術 (IT) システムの統合が、Industry 4.0 変換の主な焦点です。

オペレーション統合 (OI)

クラウドでオペレーションをモダナイズするプロセスには、準備計画、オートメーション、統合が含まれます。詳細については、オペレーション統合ガイド を参照してください。

組織の証跡

組織 AWS アカウント 内のすべての のすべてのイベント AWS CloudTrail をログに記録する、 によって作成された証跡 AWS Organizations。証跡は、組織に含まれている各 AWS アカウント に作成され、各アカウントのアクティビティを追跡します。詳細については、CloudTrail ドキュメントの組織の証跡の作成を参照してください。

組織変更管理 (OCM)

人材、文化、リーダーシップの観点から、主要な破壊的なビジネス変革を管理するためのフレームワーク。OCM は、変化の導入を加速し、移行問題に対処し、文化や組織の変化を推進することで、組織が新しいシステムと戦略の準備と移行するのを支援します。 AWS 移行戦略では、クラウド導入プロジェクトに必要な変化のスピードのため、このフレームワークは人材アクセラレーションと呼ばれます。詳細については、OCM ガイド を参照してください。

オリジンアクセスコントロール (OAC)

Amazon Simple Storage Service (Amazon S3) コンテンツを保護するための、CloudFront のアクセス制限の強化オプション。OAC は AWS リージョン、すべての S3 バケット、 AWS KMS (SSE-KMS) によるサーバー側の暗号化、S3 バケットへの動的 PUT および DELETEリクエストをサポートします。

オリジンアクセスアイデンティティ (OAI)

CloudFront の、Amazon S3 コンテンツを保護するためのアクセス制限オプション。OAI を使用すると、CloudFront が、Amazon S3 に認証可能なプリンシパルを作成します。認証されたプリンシパルは、S3 バケット内のコンテンツに、特定の CloudFront ディストリビューションを介してのみアクセスできます。OACも併せて参照してください。OAC では、より詳細な、強化されたアクセスコントロールが可能です。

ORR

「運用準備状況レビュー」を参照してください。

OT

「運用テクノロジー」を参照してください。

アウトバウンド (送信) VPC

AWS マルチアカウントアーキテクチャでは、アプリケーション内から開始されたネットワーク接続を処理する VPC。AWS Security Reference Architecture では、アプリケーションとより広範なインターネット間の双方向のインターフェイスを保護するために、インバウンド、アウトバウンド、インスペクションの各 VPC を使用してネットワークアカウントを設定することを推奨しています。

P

アクセス許可の境界

ユーザーまたはロールが使用できるアクセス許可の上限を設定する、IAM プリンシパルにアタッチされる IAM 管理ポリシー。詳細については、IAM ドキュメントのアクセス許可の境界を参照してください。

個人を特定できる情報 (PII)

直接閲覧した場合、または他の関連データと組み合わせた場合に、個人の身元を合理的に推測するために使用できる情報。PII の例には、氏名、住所、連絡先情報などがあります。

PII

個人を特定できる情報を参照してください。

プレイブック

クラウドでのコアオペレーション機能の提供など、移行に関連する作業を取り込む、事前定義された一連のステップ。プレイブックは、スクリプト、自動ランブック、またはお客様のモダナイズされた環境を運用するために必要なプロセスや手順の要約などの形式をとることができます。

PLC

「プログラム可能なロジックコントローラー」を参照してください。

PLM

「製品ライフサイクル管理」を参照してください。

ポリシー

アクセス許可を定義 (アイデンティティベースのポリシーを参照)、アクセス条件を指定 (リソースベースのポリシーを参照)、または の組織内のすべてのアカウントに対する最大アクセス許可を定義 AWS Organizations （ サービスコントロールポリシーを参照) できるオブジェクト。

多言語の永続性

データアクセスパターンやその他の要件に基づいて、マイクロサービスのデータストレージテクノロジーを個別に選択します。マイクロサービスが同じデータストレージテクノロジーを使用している場合、実装上の問題が発生したり、パフォーマンスが低下する可能性があります。マイクロサービスは、要件に最も適合したデータストアを使用すると、より簡単に実装でき、パフォーマンスとスケーラビリティが向上します。詳細については、マイクロサービスでのデータ永続性の有効化 を参照してください。

ポートフォリオ評価

移行を計画するために、アプリケーションポートフォリオの検出、分析、優先順位付けを行うプロセス。詳細については、「移行準備状況ガイド」を参照してください。

述語

`true` または を返すクエリ条件。一般的に`false`は `WHERE`句にあります。

述語プッシュダウン

転送前にクエリ内のデータをフィルタリングするデータベースクエリ最適化手法。これにより、リレーショナルデータベースから取得して処理する必要があるデータの量が減少し、クエリのパフォーマンスが向上します。

予防的コントロール

イベントの発生を防ぐように設計されたセキュリティコントロール。このコントロールは、ネットワークへの不正アクセスや好ましくない変更を防ぐ最前線の防御です。詳細については、Implementing security controls on AWSのPreventative controlsを参照してください。

プリンシパル

アクション を実行し AWS 、リソースにアクセスできる のエンティティ。このエンティティは通常、、IAM AWS アカウントロール、または ユーザーのルートユーザーです。詳細については、IAM ドキュメントのロールに関する用語と概念内にあるプリンシパルを参照してください。

プライバシーバイデザイン

開発プロセス全体を通じてプライバシーを考慮するシステムエンジニアリングアプローチ。

プライベートホストゾーン

1 つ以上の VPC 内のドメインとそのサブドメインへの DNS クエリに対し、Amazon Route 53 がどのように応答するかに関する情報を保持するコンテナ。詳細については、Route 53 ドキュメントの「プライベートホストゾーンの使用」を参照してください。

プロアクティブコントロール

非準拠リソースのデプロイを防ぐように設計されたセキュリティコントロール。これらのコントロールは、プロビジョニング前にリソースをスキャンします。リソースがコントロールに準拠していない場合、プロビジョニングされません。詳細については、 AWS Control Tower ドキュメントの「 コントロールリファレンスガイド」および「 セキュリティコントロールの実装」の「プロアクティブコントロール」を参照してください。 AWS

製品ライフサイクル管理 (PLM)

設計、開発、発売から成長と成熟まで、製品のデータとプロセスのライフサイクル全体にわたる管理。

本番環境

「環境」を参照してください。

プログラム可能なロジックコントローラー (PLC)

製造では、マシンをモニタリングし、製造プロセスを自動化する、信頼性の高い適応可能なコンピュータです。

プロンプトの連鎖

1 つの LLM プロンプトの出力を次のプロンプトの入力として使用して、より良いレスポンスを生成します。この手法は、複雑なタスクをサブタスクに分割したり、事前レスポンスを繰り返し改善または拡張したりするために使用されます。これにより、モデルのレスポンスの精度と関連性が向上し、より詳細でパーソナライズされた結果が得られます。

仮名化

データセット内の個人識別子をプレースホルダー値に置き換えるプロセス。仮名化は個人のプライバシー保護に役立ちます。仮名化されたデータは、依然として個人データとみなされます。

パブリッシュ/サブスクライブ (pub/sub)

マイクロサービス間の非同期通信を可能にするパターン。スケーラビリティと応答性を向上させます。たとえば、マイクロサービスベースの MES では、マイクロサービスは他のマイクロサー

ビスがサブスクライブできるチャネルにイベントメッセージを発行できます。システムは、公開サービスを変更せずに新しいマイクロサービスを追加できます。

Q

クエリプラン

SQL リレーショナルデータベースシステムのデータにアクセスするために使用される手順などの一連のステップ。

クエリプランのリグレッション

データベースサービスのオプティマイザーが、データベース環境に特定の変更が加えられる前に選択されたプランよりも最適性の低いプランを選択すること。これは、統計、制限事項、環境設定、クエリパラメータのバインディングの変更、およびデータベースエンジンの更新などが原因である可能性があります。

R

RACI マトリックス

責任、説明責任、相談、通知 (RACI) を参照してください。

RAG

「取得拡張生成」を参照してください。

ランサムウェア

決済が完了するまでコンピュータシステムまたはデータへのアクセスをブロックするように設計された、悪意のあるソフトウェア。

RASCI マトリックス

責任、説明責任、相談、情報 (RACI) を参照してください。

RCAC

「行と列のアクセスコントロール」を参照してください。

リードレプリカ

読み取り専用に使用されるデータベースのコピー。クエリをリードレプリカにルーティングして、プライマリデータベースへの負荷を軽減できます。

再設計

「7 Rs」を参照してください。

目標復旧時点 (RPO)

最後のデータリカバリポイントからの最大許容時間です。これにより、最後の回復時点からサービスが中断されるまでの間に許容できるデータ損失の程度が決まります。

目標復旧時間 (RTO)

サービスの中断から復旧までの最大許容遅延時間。

リファクタリング

「7 Rs」を参照してください。

リージョン

地理的エリア内の AWS リソースのコレクション。各 AWS リージョン は、耐障害性、安定性、耐障害性を提供するために、他の とは独立しています。詳細については、AWS リージョン「アカウントで使用できる を指定する」を参照してください。

回帰

数値を予測する機械学習手法。例えば、「この家はどれくらいの値段で売れるでしょうか?」という問題を解決するために、機械学習モデルは、線形回帰モデルを使用して、この家に関する既知の事実 (平方フィートなど) に基づいて家の販売価格を予測できます。

リホスト

「7 Rs」を参照してください。

リリース

デプロイプロセスで、変更を本番環境に昇格させること。

再配置

「7 Rs」を参照してください。

プラットフォーム変更

「7 Rs」を参照してください。

再購入

「7 Rs」を参照してください。

回復性

中断に抵抗または回復するアプリケーションの機能。高可用性とディザスタリカバリは、 で回復性を計画する際の一般的な考慮事項です AWS クラウド。詳細については、AWS クラウド「レジリエンス」を参照してください。

リソースベースのポリシー

Amazon S3 バケット、エンドポイント、暗号化キーなどのリソースにアタッチされたポリシー。このタイプのポリシーは、アクセスが許可されているプリンシパル、サポートされているアクション、その他の満たすべき条件を指定します。

実行責任者、説明責任者、協業先、報告先 (RACI) に基づくマトリックス

移行活動とクラウド運用に関わるすべての関係者の役割と責任を定義したマトリックス。マトリックスの名前は、マトリックスで定義されている責任の種類、すなわち責任 (R) 、説明責任 (A) 、協議 (C) 、情報提供 (I) に由来します。サポート (S) タイプはオプションです。サポートを含めると、そのマトリックスは RASCI マトリックスと呼ばれ、サポートを除外すると RACI マトリックスと呼ばれます。

レスポンシブコントロール

有害事象やセキュリティベースラインからの逸脱について、修復を促すように設計されたセキュリティコントロール。詳細については、Implementing security controls on AWSのResponsive controlsを参照してください。

保持

「7 Rs」を参照してください。

廃止

「7 Rs」を参照してください。

取得拡張生成 (RAG)

LLM がレスポンスを生成する前にトレーニングデータソースの外部にある信頼できるデータソースを参照する生成 AI テクノロジー。たとえば、RAG モデルは、組織のナレッジベースまたはカスタムデータのセマンティック検索を実行する場合があります。詳細については、「RAG とは」を参照してください。

ローテーション

攻撃者が認証情報にアクセスすることをより困難にするために、シークレットを定期的に更新するプロセス。

行と列のアクセス制御 (RCAC)

アクセスルールが定義された、基本的で柔軟な SQL 表現の使用。RCAC は行権限と列マスクで構成されています。

RPO

「目標復旧時点」を参照してください。

RTO

目標復旧時間を参照してください。

ランブック

特定のタスクを実行するために必要な手動または自動化された一連の手順。これらは通常、エラー率の高い反復操作や手順を合理化するために構築されています。

S

SAML 2.0

多くの ID プロバイダー (IdP) が使用しているオープンスタンダード。この機能を使用すると、フェデレーティッドシングルサインオン (SSO) が有効になるため、ユーザーは組織内のすべてのユーザーを IAM で作成しなくても、 AWS Management Console にログインしたり AWS 、 API オペレーションを呼び出すことができます。SAML 2.0 ベースのフェデレーションの詳細については、IAM ドキュメントのSAML 2.0 ベースのフェデレーションについてを参照してください。

SCADA

「監視コントロールとデータ取得」を参照してください。

SCP

「サービスコントロールポリシー」を参照してください。

シークレット

暗号化された形式で保存する AWS Secrets Managerパスワードやユーザー認証情報などの機密情報または制限付き情報。シークレット値とそのメタデータで構成されます。シークレット値は、バイナリ、単一の文字列、または複数の文字列にすることができます。詳細については、Secrets Manager ドキュメントの「Secrets Manager シークレットの内容」を参照してください。

設計によるセキュリティ

開発プロセス全体でセキュリティを考慮するシステムエンジニアリングアプローチ。

セキュリティコントロール

脅威アクターによるセキュリティ脆弱性の悪用を防止、検出、軽減するための、技術上または管理上のガードレール。セキュリティコントロールには、予防的、検出的、応答的、プロアクティブの 4 つの主なタイプがあります。

セキュリティ強化

アタックサーフェスを狭めて攻撃への耐性を高めるプロセス。このプロセスには、不要になったリソースの削除、最小特権を付与するセキュリティのベストプラクティスの実装、設定ファイル内の不要な機能の無効化、といったアクションが含まれています。

Security Information and Event Management (SIEM) システム

セキュリティ情報管理 (SIM) とセキュリティイベント管理 (SEM) のシステムを組み合わせたツールとサービス。SIEM システムは、サーバー、ネットワーク、デバイス、その他ソースからデータを収集、モニタリング、分析して、脅威やセキュリティ違反を検出し、アラートを発信します。

セキュリティレスポンスの自動化

セキュリティイベントに自動的に応答または修復するように設計された、事前定義されたプログラムされたアクション。これらの自動化は、セキュリティのベストプラクティスを実装するのに役立つ検出的または応答的な AWS セキュリティコントロールとして機能します。自動応答アクションの例としては、VPC セキュリティグループの変更、Amazon EC2 インスタンスへのパッチ適用、認証情報の更新などがあります。

サーバー側の暗号化

送信先にあるデータの、それ AWS のサービス を受け取る による暗号化。

サービスコントロールポリシー (SCP)

AWS Organizationsの組織内の、すべてのアカウントのアクセス許可を一元的に管理するポリシー。SCP は、管理者がユーザーまたはロールに委任するアクションに、ガードレールを定義したり、アクションの制限を設定したりします。SCP は、許可リストまたは拒否リストとして、許可または禁止するサービスやアクションを指定する際に使用できます。詳細については、 AWS Organizations ドキュメントの「サービスコントロールポリシー」を参照してください。

サービスエンドポイント

のエントリポイントの URL AWS のサービス。ターゲットサービスにプログラムで接続するには、エンドポイントを使用します。詳細については、AWS 全般のリファレンスの「AWS のサービス エンドポイント」を参照してください。

サービスレベルアグリーメント (SLA)

サービスのアップタイムやパフォーマンスなど、IT チームがお客様に提供すると約束したものを明示した合意書。

サービスレベルインジケータ (SLI)

エラー率、可用性、スループットなど、サービスのパフォーマンス側面の測定。

サービスレベルの目標 (SLO)

サービスレベルのインジケータによって測定される、サービスの状態を表すターゲットメトリクス。

責任共有モデル

クラウドのセキュリティとコンプライアンス AWS について と共有する責任を説明するモデル。AWS はクラウドのセキュリティを担当しますが、お客様はクラウドのセキュリティを担当します。詳細については、責任共有モデルを参照してください。

SIEM

セキュリティ情報とイベント管理システムを参照してください。

単一障害点 (SPOF)

システムを中断する可能性のあるアプリケーションの 1 つの重要なコンポーネントの障害。

SLA

「サービスレベルの契約」を参照してください。

SLI

「サービスレベルインジケータ」を参照してください。

SLO

「サービスレベルの目標」を参照してください。

スプリットアンドシードモデル

モダナイゼーションプロジェクトのスケーリングと加速のためのパターン。新機能と製品リリースが定義されると、コアチームは解放されて新しい製品チームを作成します。これにより、お客様の組織の能力とサービスの拡張、デベロッパーの生産性の向上、迅速なイノベーションのサポートに役立ちます。詳細については、『』の「アプリケーションをモダナイズするための段階的アプローチ AWS クラウド」を参照してください。

SPOF

単一障害点を参照してください。

スタースキーマ

1 つの大きなファクトテーブルを使用してトランザクションデータまたは測定データを保存し、1 つ以上の小さなディメンションテーブルを使用してデータ属性を保存するデータベース組織構造。この構造は、データウェアハウスまたはビジネスインテリジェンスの目的で使用するように設計されています。

strangler fig パターン

レガシーシステムが廃止されるまで、システム機能を段階的に書き換えて置き換えることにより、モノリシックシステムをモダナイズするアプローチ。このパターンは、宿主の樹木から根を成長させ、最終的にその宿主を包み込み、宿主に取って代わるイチジクのつるを例えています。そのパターンは、モノリシックシステムを書き換えるときのリスクを管理する方法として Martin Fowler により提唱されました。このパターンの適用方法の例については、コンテナと Amazon API Gateway を使用して、従来の Microsoft ASP.NET (ASMX) ウェブサービスを段階的にモダナイズを参照してください。

サブネット

VPC 内の IP アドレスの範囲。サブネットは、1 つのアベイラビリティーゾーンに存在する必要があります。

監視制御とデータ収集 (SCADA)

製造では、ハードウェアとソフトウェアを使用して物理アセットと本番稼働をモニタリングするシステム。

対称暗号化

データの暗号化と復号に同じキーを使用する暗号化のアルゴリズム。

合成テスト

ユーザーとのやり取りをシミュレートして潜在的な問題を検出したり、パフォーマンスをモニタリングしたりする方法でシステムをテストします。Amazon CloudWatch Synthetics を使用して、これらのテストを作成できます。

システムプロンプト

LLM にコンテキスト、指示、またはガイドラインを提供して動作を指示する手法。システムプロンプトは、コンテキストを設定し、ユーザーとのやり取りのルールを確立するのに役立ちます。

T

tags

AWS リソースを整理するためのメタデータとして機能するキーと値のペア。タグは、リソースの管理、識別、整理、検索、フィルタリングに役立ちます。詳細については、「AWS リソースのタグ付け」を参照してください。

ターゲット変数

監督された機械学習でお客様が予測しようとしている値。これは、結果変数 のことも指します。例えば、製造設定では、ターゲット変数が製品の欠陥である可能性があります。

タスクリスト

ランブックの進行状況を追跡するために使用されるツール。タスクリストには、ランブックの概要と完了する必要のある一般的なタスクのリストが含まれています。各一般的なタスクには、推定所要時間、所有者、進捗状況が含まれています。

テスト環境

「環境」を参照してください。

トレーニング

お客様の機械学習モデルに学習するデータを提供すること。トレーニングデータには正しい答えが含まれている必要があります。学習アルゴリズムは入力データ属性をターゲット (お客様が予測したい答え) にマッピングするトレーニングデータのパターンを検出します。これらのパターンをキャプチャする機械学習モデルを出力します。そして、お客様が機械学習モデルを使用して、ターゲットがわからない新しいデータでターゲットを予測できます。

トランジットゲートウェイ

VPC とオンプレミスネットワークを相互接続するために使用できる、ネットワークの中継ハブ。詳細については、 AWS Transit Gateway ドキュメントの「トランジットゲートウェイとは」を参照してください。

トランクベースのワークフロー

デベロッパーが機能ブランチで機能をローカルにビルドしてテストし、その変更をメインブランチにマージするアプローチ。メインブランチはその後、開発環境、本番前環境、本番環境に合わせて順次構築されます。

信頼されたアクセス

ユーザーに代わって AWS Organizations およびそのアカウントで組織内でタスクを実行するために指定したサービスにアクセス許可を付与します。信頼されたサービスは、サービスにリンクされたロールを必要なときに各アカウントに作成し、ユーザーに代わって管理タスクを実行します。詳細については、 ドキュメントの「 を他の AWS のサービス AWS Organizations で使用する AWS Organizations 」を参照してください。

チューニング

機械学習モデルの精度を向上させるために、お客様のトレーニングプロセスの側面を変更する。例えば、お客様が機械学習モデルをトレーニングするには、ラベル付けセットを生成し、ラベルを追加します。これらのステップを、異なる設定で複数回繰り返して、モデルを最適化します。

ツーピザチーム

2 枚のピザで養うことができるくらい小さな DevOps チーム。ツーピザチームの規模では、ソフトウェア開発におけるコラボレーションに最適な機会が確保されます。

U

不確実性

予測機械学習モデルの信頼性を損なう可能性がある、不正確、不完全、または未知の情報を指す概念。不確実性には、次の 2 つのタイプがあります。認識論的不確実性は、限られた、不完全なデータによって引き起こされ、弁論的不確実性 は、データに固有のノイズとランダム性によって引き起こされます。詳細については、深層学習システムにおける不確実性の定量化 ガイドを参照してください。

未分化なタスク

ヘビーリフティングとも呼ばれ、アプリケーションの作成と運用には必要だが、エンドユーザーに直接的な価値をもたらさなかったり、競争上の優位性をもたらしたりしない作業です。未分化なタスクの例としては、調達、メンテナンス、キャパシティプランニングなどがあります。

上位環境

???「環境」を参照してください。

V

バキューミング

ストレージを再利用してパフォーマンスを向上させるために、増分更新後にクリーンアップを行うデータベースのメンテナンス操作。

バージョンコントロール

リポジトリ内のソースコードへの変更など、変更を追跡するプロセスとツール。

VPC ピアリング

プライベート IP アドレスを使用してトラフィックをルーティングできる、2 つの VPC 間の接続。詳細については、Amazon VPC ドキュメントの「VPC ピア機能とは」を参照してください。

脆弱性

システムのセキュリティを脅かすソフトウェアまたはハードウェアの欠陥。

W

ウォームキャッシュ

頻繁にアクセスされる最新の関連データを含むバッファキャッシュ。データベースインスタンスはバッファキャッシュから、メインメモリまたはディスクからよりも短い時間で読み取りを行うことができます。

ウォームデータ

アクセス頻度の低いデータ。この種類のデータをクエリする場合、通常は適度に遅いクエリでも問題ありません。

ウィンドウ関数

現在のレコードに何らかの形で関連する行のグループに対して計算を実行する SQL 関数。ウィンドウ関数は、移動平均の計算や、現在の行の相対位置に基づく行の値へのアクセスなどのタスクの処理に役立ちます。

ワークロード

ビジネス価値をもたらすリソースとコード (顧客向けアプリケーションやバックエンドプロセスなど) の総称。

ワークストリーム

特定のタスクセットを担当する移行プロジェクト内の機能グループ。各ワークストリームは独立していますが、プロジェクト内の他のワークストリームをサポートしています。たとえば、ポートフォリオワークストリームは、アプリケーションの優先順位付け、ウェーブ計画、および移行メタデータの収集を担当します。ポートフォリオワークストリームは、これらの設備を移行ワークストリームで実現し、サーバーとアプリケーションを移行します。

WORM

「Write Once」、「Read Many」を参照してください。

WQF

AWS 「ワークロード認定フレームワーク」を参照してください。

Write Once, Read Many (WORM)

データを 1 回書き込み、データの削除や変更を防ぐストレージモデル。承認されたユーザーは、必要な回数だけデータを読み取ることができますが、変更することはできません。このデータストレージインフラストラクチャはイミュータブルと見なされます。

Z

ゼロデイエクスプロイト

ゼロデイ脆弱性を利用する攻撃、通常はマルウェア。

ゼロデイ脆弱性

実稼働システムにおける未解決の欠陥または脆弱性。脅威アクターは、このような脆弱性を利用してシステムを攻撃する可能性があります。開発者は、よく攻撃の結果で脆弱性に気付きます。

ゼロショットプロンプト

LLM にタスクを実行する手順を提供しますが、タスクのガイドに役立つ例 (ショット) はありません。LLM は、事前トレーニング済みの知識を使用してタスクを処理する必要があります。ゼロショットプロンプトの有効性は、タスクの複雑さとプロンプトの品質によって異なります。「数ショットプロンプト」も参照してください。

ゾンビアプリケーション

平均 CPU およびメモリ使用率が 5% 未満のアプリケーション。移行プロジェクトでは、これらのアプリケーションを廃止するのが一般的です。

翻訳は機械翻訳により提供されています。提供された翻訳内容と英語版の間で齟齬、不一致または矛盾がある場合、英語版が優先します。