Création de solutions de génération augmentée de récupération AWS pour les soins de santé

AWS Conseils prescriptifs

AWS Conseils prescriptifs: Création de solutions de génération augmentée de récupération AWS pour les soins de santé

Table of Contents

Introduction 1
Soins aux patients et productivité 2
Gestion des talents 2
Opportunités et défis 4
Opportunités d'applications d'IA générative dans le secteur de la santé 4
Analyse d'image avancée 5
Défis liés à l'industrialisation des solutions 5
Cas d'utilisation : création d'une application de renseignement médical 6
Présentation de la solution 6
Étape 1 : Découverte des données 8
Étape 2 : Création d'un graphe des connaissances médicales 9
Étape 3 : Création d'agents de récupération du contexte 15
Agents Amazon Bedrock 15
LangChain agents 17
Étape 4 : Création d'une base de connaissances 18
Utilisation du OpenSearch service 19
Création d'une architecture RAG 20
Étape 5 : Génération de réponses 23
Alignement sur le framework AWS Well-Architected 25
Cas d'utilisation : prévision des taux de réadmission 26
Présentation de la solution 26
Étape 1 : Prédire les résultats pour les patients 29
Étape 2 : Prédire le comportement du patient 31
Étape 3 : Prédire la réadmission du patient 33
Étape 4 : Calcul du score de propension 36
Alignement sur le framework AWS Well-Architected 39
Cas d'utilisation : gestion des talents 40
Présentation de la solution 41
Étape 1 : Création d'un profil de compétences 43
Étape 2 : Découverte de la role-to-skill pertinence 44
Étape 3 : Recommander une formation 45
Alignement sur le framework AWS Well-Architected 46
Développement de solutions 48
Amazon Q Developer 48

Design RAG à plusieurs récupérateurs 49
ReAct agents 51
Évaluation des solutions 53
Évaluation de l'extraction d'informations 53
Évaluation de plusieurs récupérateurs 54
Utiliser un LLM 54
Ressources 56
AWS documentation 56
AWS articles de blog 56
Autres ressources 56
Collaborateurs 57
Conception 57
Révision 57
Rédaction technique 57
Historique du document 58
Glossaire 59
..... 59
A 60
B 63
C 65
D 68
E 73
F 75
G 77
H 78
I 80
L 82
M 84
O 88
P 91
Q 94
R 94
S 97
T 101
U 103
V 103

W 104
Z 105
..... cvi

Création de solutions de génération augmentée de récupération AWS pour les soins de santé

Amazon Web Services, Accenture, et Cadiem (contributeurs)

Mars 2025 (historique du document)

Avant les grands modèles de langage (LLMs) et l'IA générative, le développement d'applications automatisées et de haute précision dans le secteur de la santé était un défi. Les méthodes traditionnelles reposaient largement sur la saisie et l'analyse manuelles des données. La complexité de l'analyse de l'imagerie médicale et des dossiers des patients a nécessité une intervention humaine importante, ce qui a souvent entraîné des flux de travail fragmentés et inefficaces. Les progrès des technologies d'intelligence artificielle vous aident à créer des applications hyperpersonnalisées à grande échelle. Les applications de santé peuvent désormais s'intégrer aux bases de connaissances médicales, interpréter les images diagnostiques avec une précision accrue et prévoir les résultats des patients à l'aide de modèles prédictifs.

Ce guide explique comment LLMs révolutionnent les soins de santé grâce à des applications de génération augmentée de récupération avec lesquelles vous pouvez créer. Services AWSLa génération augmentée de récupération (RAG) est une technologie d'IA générative dans laquelle un LLM fait référence à une source de données faisant autorité qui se trouve en dehors de ses sources de données de formation avant de générer une réponse. Les applications RAG fondent les résultats du modèle sur des connaissances du monde réel, ce qui réduit les hallucinations et augmente la pertinence des réponses. Dans le secteur de la santé, le RAG peut être utilisé pour fournir des informations up-to-date médicales précises, garantissant ainsi que les prestataires de soins de santé ont accès aux dernières recherches et directives cliniques. En transformant les données en informations exploitables et en automatisant des processus complexes, ces technologies contribuent à améliorer les soins aux patients, à rationaliser les opérations et à améliorer la productivité des professionnels de santé.

Dans Amazon Bedrock, vous pouvez les affiner LLMs et les intégrer à des agents intelligents pour créer des solutions de santé avancées. Soulignant la synergie entre Amazon OpenSearch Service et Amazon Neptune, le guide montre comment ces services améliorent les solutions RAG grâce à une pertinence de recherche améliorée et à une extraction avancée de données multi-sources. Vous pouvez orchestrer des solutions Amazon Bedrock complètes qui utilisent les agents Amazon Bedrock et LangChainpour coordonner de manière fluide les interactions entre les différents référentiels de

données. Cette intégration démontre le pouvoir de combiner des services spécialisés pour créer des systèmes pilotés par l'IA plus efficaces et efficients.

Soins aux patients et productivité

Ce guide présente deux cas d'utilisation concrets pour les soins aux patients et la productivité : l'augmentation des données sur les patients et la prévision des risques de réadmission. Il fournit des plans stratégiques pour la mise en œuvre de ces solutions à grande échelle, offrant aux établissements de santé une voie claire vers l'industrialisation des processus pilotés par l'IA. Grâce à ces informations, les établissements de santé peuvent utiliser des technologies d'intelligence artificielle avancées pour créer des flux de travail plus efficaces et plus intelligents.

Gestion des talents

Ce guide décrit également des stratégies visant à requalifier les professionnels de santé et à leur donner les moyens d'intégrer facilement l'IA générative dans leur routine quotidienne. Cela peut améliorer à la fois la productivité et la qualité des soins aux patients. En dotant leur personnel des compétences nécessaires pour utiliser efficacement les outils avancés d'intelligence artificielle, les établissements de santé peuvent optimiser leur retour sur investissement et stimuler l'innovation dans les soins aux patients.

Cette solution de gestion des talents basée sur l'IA inclut les fonctionnalités clés suivantes :

- Analyseur intelligent de CV de talents : en utilisant les fonctionnalités avancées LLMs disponibles dans Amazon Bedrock, cet outil extrait et analyse efficacement les compétences et attributs essentiels des CV en matière de talents. Cet outil permet de rationaliser le processus de recrutement.
- Base de connaissances sur les talents : alimentée par Amazon Neptune, cette base de données dynamique fournit des informations en temps réel sur les niveaux de personnel, la répartition des compétences et les tendances du secteur. Cela vous aide à prendre des décisions basées sur les données concernant la gestion des effectifs.
- Moteur de recommandation d'apprentissage — Cet outil piloté par l'IA identifie les lacunes en matière de compétences au sein de l'organisation et recommande des programmes de formation personnalisés pour le personnel médical. Cet outil favorise le développement professionnel continu et aide votre personnel à s'adapter à l'évolution des technologies de santé.

Ensemble, ces fonctionnalités basées sur l'IA aident à optimiser les performances du personnel, révolutionnant ainsi la gestion des talents grâce à une intelligence et une efficacité accrues.

Opportunités et défis

Amazon Bedrock peut améliorer la productivité, l'évolutivité, la rentabilité et fournir des informations basées sur les données. Amazon Bedrock permet aux établissements de santé de les utiliser LLMs efficacement dans différents cas d'utilisation, qu'il s'agisse de création de contenu, d'analyse de données ou de prise de décision automatisée. Ce guide propose des approches pour surmonter les défis courants liés à l'IA générative, tels que les problèmes de qualité des données, l'évolutivité de l'infrastructure, le maintien des performances des modèles et les exigences d'amélioration continue pendant la transition de la preuve de concept à la production.

Opportunités d'applications d'IA générative dans le secteur de la santé

Le secteur de la santé est sur le point de connaître un changement transformateur, stimulé par les opportunités offertes par les applications d'IA générative. L'IA générative a le potentiel d'améliorer les soins aux patients, de rationaliser les opérations et d'accélérer la recherche médicale. En utilisant des modèles d'IA avancés, les prestataires de soins de santé peuvent automatiser l'augmentation des dossiers médicaux. Les dossiers complets et les antécédents des up-to-date patients facilitent des diagnostics et des plans de traitement plus précis. L'analyse d'images pilotée par l'IA, telle que l'interprétation des sonogrammes et d'autres images médicales, peut fournir des informations rapides et précises, réduisant ainsi la charge de travail des professionnels de santé et minimisant le risque d'erreur humaine.

Au-delà du diagnostic et du traitement, l'IA générative peut jouer un rôle essentiel dans l'analyse prédictive. L'analyse prédictive aide les établissements de santé à anticiper les résultats pour les patients et à personnaliser les plans de soins en conséquence. Cette technologie peut également optimiser les processus administratifs, qu'il s'agisse de gérer les données des patients ou de rationaliser la communication entre les prestataires et les patients. En intégrant des solutions d'IA générative aux systèmes de santé existants, les établissements médicaux peuvent gagner en efficacité, réduire les coûts et, en fin de compte, fournir des soins de meilleure qualité. L'intégration de l'IA aux soins de santé n'est pas simplement une amélioration, mais une évolution fondamentale vers des soins plus intelligents, réactifs et centrés sur le patient.

Analyse d'image avancée

La combinaison d'Amazon Bedrock avec des magasins de données, tels qu'Amazon Neptune et OpenSearch Amazon Service, peut vous aider à surmonter les difficultés liées à l'analyse avancée d'images dans le secteur de la santé. Les solutions de récupération d'informations peuvent améliorer le processus de découverte des maladies et améliorer la précision de l'interprétation en évaluant les images diagnostiques et en interprétant les échographies. La solution peut intégrer les données d'évaluation visuelles et textuelles à l'examen manuel de l'évaluation des patients par les médecins.

Défis liés à l'industrialisation des solutions

Les principaux obstacles à surmonter lors de l'industrialisation des solutions d'IA dans le secteur de la santé sont la qualité et la disponibilité des données. Les données de santé existent souvent dans des formats fragmentés et incohérents. Il est essentiel de s'assurer que les modèles d'IA ont accès à des données propres, structurées et représentatives pour maintenir les performances dans des scénarios réels. L'évolutivité de l'infrastructure peut devenir un défi en raison des environnements de production. Ces environnements doivent gérer d'importants volumes de données sur les patients en temps réel tout en offrant des temps de réponse rapides et en respectant les réglementations relatives à la confidentialité des données, telles que la loi HIPAA (Health Insurance Portability and Accountability Act). De plus, compte tenu des nouvelles informations médicales et des données sur les patients qui évoluent au fil du temps, les modèles d'IA doivent être réentraînés et mis à jour pour rester pertinents et donner des recommandations précises. Enfin, l'intégration de ces solutions d'IA dans les systèmes de santé existants peut s'avérer complexe en raison de problèmes d'interopérabilité et de la nécessité de les aligner sur les flux de travail cliniques actuels. Cette intégration nécessite des modifications techniques et opérationnelles.

Cas d'utilisation : création d'une application d'intelligence médicale à partir de données augmentées sur les patients

L'IA générative peut contribuer à améliorer les soins aux patients et la productivité du personnel en améliorant les fonctions cliniques et administratives. L'analyse d'images pilotée par l'IA, telle que l'interprétation des sonogrammes, accélère les processus de diagnostic et améliore la précision. Il peut fournir des informations essentielles qui soutiennent des interventions médicales en temps opportun.

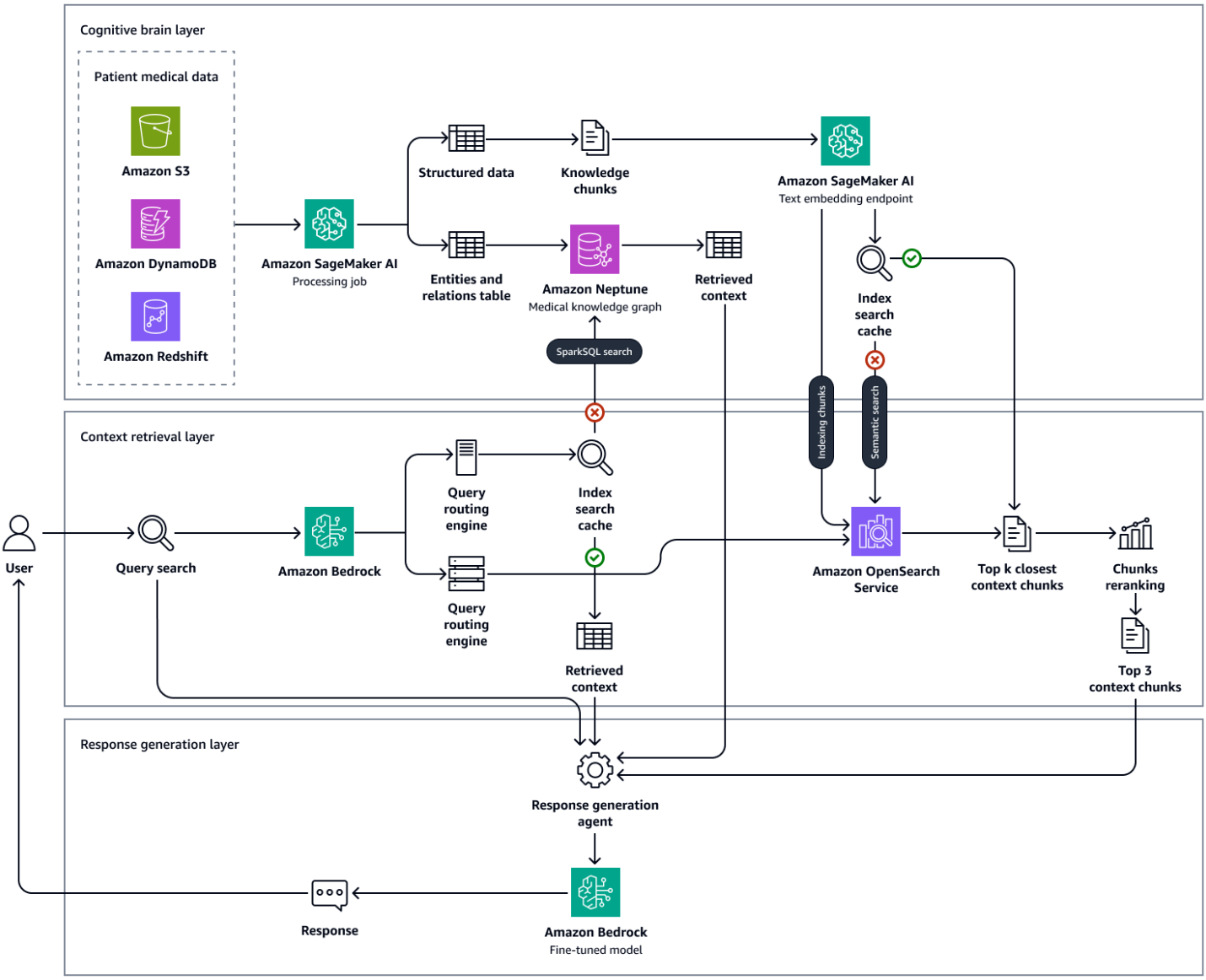
Lorsque vous combinez des modèles d'IA génératifs avec des graphes de connaissances, vous pouvez automatiser l'organisation chronologique des dossiers médicaux électroniques. Cela vous permet d'intégrer les données en temps réel issues des interactions médecin-patient, des symptômes, des diagnostics, des résultats de laboratoire et de l'analyse d'images. Cela fournit au médecin des données complètes sur le patient. Ces données aident le médecin à prendre des décisions médicales plus précises et plus rapides, améliorant à la fois les résultats pour les patients et la productivité des prestataires de soins de santé.

Présentation de la solution

L'IA peut renforcer les capacités des médecins et des cliniciens en synthétisant les données des patients et les connaissances médicales afin de fournir des informations précieuses. Cette solution RAG (Retrieval Augmented Generation) est un moteur d'intelligence médicale qui utilise un ensemble complet de données et de connaissances sur les patients issues de millions d'interactions cliniques. Il exploite le pouvoir de l'IA générative pour créer des informations fondées sur des preuves afin d'améliorer les soins aux patients. Il est conçu pour améliorer les flux de travail cliniques, réduire les erreurs et améliorer les résultats pour les patients.

La solution inclut une capacité de traitement d'image automatique alimentée par. LLMs Cette fonctionnalité réduit le temps que le personnel médical doit consacrer à la recherche manuelle d'images diagnostiques similaires et à l'analyse des résultats de diagnostic.

L'image suivante montre la solution end-to-end-workflow pour cette solution. Il utilise Amazon Neptune, Amazon SageMaker AI, Amazon OpenSearch Service et un modèle de base dans Amazon Bedrock. Pour l'agent de récupération du contexte qui interagit avec le graphe des connaissances médicales dans Neptune, vous pouvez choisir entre un agent Amazon Bedrock et un LangChain agent.



Lors de nos expériences avec des exemples de questions médicales, nous avons observé que les réponses finales générées par notre approche utilisant le graphe de connaissances géré dans Neptune, une base de données OpenSearch vectorielle hébergeant la base de connaissances cliniques et Amazon Bedrock LLMs étaient fondées sur des faits et étaient bien plus précises en réduisant le nombre de faux positifs et en augmentant le nombre de vrais positifs. Cette solution peut générer des informations factuelles sur l'état de santé du patient et vise à améliorer les flux de travail cliniques, à réduire les erreurs et à améliorer les résultats pour les patients.

La création de cette solution comprend les étapes suivantes :

- Étape 1 : Découverte des données

- Étape 2 : Création d'un graphe des connaissances médicales
- Étape 3 : Création d'agents de récupération du contexte pour interroger le graphe des connaissances médicales
- Étape 4 : Création d'une base de connaissances contenant des données descriptives en temps réel
- Étape 5 : Utilisation LLMs pour répondre à des questions médicales

Étape 1 : Découverte des données

Il existe de nombreux ensembles de données médicales open source que vous pouvez utiliser pour soutenir le développement d'une solution de santé basée sur l'IA. L'un de ces ensembles de données est le jeu de données MIMIC-IV, qui est un ensemble de données de dossiers de santé électroniques (DSE) accessible au public qui est largement utilisé dans le milieu de la recherche en santé. Le MIMIC-IV contient des informations cliniques détaillées, notamment des notes de sortie en texte libre extraites des dossiers des patients. Vous pouvez utiliser ces enregistrements pour expérimenter des techniques de sommation de texte et d'extraction d'entités. Ces techniques vous aident à extraire des informations médicales (telles que les symptômes du patient, les médicaments administrés et les traitements prescrits) à partir de texte non structuré.

Vous pouvez également utiliser un ensemble de données qui fournit des résumés de sortie de patients annotés et anonymisés, spécialement conçus à des fins de recherche. Un ensemble de données récapitulatives sur les sorties peut vous aider à expérimenter l'extraction d'entités, ce qui vous permet d'identifier les entités médicales clés (telles que les affections, les procédures et les médicaments) à partir du texte. Étape 2 : Création d'un graphe des connaissances médicalesdans ce guide, vous explique comment utiliser les données structurées extraites des ensembles de données MIMIC-IV et récapitulatifs des sorties pour créer un graphique des connaissances médicales. Ce graphe de connaissances médicales sert de colonne vertébrale aux systèmes avancés d'interrogation et d'aide à la décision destinés aux professionnels de santé.

Outre les jeux de données textuels, vous pouvez utiliser des jeux de données d'images. Par exemple, le jeu de données sur les radiographies musculosquelettiques (MURA), qui est une base de données complète d'images radiographiques multivues des os. Utilisez ces ensembles de données d'images pour expérimenter l'évaluation diagnostique à l'aide de techniques de décodage d'images médicales. Ces techniques de décodage sont cruciales pour le diagnostic précoce de maladies telles que les maladies musculosquelettiques, les maladies cardiovasculaires et l'ostéoporose. En affinant les modèles de base de la vision et du langage sur le jeu de données d'images médicales, vous

pouvez détecter des anomalies dans les images diagnostiques. Cela permet au système de fournir aux cliniciens des informations diagnostiques précoces et précises. En utilisant des ensembles de données d'images et de textes, vous pouvez créer une application de santé pilotée par l'IA capable de traiter à la fois du texte et des données d'image pour améliorer les soins aux patients.

Étape 2 : Création d'un graphe des connaissances médicales

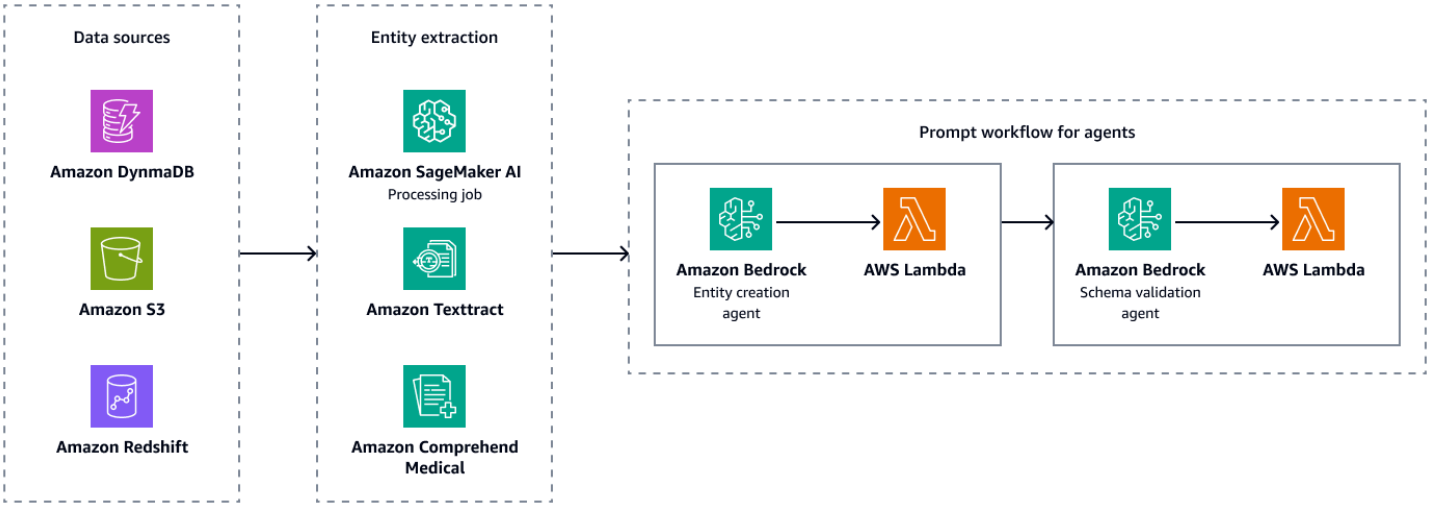
Pour tout établissement de santé qui souhaite créer un système d'aide à la décision basé sur une base de connaissances massive, l'un des principaux défis consiste à localiser et à extraire les entités médicales présentes dans les notes cliniques, les revues médicales, les résumés de sortie et d'autres sources de données. Vous devez également capturer les relations temporelles, les sujets et les évaluations de certitude de ces dossiers médicaux afin d'utiliser efficacement les entités, les attributs et les relations extraits.

La première étape consiste à extraire des concepts médicaux d'un texte médical non structuré en utilisant quelques instructions pour un modèle de base, tel que Llama 3 dans Amazon Bedrock. L'invite instantanée consiste à fournir à un LLM un petit nombre d'exemples illustrant la tâche et le résultat souhaité avant de lui demander d'effectuer une tâche similaire. À l'aide d'un extracteur d'entités médicales basé sur le LLM, vous pouvez analyser le texte médical non structuré, puis générer une représentation des données structurées des entités de connaissances médicales. Vous pouvez également stocker les attributs du patient à des fins d'analyse et d'automatisation en aval. Le processus d'extraction des entités inclut les actions suivantes :

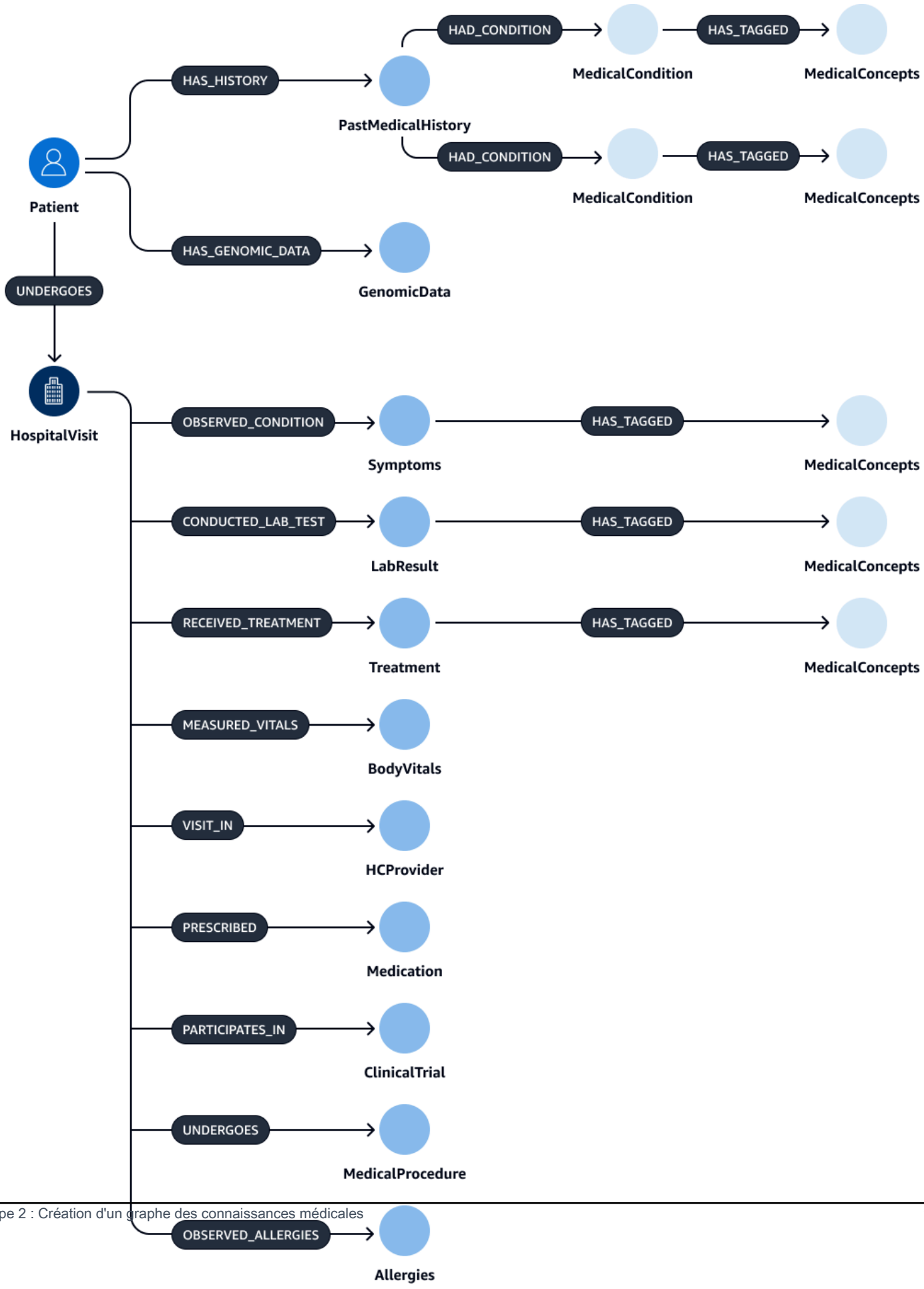
- Extrayez des informations sur des concepts médicaux, tels que les maladies, les médicaments, les dispositifs médicaux, la posologie, la fréquence des médicaments, la durée des médicaments, les symptômes, les procédures médicales et leurs attributs cliniquement pertinents.
- Capturez les caractéristiques fonctionnelles, telles que les relations temporelles entre les entités extraites, les sujets et les évaluations de certitude.
- Élargissez les vocabulaires médicaux standard, tels que les suivants :
  - Identifiants conceptuels (RxCUI) issus de la base de données RxNorm
  - Codes de la classification internationale des maladies, 10e révision, modification clinique (ICD-10-CM)
  - Termes tirés des rubriques médicales (MeSH)
  - Concepts issus de la nomenclature systématisée de la médecine, termes cliniques (SNOMED CT)
  - Codes issus du système de langage médical unifié (UMLS)

- Résumez les notes de sortie et tirez des informations médicales des transcriptions.

La figure suivante montre les étapes d'extraction d'entités et de validation du schéma pour créer des combinaisons appariées valides d'entités, d'attributs et de relations. Vous pouvez stocker des données non structurées, telles que des résumés de sortie ou des notes de patients, dans Amazon Simple Storage Service (Amazon S3). Vous pouvez stocker des données structurées, telles que les données de planification des ressources d'entreprise (ERP), les dossiers médicaux électroniques et les systèmes d'information de laboratoire, dans Amazon Redshift et Amazon DynamoDB. Vous pouvez créer un agent de création d'entités Amazon Bedrock. Cet agent peut intégrer des services, tels que les pipelines d'extraction de données Amazon SageMaker AI, Amazon Textract et Amazon Comprehend Medical, pour extraire des entités, des relations et des attributs à partir de sources de données structurées et non structurées. Enfin, vous utilisez un agent de validation de schéma Amazon Bedrock pour vous assurer que les entités et les relations extraites sont conformes au schéma de graphe prédéfini et préservent l'intégrité des connexions au bord du nœud et des propriétés associées.



Après extraction et validation des entités, des relations et des attributs, vous pouvez les lier pour créer un subject-object-predicate triplet. Vous ingérez ces données dans une base de données de graphes Amazon Neptune, comme illustré dans la figure suivante. Les bases de données de graphes sont optimisées pour stocker et interroger les relations entre les éléments de données.


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

Vous pouvez créer un graphe de connaissances complet à partir de ces données. Un graphe de connaissances vous aide à organiser et à interroger toutes sortes d'informations connectées. Par exemple, vous pouvez créer un graphe de connaissances comportant les nœuds principaux suivants : `HospitalVisitPastMedicalHistory`,`Symptoms`,`Medication`,`MedicalProcedures`, et`Treatment`.

Les tableaux suivants répertorient les entités et leurs attributs que vous pouvez extraire des notes de décharge.

Entité	Attributs
`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`

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

Le tableau suivant répertorie les relations que les entités peuvent avoir et leurs attributs correspondants. Par exemple, l'`Patient`entité peut se connecter à l'`HospitalVisit`entité associée à la `[UNDERGOES]` relation. L'attribut de cette relation est`VisitDate`.

Entité visée	Relation	Entité d'objet	Attributs
`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`

Entité visée	Relation	Entité d'objet	Attributs
HospitalVisit	[PRESCRIBED]	Medication	Duration, Dosage, Adherence , VisitDate
Patient	[HAS_HISTORY]	PastMedicalHistory	Aucun
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	Aucun
LabResult	[HAS_TAGGED]	MedicalConcepts	Aucun
Treatment	[HAS_TAGGED]	MedicalConcepts	Aucun
Symptoms	[HAS_TAGGED]	MedicalConcepts	Aucun

Étape 3 : Création d'agents de récupération du contexte pour interroger le graphe des connaissances médicales

Après avoir créé la base de données de graphes médicaux, l'étape suivante consiste à créer des agents pour l'interaction graphique. Les agents récupèrent le contexte correct et requis pour la requête saisie par un médecin ou un clinicien. Il existe plusieurs options pour configurer ces agents qui récupèrent le contexte à partir du Knowledge Graph :

- Agents Amazon Bedrock
- LangChain agents

Agents Amazon Bedrock pour l'interaction graphique

Les agents Amazon Bedrock fonctionnent parfaitement avec les bases de données graphiques Amazon Neptune. Vous pouvez effectuer des interactions avancées via les groupes d'actions Amazon Bedrock. Le groupe d'actions lance le processus en appelant une AWS Lambda fonction qui exécute les requêtes Neptune OpenCypher.

Pour interroger un graphe de connaissances, vous pouvez utiliser deux approches distinctes : exécution directe de requêtes ou interrogation avec intégration de contexte. Ces approches peuvent être appliquées indépendamment ou combinées, en fonction de votre cas d'utilisation spécifique et de vos critères de classement. En combinant les deux approches, vous pouvez fournir un contexte plus complet au LLM, ce qui peut améliorer les résultats. Les deux approches d'exécution des requêtes sont les suivantes :

- Exécution directe de requêtes Cypher sans intégration — La fonction Lambda exécute des requêtes directement sur Neptune sans aucune recherche basée sur les intégrations. Voici un exemple de cette approche :

```
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
```

- Exécution directe de requêtes Cypher à l'aide de la recherche intégrée — La fonction Lambda utilise la recherche incorporée pour améliorer les résultats des requêtes. Cette approche améliore l'exécution des requêtes en incorporant des intégrations, qui sont des représentations vectorielles denses de données. Les intégrations sont particulièrement utiles lorsque la requête nécessite une

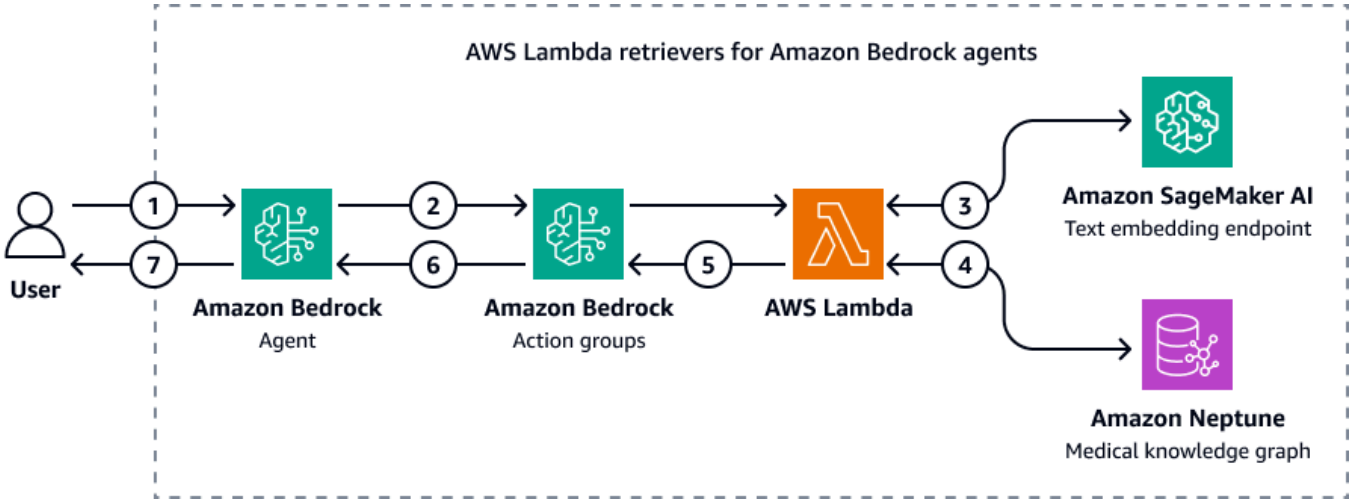
similitude sémantique ou une compréhension plus large au-delà des correspondances exactes. Vous pouvez utiliser des modèles pré-entraînés ou personnalisés pour générer des intégrations pour chaque condition médicale. Voici un exemple de cette approche :

```
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
```

Dans cet exemple, la `search_embedding("Acute Diabetes")` fonction récupère des affections dont la sémantique est proche du « diabète aigu ». Cela permet à la requête de trouver également des patients atteints de maladies telles que le prédiabète ou le syndrome métabolique.

L'image suivante montre comment les agents Amazon Bedrock interagissent avec Amazon Neptune afin d'exécuter une requête Cypher sur un graphe de connaissances médicales.



Le schéma suivant illustre le flux de travail suivant :

1. L'utilisateur soumet une question à l'agent Amazon Bedrock.
2. L'agent Amazon Bedrock transmet la question et les variables du filtre d'entrée aux groupes d'action Amazon Bedrock. Ces groupes d'actions contiennent une AWS Lambda fonction qui interagit avec le point de terminaison d'intégration de texte Amazon SageMaker AI et le graphe de connaissances médicales Amazon Neptune.

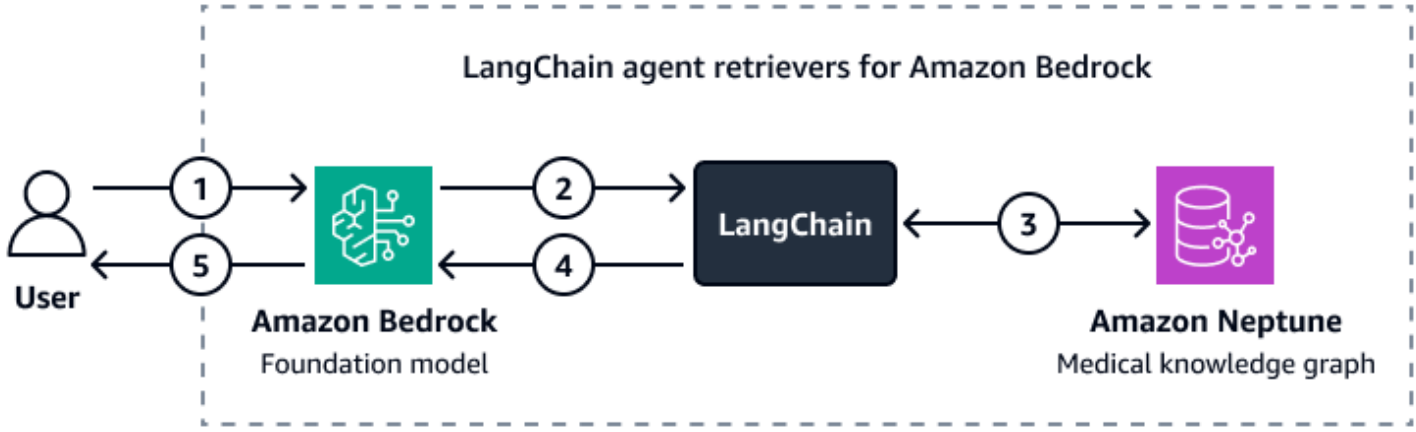
3. La fonction Lambda s'intègre au point de terminaison d'intégration de texte SageMaker AI pour effectuer une recherche sémantique dans la requête OpenCypher. Il convertit la requête en langage naturel en une requête OpenCypher en utilisant le sous-jacent LangChain agents.
4. La fonction Lambda interroge le graphe de connaissances médicales de Neptune pour trouver le jeu de données correct et reçoit le résultat du graphe de connaissances médicales de Neptune.
5. La fonction Lambda renvoie les résultats de Neptune aux groupes d'actions Amazon Bedrock.
6. Les groupes d'action Amazon Bedrock envoient le contexte récupéré à l'agent Amazon Bedrock.
7. L'agent Amazon Bedrock génère la réponse en utilisant la requête utilisateur d'origine et le contexte extrait du graphe de connaissances.

LangChain agents pour l'interaction graphique

Vous pouvez intégrer LangChain avec Neptune pour permettre des requêtes et des extractions basées sur des graphes. Cette approche peut améliorer les flux de travail basés sur l'IA en utilisant les fonctionnalités de la base de données de graphes de Neptune. La coutume LangChain retriever joue le rôle d'intermédiaire. Le modèle de base d'Amazon Bedrock peut interagir avec Neptune en utilisant à la fois des requêtes Cypher directes et des algorithmes de graphes plus complexes.

Vous pouvez utiliser le récupérateur personnalisé pour affiner la façon dont LangChain l'agent interagit avec les algorithmes du graphe Neptune. Par exemple, vous pouvez utiliser des instructions ponctuelles, qui vous permettent d'adapter les réponses du modèle de base en fonction de modèles ou d'exemples spécifiques. Vous pouvez également appliquer des filtres identifiés par le LLM pour affiner le contexte et améliorer la précision des réponses. Cela peut améliorer l'efficacité et la précision de l'ensemble du processus de récupération lors de l'interaction avec des données graphiques complexes.

L'image suivante montre comment une personnalisation LangChain un agent orchestre l'interaction entre un modèle de fondation Amazon Bedrock et un graphe de connaissances médicales Amazon Neptune.



Le schéma suivant illustre le flux de travail suivant :

1. Un utilisateur soumet une question à Amazon Bedrock et au LangChain agent.
2. Le modèle de fondation Amazon Bedrock utilise le schéma Neptune, fourni par LangChain agent, pour générer une requête pour la question de l'utilisateur.
3. Le LangChain l'agent exécute la requête sur le graphe de connaissances médicales Amazon Neptune.
4. Le LangChain l'agent envoie le contexte récupéré au modèle de fondation Amazon Bedrock.
5. Le modèle Amazon Bedrock Foundation utilise le contexte extrait pour générer une réponse à la question de l'utilisateur.

Étape 4 : Création d'une base de connaissances contenant des données descriptives en temps réel

Ensuite, vous créez une base de connaissances contenant des notes descriptives en temps réel sur les interactions médecin-patient, des évaluations d'images diagnostiques et des rapports d'analyse de laboratoire. Cette base de connaissances est une base de données vectorielle. En utilisant une base de données vectorielle, qui peut stocker des connaissances médicales descriptives sous une forme indexée et vectorisée, les prestataires de soins de santé peuvent consulter et accéder efficacement aux informations pertinentes à partir d'un vaste référentiel. Ces représentations vectorisées vous aident à récupérer des données sémantiquement similaires. Les prestataires de soins peuvent parcourir rapidement les notes cliniques, les images médicales et les résultats de laboratoire. Cela accélère la prise de décision éclairée en offrant un accès instantané à des

informations contextuelles pertinentes, améliorant ainsi la précision et la rapidité des diagnostics et des plans de traitement.

Utilisation d'une base de connaissances médicales de OpenSearch service

Amazon OpenSearch Service peut gérer de gros volumes de données médicales de grande dimension. Il s'agit d'un service géré qui facilite les recherches de haute performance et les analyses en temps réel. Elle convient parfaitement comme base de données vectorielle pour les applications RAG. OpenSearch Le service agit comme un outil principal pour gérer de grandes quantités de données non structurées ou semi-structurées, telles que les dossiers médicaux, les articles de recherche et les notes cliniques. Ses fonctionnalités avancées de recherche sémantique vous aident à récupérer des informations contextuellement pertinentes. Cela le rend particulièrement utile dans des applications telles que les systèmes d'aide à la décision clinique, les outils de résolution des requêtes des patients et les systèmes de gestion des connaissances dans le domaine de la santé. Par exemple, un clinicien peut rapidement trouver des données pertinentes sur les patients ou des études de recherche correspondant à des symptômes ou à des protocoles de traitement spécifiques. Cela aide les cliniciens à prendre des décisions fondées sur les informations les plus pertinentes up-to-date et les plus pertinentes.

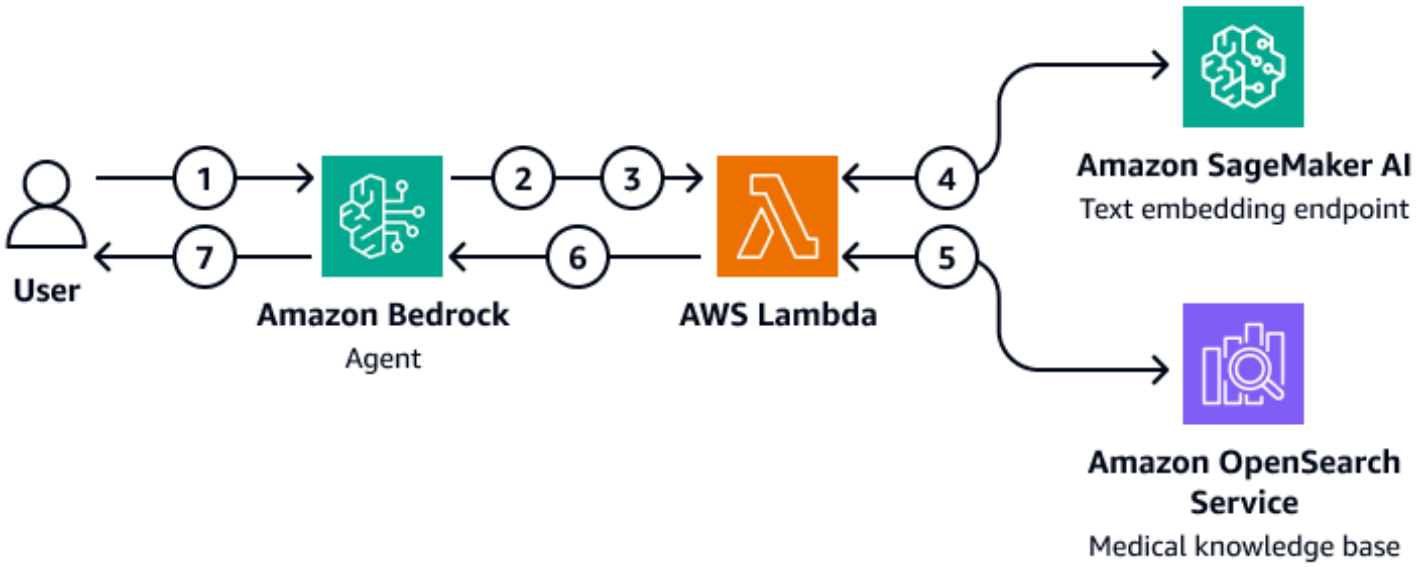
OpenSearch Le service peut évoluer et gérer l'indexation et l'interrogation des données en temps réel. Il est donc idéal pour les environnements de santé dynamiques où l'accès rapide à des informations précises est essentiel. En outre, il dispose de fonctionnalités de recherche multimodales optimales pour les recherches nécessitant plusieurs entrées, telles que des images médicales et des notes de médecin. Lors de la mise en œuvre du OpenSearch service pour les applications de santé, il est essentiel de définir des champs et des mappages précis afin d'optimiser l'indexation et la récupération des données. Les champs représentent les différents éléments de données, tels que les dossiers des patients, les antécédents médicaux et les codes de diagnostic. Les mappages définissent la manière dont ces champs sont stockés (sous forme incorporée ou sous forme originale) et interrogés. Pour les applications de santé, il est essentiel d'établir des mappages adaptés à différents types de données, notamment les données structurées (telles que les résultats de tests numériques), les données semi-structurées (telles que les notes des patients) et les données non structurées (telles que les images médicales)

Dans OpenSearch Service, vous pouvez effectuer des requêtes de recherche neurale en texte intégral à l'aide d'instructions organisées pour effectuer des recherches dans les dossiers médicaux, les notes cliniques ou les documents de recherche afin de trouver rapidement des informations pertinentes sur des symptômes, des traitements ou des antécédents de patients spécifiques. Les requêtes de recherche neuronale gèrent automatiquement l'intégration de l'invite d'entrée et des

images à l'aide de modèles de réseaux neuronaux intégrés. Cela l'aide à comprendre et à saisir les relations sémantiques les plus profondes des données multimodales, offrant ainsi des résultats de recherche plus précis et plus sensibles au contexte par rapport à d'autres algorithmes de requête de recherche, tels que la recherche K-Nearest Neighbor (K-nn).

Création d'une architecture RAG

Vous pouvez déployer une solution RAG personnalisée qui utilise les agents Amazon Bedrock pour interroger une base de connaissances médicales dans OpenSearch Service. Pour ce faire, vous créez une AWS Lambda fonction qui peut interagir avec le OpenSearch service et l'interroger. La fonction Lambda intègre la question saisie par l'utilisateur en accédant à un point de terminaison d'intégration de texte basé sur SageMaker l'IA. L'agent Amazon Bedrock transmet des paramètres de requête supplémentaires en tant qu'entrées à la fonction Lambda. La fonction interroge la base de connaissances médicales dans OpenSearch Service, qui renvoie le contenu médical pertinent. Après avoir configuré la fonction Lambda, ajoutez-la en tant que groupe d'action dans l'agent Amazon Bedrock. L'agent Amazon Bedrock prend les données saisies par l'utilisateur, identifie les variables nécessaires, transmet les variables et la question à la fonction Lambda, puis lance la fonction. La fonction renvoie un contexte qui aide le modèle de base à fournir une réponse plus précise à la question de l'utilisateur.

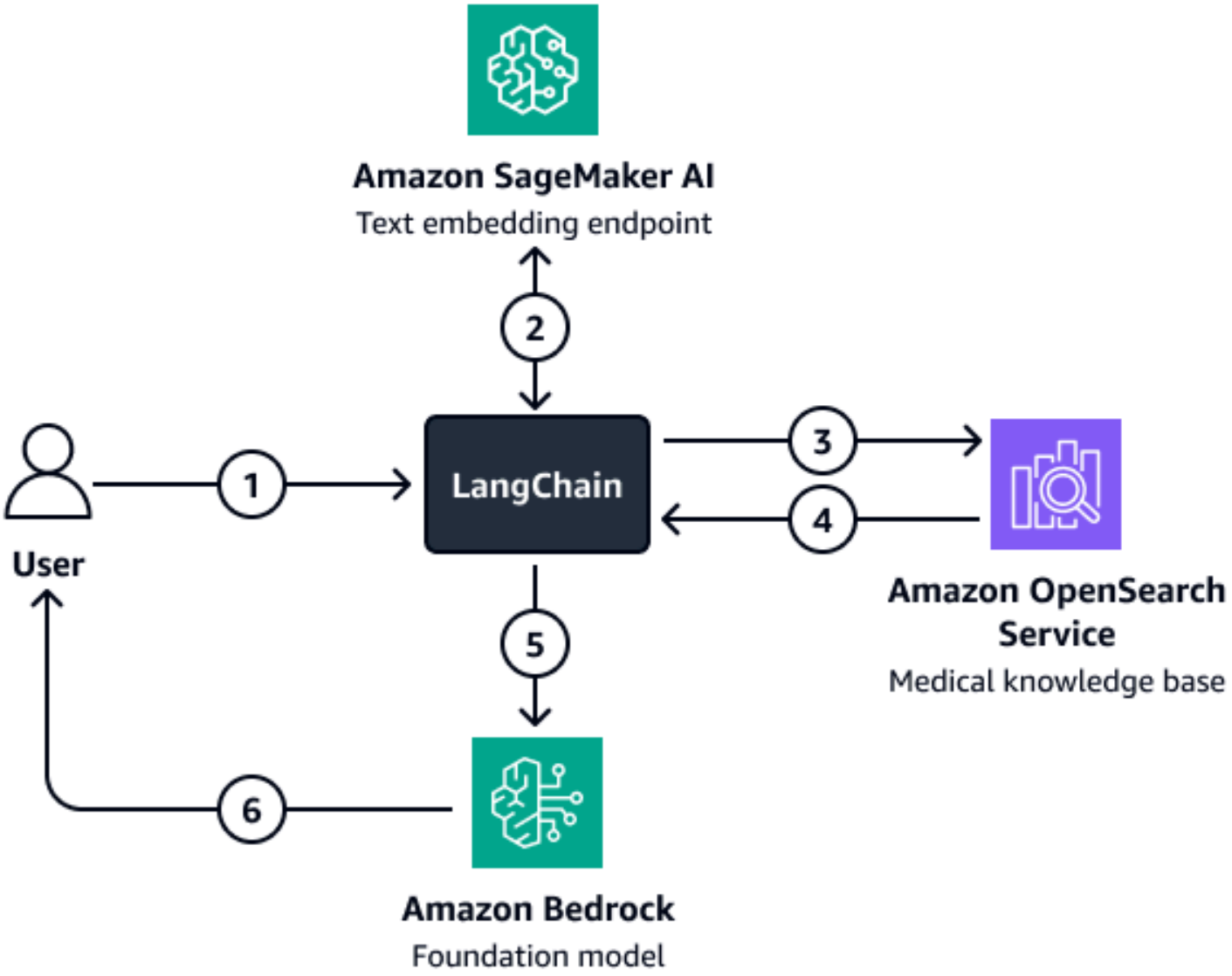


Le schéma suivant illustre le flux de travail suivant :

1. Un utilisateur soumet une question à l'agent Amazon Bedrock.
2. L'agent Amazon Bedrock sélectionne le groupe d'action à lancer.

3. L'agent Amazon Bedrock lance une AWS Lambda fonction et lui transmet des paramètres.
4. La fonction Lambda lance le modèle d'intégration de texte Amazon SageMaker AI pour intégrer la question de l'utilisateur.
5. La fonction Lambda transmet le texte intégré ainsi que les paramètres et filtres supplémentaires à Amazon OpenSearch Service. Amazon OpenSearch Service interroge la base de connaissances médicales et renvoie les résultats à la fonction Lambda.
6. La fonction Lambda transmet les résultats à l'agent Amazon Bedrock.
7. Le modèle de base de l'agent Amazon Bedrock génère une réponse basée sur les résultats et renvoie la réponse à l'utilisateur.

Pour les situations impliquant un filtrage plus complexe, vous pouvez utiliser une option personnalisée LangChain retriever. Créez ce récupérateur en configurant un client de recherche vectorielle de OpenSearch service qui est chargé directement dans LangChain. Cette architecture vous permet de transmettre davantage de variables afin de créer les paramètres du filtre. Une fois le récupérateur configuré, utilisez le modèle et le récupérateur Amazon Bedrock pour configurer une chaîne de questions-réponses. Cette chaîne orchestre l'interaction entre le modèle et le récupérateur en transmettant les entrées utilisateur et les filtres potentiels au récupérateur. Le récupérateur renvoie un contexte pertinent qui aide le modèle de base à répondre à la question de l'utilisateur.



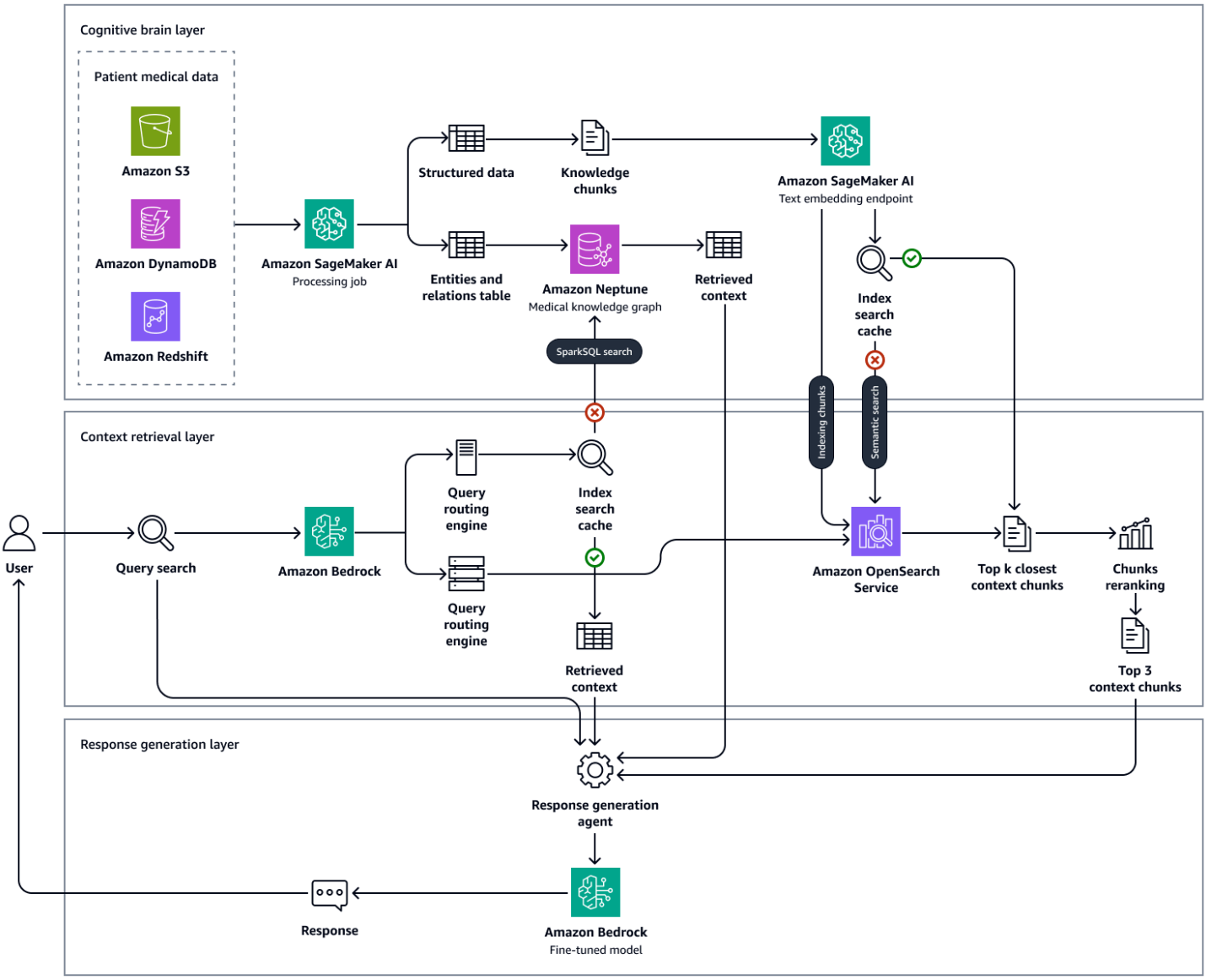
Le schéma suivant illustre le flux de travail suivant :

1. Un utilisateur soumet une question au LangChain agent retriever.
2. Le LangChain l'agent retriever envoie la question au point de terminaison d'intégration de texte Amazon SageMaker AI pour intégrer la question.
3. Le LangChain l'agent retriever transmet le texte intégré à Amazon OpenSearch Service.
4. Amazon OpenSearch Service renvoie les documents récupérés au LangChain agent retriever.
5. Le LangChain l'agent retriever transmet la question de l'utilisateur et le contexte récupéré au modèle Amazon Bedrock Foundation.
6. Le modèle de base génère une réponse et l'envoie à l'utilisateur.

Étape 5 : Utilisation LLMs pour répondre à des questions médicales

Les étapes précédentes vous aident à créer une application de renseignement médical capable de récupérer les dossiers médicaux d'un patient et de résumer les médicaments pertinents et les diagnostics potentiels. À présent, vous créez la couche de génération. Cette couche utilise les capacités génératives d'un LLM dans Amazon Bedrock, comme Llama 3, pour augmenter le rendement de l'application.

Lorsqu'un clinicien saisit une requête, la couche de récupération du contexte de l'application exécute le processus de récupération à partir du graphe de connaissances et renvoie les principaux enregistrements relatifs à l'historique, à la démographie, aux symptômes, au diagnostic et aux résultats du patient. À partir de la base de données vectorielle, il extrait également des notes descriptives en temps réel sur les interactions médecin-patient, des informations sur l'évaluation des images diagnostiques, des résumés de rapports d'analyse de laboratoire et des informations issues d'un vaste corpus de recherches médicales et de livres universitaires. Les meilleurs résultats obtenus, la requête du clinicien et les instructions (qui sont conçues pour sélectionner les réponses en fonction de la nature de la requête) sont ensuite transmis au modèle de base d'Amazon Bedrock. Il s'agit de la couche de génération de réponses. Le LLM utilise le contexte récupéré pour générer une réponse à la requête du clinicien. La figure suivante montre le end-to-end flux de travail des étapes de cette solution.



Vous pouvez utiliser un modèle de base préformé dans Amazon Bedrock, tel que Llama 3, pour une gamme de cas d'utilisation que l'application de renseignement médical doit gérer. Le LLM le plus efficace pour une tâche donnée varie en fonction du cas d'utilisation. Par exemple, un modèle préformé peut être suffisant pour résumer les conversations patient-médecin, effectuer des recherches dans les médicaments et les antécédents des patients, et extraire des informations à partir d'ensembles de données médicales internes et de corpus de connaissances scientifiques. Cependant, un LLM affiné peut être nécessaire pour d'autres cas d'utilisation complexes, tels que les évaluations de laboratoire en temps réel, les recommandations de procédures médicales et les prédictions des résultats pour les patients. Vous pouvez affiner un LLM en l'entraînant sur des ensembles de données du domaine médical. Les exigences spécifiques ou complexes en matière de santé et de sciences de la vie orientent le développement de ces modèles affinés.

Pour plus d'informations sur la mise au point d'un LLM ou sur le choix d'un LLM existant qui a été formé sur les données du domaine médical, voir Utilisation de grands modèles linguistiques pour les cas d'utilisation dans les domaines de la santé et des sciences de la vie.

Alignement sur le framework AWS Well-Architected

La solution s'aligne sur les six piliers du AWS Well-Architected Framework comme suit :

- Excellence opérationnelle — L'architecture est découplée pour une surveillance et des mises à jour efficaces. Les agents Amazon Bedrock vous AWS Lambda aident à déployer et à annuler rapidement des outils.
- Sécurité — Cette solution est conçue pour se conformer aux réglementations sanitaires, telles que la loi HIPAA. Vous pouvez également implémenter le chiffrement, un contrôle d'accès précis et des garde-corps Amazon Bedrock pour protéger les données des patients.
- Fiabilité : les services AWS gérés, tels qu'Amazon OpenSearch Service et Amazon Bedrock, fournissent l'infrastructure nécessaire à une interaction continue avec les modèles.
- Efficacité des performances — La solution RAG récupère rapidement les données pertinentes en utilisant une recherche sémantique optimisée et des requêtes Cypher, tandis qu'un agent-routeur identifie les itinéraires optimaux pour les requêtes des utilisateurs.
- Optimisation des coûts — Le pay-per-token modèle d'Amazon Bedrock et de l'architecture RAG réduit les coûts d'inférence et de pré-formation.
- Durabilité — L'utilisation d'une infrastructure et de pay-per-token calcul sans serveur minimise l'utilisation des ressources et améliore la durabilité.

Cas d'utilisation : prévision des résultats pour les patients et des taux de réadmission

Les analyses prédictives basées sur l'IA offrent d'autres avantages en prévoyant les résultats pour les patients et en permettant des plans de traitement personnalisés. Cela peut améliorer la satisfaction des patients et les résultats de santé. En intégrant ces capacités d'IA à Amazon Bedrock et à d'autres technologies, les prestataires de soins de santé peuvent réaliser des gains de productivité significatifs, réduire les coûts et améliorer la qualité globale des soins aux patients.

Vous pouvez stocker des données médicales, telles que les antécédents des patients, les notes cliniques, les médicaments et les traitements, dans un graphe de connaissances. En combinant la compréhension contextuelle approfondie des LLMs données temporelles structurées d'un graphe de connaissances médicales, les prestataires de soins de santé peuvent obtenir des informations supplémentaires sur les modèles individuels des patients. Grâce à l'analyse prédictive, vous pouvez identifier à un stade précoce les cas potentiels de non-observance ou de complications liées au traitement et générer des scores personnalisés de propension à la réadmission.

Cette solution vous permet de prévoir la probabilité d'une réadmission. Ces prévisions peuvent améliorer les résultats pour les patients et réduire les coûts des soins de santé. Cette solution peut également aider les cliniciens et les administrateurs des hôpitaux à concentrer leur attention sur les patients présentant un risque élevé de réadmission. Cela les aide également à lancer des interventions proactives auprès de ces patients par le biais d'alertes, de libre-service et d'actions basées sur les données.

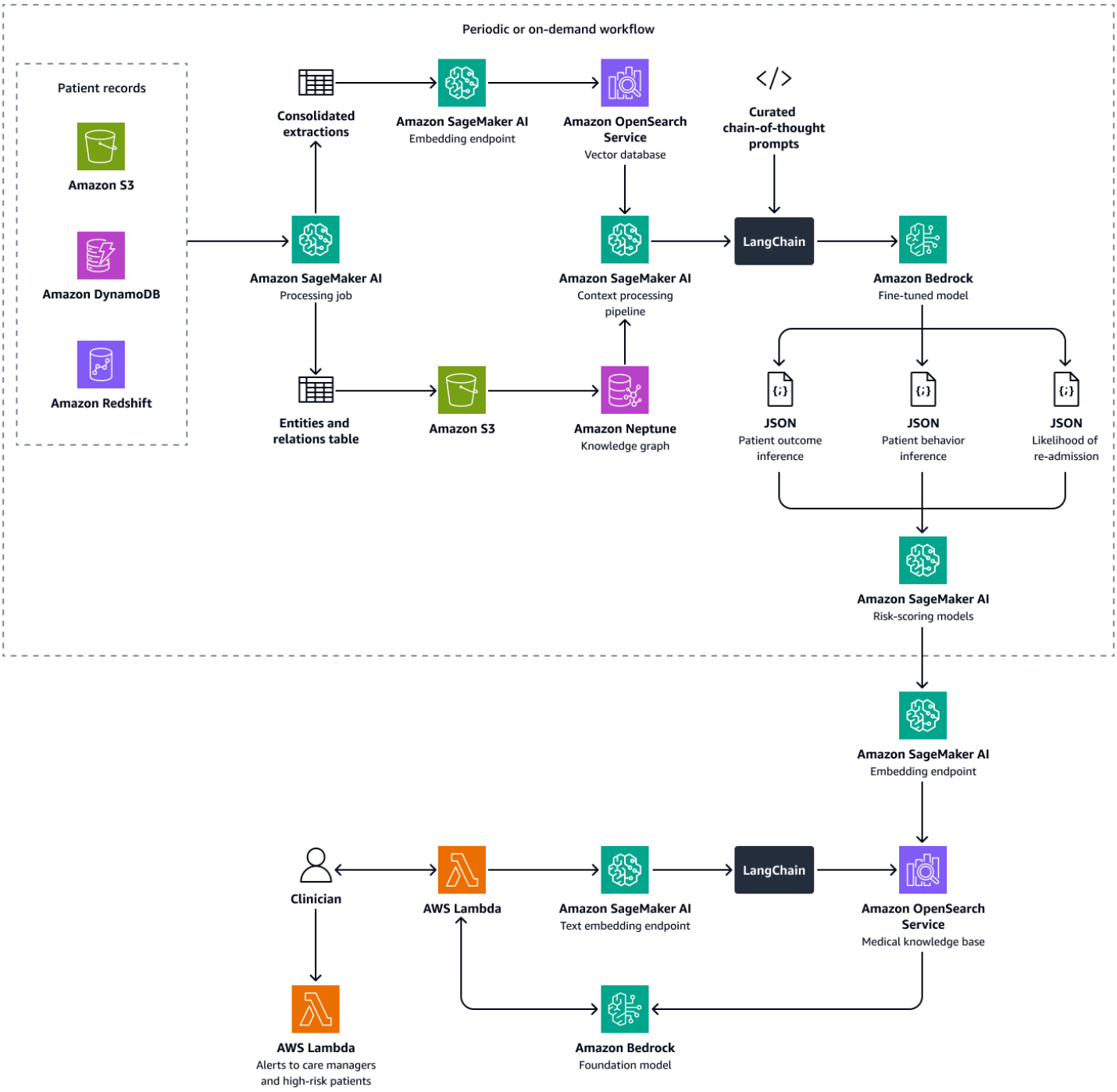
Présentation de la solution

Cette solution utilise un framework de génération augmentée de récupération multiple (RAG) pour analyser les données des patients. Il prédit la probabilité de réadmission à l'hôpital pour chaque patient et vous aide à calculer un score de propension à la réadmission au niveau de l'hôpital. Cette solution intègre les fonctionnalités suivantes :

- Graphe de connaissances — Stocke des données structurées et chronologiques sur les patients, telles que les consultations à l'hôpital, les réadmissions précédentes, les symptômes, les résultats de laboratoire, les traitements prescrits et l'historique d'observance du traitement médicamenteux

- Base de données vectorielle — Stocke des données cliniques non structurées, telles que les résumés des sorties, les notes du médecin et les enregistrements des rendez-vous manqués ou des effets secondaires signalés des médicaments
- LLM affiné — Consomme à la fois les données structurées du graphe de connaissances et les données non structurées de la base de données vectorielles afin de générer des inférences sur le comportement du patient, l'observance du traitement et la probabilité de réadmission

Les modèles de notation des risques quantifient les inférences du LLM sous forme de scores numériques. Vous pouvez agréger les scores pour obtenir un score de propension à la réadmission au niveau de l'hôpital. Ce score définit l'exposition au risque de chaque patient, et vous pouvez le calculer périodiquement ou selon les besoins. Toutes les inférences et les scores de risque sont indexés et stockés dans Amazon OpenSearch Service afin que les responsables de soins et les cliniciens puissent les récupérer. En intégrant un agent d'intelligence artificielle conversationnel à cette base de données vectorielle, les cliniciens et les responsables de soins peuvent extraire facilement des informations au niveau du patient, de l'ensemble de l'établissement ou de la spécialité médicale. Vous pouvez également configurer des alertes automatisées en fonction des scores de risque, ce qui encourage les interventions proactives.



La création de cette solution comprend les étapes suivantes :

- Étape 1 : Prédire les résultats des patients à l'aide d'un graphe de connaissances médicales
- Étape 2 : Prédire le comportement du patient à l'égard des médicaments ou traitements prescrits
- Étape 3 : Prédire la probabilité de réadmission du patient

- Étape 4 : Calcul du score de propension à la réadmission à l'hôpital

Étape 1 : Prédire les résultats des patients à l'aide d'un graphe de connaissances médicales

Dans Amazon Neptune, vous pouvez utiliser un graphe de connaissances pour stocker des informations temporelles sur les visites des patients et les résultats obtenus au fil du temps. Le moyen le plus efficace de créer et de stocker un graphe de connaissances consiste à utiliser un modèle de graphe et une base de données de graphes. Les bases de données de graphes sont spécialement conçues pour stocker et parcourir les relations. Les bases de données graphiques facilitent la modélisation et la gestion de données hautement connectées et disposent de schémas flexibles.

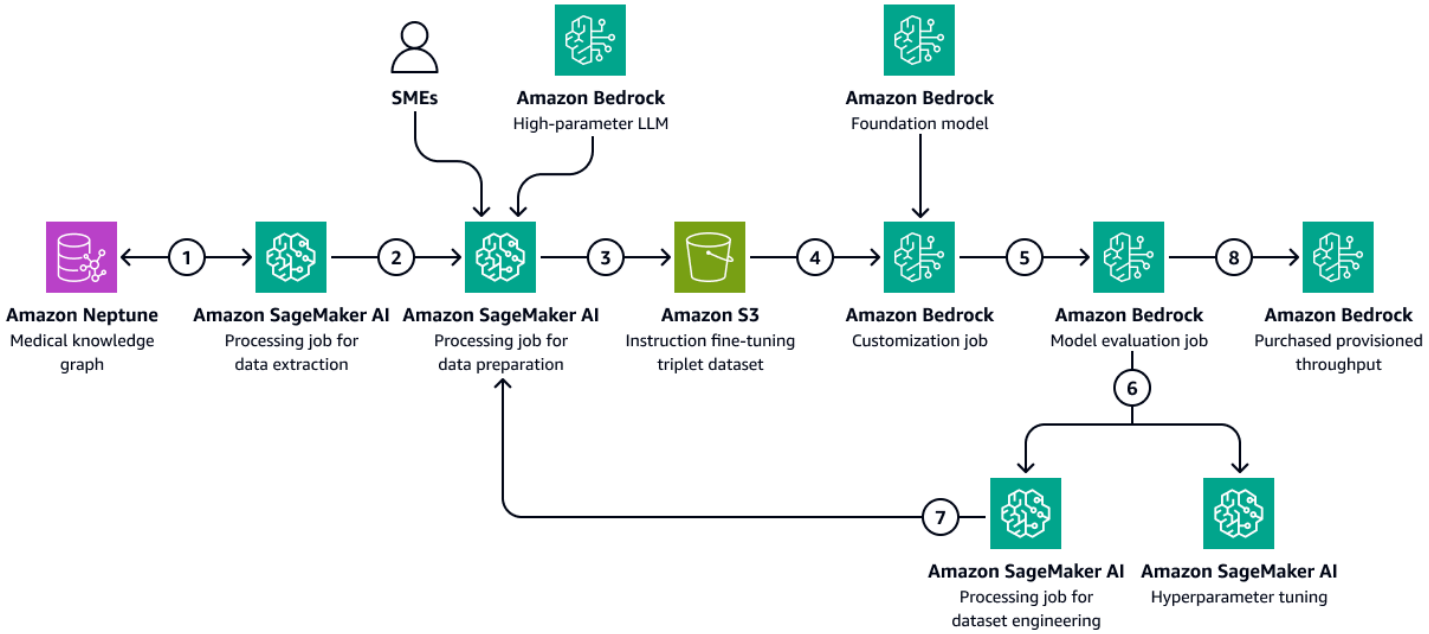
Le graphe de connaissances vous permet d'effectuer une analyse de séries chronologiques. Les éléments clés de la base de données de graphes utilisés pour la prédiction temporelle des résultats pour les patients sont les suivants :

- Données historiques — Diagnostics antérieurs, traitement continu, médicaments déjà utilisés et résultats de laboratoire pour le patient
- Visites des patients (chronologique) — Dates des visites, symptômes, allergies observées, notes cliniques, diagnostics, procédures, traitements, médicaments prescrits et résultats de laboratoire
- Symptômes et paramètres cliniques — Informations cliniques et basées sur les symptômes, notamment la gravité, les schémas de progression et la réponse du patient au médicament

Vous pouvez utiliser les informations du graphique des connaissances médicales pour peaufiner un LLM dans Amazon Bedrock, tel que Llama 3. Vous affinez le LLM à l'aide de données séquentielles du patient concernant la réponse du patient à un ensemble de médicaments ou de traitements au fil du temps. Utilisez un ensemble de données étiqueté qui classe un ensemble de médicaments ou de traitements et de données d'interaction patient-clinique dans des catégories prédéfinies indiquant l'état de santé d'un patient. La détérioration de l'état de santé, l'amélioration ou la stabilité des progrès sont des exemples de ces catégories. Lorsque le clinicien saisit un nouveau contexte concernant le patient et ses symptômes, le LLM affiné peut utiliser les modèles de l'ensemble de données de formation afin de prédire l'issue potentielle du patient.

L'image suivante montre les étapes séquentielles nécessaires pour peaufiner un LLM dans Amazon Bedrock à l'aide d'un ensemble de données de formation spécifique au secteur de la santé. Ces

données peuvent inclure l'état de santé des patients et les réponses aux traitements au fil du temps. Cet ensemble de données de formation aiderait le modèle à faire des prédictions généralisées sur les résultats pour les patients.



Le schéma suivant illustre le flux de travail suivant :

1. La tâche d'extraction de données Amazon SageMaker AI interroge le graphe de connaissances pour récupérer des données chronologiques sur les réponses des différents patients à un ensemble de médicaments ou de traitements au fil du temps.
2. Le travail de préparation des données basé sur l' SageMaker IA intègre un programme de maîtrise en droit d'Amazon Bedrock et des contributions d'experts en la matière ()SMEs. Le travail classe les données extraites du graphe de connaissances dans des catégories prédéfinies (telles que détérioration de l'état de santé, amélioration ou progrès stables) qui indiquent l'état de santé de chaque patient.
3. Le travail crée un ensemble de données de réglage précis qui inclut les informations extraites du graphe de connaissances, des chain-of-thought instructions et de la catégorie de résultats pour le patient. Il télécharge cet ensemble de données de formation dans un compartiment Amazon S3.
4. Une tâche de personnalisation d'Amazon Bedrock utilise cet ensemble de données de formation pour peaufiner un LLM.
5. Le travail de personnalisation d'Amazon Bedrock intègre le modèle fondamental de choix d'Amazon Bedrock dans l'environnement de formation. Il démarre le travail de réglage fin et utilise le jeu de données d'entraînement et les hyperparamètres d'entraînement que vous configurez.

6. Une tâche d'évaluation Amazon Bedrock évalue le modèle affiné à l'aide d'un cadre d'évaluation de modèle prédéfini.
7. Si le modèle doit être amélioré, la tâche de formation s'exécute à nouveau avec davantage de données après un examen attentif de l'ensemble de données d'apprentissage. Si le modèle ne montre pas d'amélioration progressive des performances, envisagez également de modifier les hyperparamètres d'entraînement.
8. Une fois que l'évaluation du modèle répond aux normes définies par les parties prenantes de l'entreprise, vous hébergez le modèle affiné sur le débit provisionné par Amazon Bedrock.

Étape 2 : Prédire le comportement du patient à l'égard des médicaments ou traitements prescrits

FineTuned LLMs peut traiter les notes cliniques, les résumés des sorties et d'autres documents spécifiques au patient à partir du graphe temporel des connaissances médicales. Ils peuvent évaluer si le patient est susceptible de suivre les médicaments ou traitements prescrits.

Cette étape utilise le graphe de connaissances créé dansÉtape 1 : Prédire les résultats des patients à l'aide d'un graphe de connaissances médicales. Le graphe de connaissances contient des données issues du profil du patient, y compris l'historique d'adhésion du patient en tant que nœud. Cela inclut également les cas de non-observance des médicaments ou des traitements, les effets secondaires des médicaments, le manque d'accès aux médicaments ou les obstacles liés au coût de ceux-ci, ou les schémas posologiques complexes comme caractéristiques de ces nœuds.

Finetuned LLMs peut utiliser les anciennes données d'exécution des ordonnances issues du graphe des connaissances médicales et les résumés descriptifs des notes cliniques provenant d'une base de données vectorielle Amazon OpenSearch Service. Ces notes cliniques peuvent mentionner des rendez-vous fréquemment manqués ou le non-respect des traitements. Le LLM peut utiliser ces notes pour prédire la probabilité d'une future non-conformité.

1. Préparez les données d'entrée comme suit :
   - Données structurées — Extrayez les données récentes des patients, telles que les trois dernières visites et les résultats de laboratoire, à partir du graphe des connaissances médicales.
   - Données non structurées : récupérez les notes cliniques récentes de la base de données vectorielle Amazon OpenSearch Service.

2. Créez une invite de saisie qui inclut les antécédents du patient et le contexte actuel. Voici un exemple d'invite :

```
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. Passez l'invite au LLM peaufiné. Le LLM traite l'invite et prédit le résultat. Voici un exemple de réponse du LLM :

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

4. Analysez la réponse du modèle pour extraire la catégorie de résultats prédite. Par exemple, la catégorie de l'exemple de réponse de l'étape précédente peut être une forte probabilité de non-conformité.
5. (Facultatif) Utilisez des logits modèles ou des méthodes supplémentaires pour attribuer des scores de confiance. Les logits sont les probabilités non normalisées que l'objet appartienne à une certaine classe ou catégorie.

Étape 3 : Prédire la probabilité de réadmission du patient

Les réadmissions à l'hôpital sont une préoccupation majeure en raison du coût élevé de l'administration des soins de santé et de leur impact sur le bien-être du patient. Le calcul des taux de réadmission dans les hôpitaux est un moyen de mesurer la qualité des soins aux patients et les performances d'un professionnel de santé.

Pour calculer le taux de réadmission, vous avez défini un indicateur, tel qu'un taux de réadmission sur 7 jours. Cet indicateur est le pourcentage de patients admis qui retournent à l'hôpital pour une visite imprévue dans les sept jours suivant leur sortie. Pour prévoir les chances de réadmission d'un patient, un LLM affiné peut utiliser les données temporelles du graphe de connaissances médicales que vous avez créé dans. Étape 1 : Prédire les résultats des patients à l'aide d'un graphe de connaissances médicales Ce graphique de connaissances conserve des enregistrements chronologiques des rencontres avec les patients, des procédures, des médicaments et des symptômes. Ces enregistrements de données contiennent les éléments suivants :

- Durée écoulée depuis la dernière sortie du patient
- La réponse du patient aux traitements et médicaments antérieurs
- L'évolution des symptômes ou des affections au fil du temps

Vous pouvez traiter ces événements chronologiques pour prédire la probabilité de réadmission d'un patient grâce à une invite du système organisée. L'invite transmet la logique de prédiction au LLM affiné.

1. Préparez les données d'entrée comme suit :
   - Historique d'observance — Extrayez les dates de prise des médicaments, les fréquences de renouvellement des médicaments, les détails du diagnostic et des médicaments, les antécédents médicaux chronologiques et d'autres informations du graphique des connaissances médicales.
   - Indicateurs comportementaux — Récupérez et incluez les notes cliniques concernant les rendez-vous manqués et les effets secondaires signalés par les patients.
2. Créez une invite de saisie qui inclut l'historique d'adhésion et les indicateurs comportementaux. Voici un exemple d'invite :

```
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. Passez l'invite au LLM peaufiné. Le LLM traite l'invite et prédit la probabilité et les raisons de la réadmission. Voici un exemple de réponse du LLM :

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

4. Catégorisez la prédiction selon une échelle standardisée, telle que faible, moyenne ou élevée.
```

5. Passez en revue le raisonnement fourni par le LLM et identifiez les facteurs clés qui contribuent à la prédiction.
6. Associez les résultats qualitatifs aux scores quantitatifs. Par exemple, une probabilité très élevée peut être égale à 0,9.
7. Utilisez des ensembles de données de validation pour calibrer les résultats du modèle par rapport aux taux de réadmission réels.

Étape 4 : Calcul du score de propension à la réadmission à l'hôpital

Ensuite, vous calculez un score de propension à la réadmission à l'hôpital par patient. Ce score reflète l'impact net des trois analyses effectuées au cours des étapes précédentes : les résultats potentiels pour les patients, le comportement du patient à l'égard des médicaments et des traitements, et la probabilité de réadmission du patient. En agrégeant le score de propension à la réadmission du patient au niveau de la spécialité, puis au niveau de l'hôpital, vous pouvez obtenir des informations utiles aux cliniciens, aux responsables des soins et aux administrateurs. Le score de propension à la réadmission à l'hôpital vous permet d'évaluer le rendement global par établissement, par spécialité ou par affection. Vous pouvez ensuite utiliser ce score pour mettre en œuvre des interventions proactives.

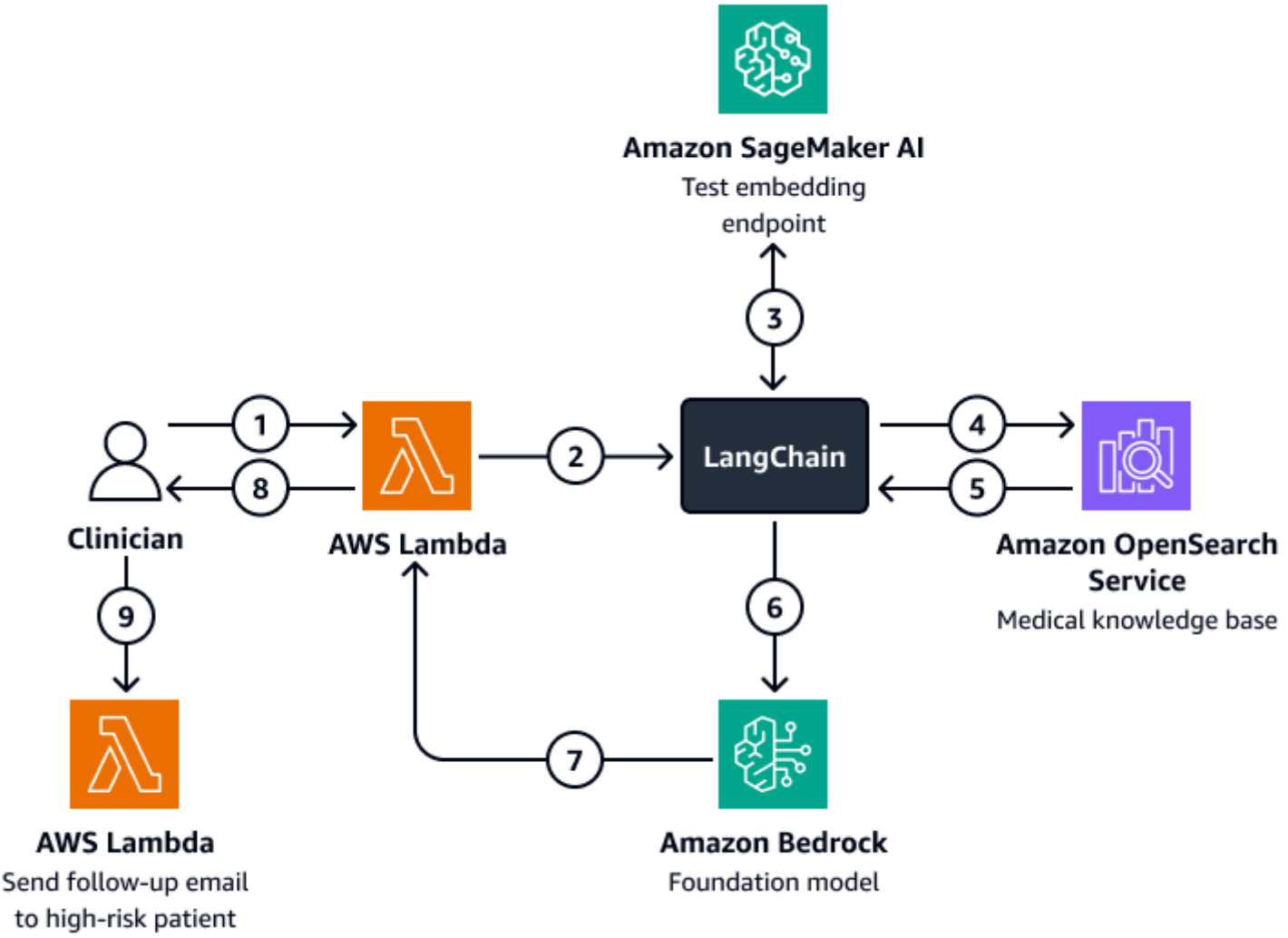
1. Attribuez des poids à chacun des différents facteurs (prédiction des résultats, probabilité d'adhésion, réadmission). Voici des exemples de poids :
   - Poids de prédiction des résultats : 0,4
   - Poids prévisionnel d'adhérence : 0,3
   - Poids de la probabilité de réadmission : 0,3
2. Utilisez le calcul suivant pour calculer le score composite :

   `ReadadmissionPropensityScore` = (`OutcomeScore` × `OutcomeWeight`) + (`AdherenceScore` × `AdherenceWeight`) + (`ReadmissionLikelihoodScore` × `ReadmissionLikelihoodWeight`)
3. Assurez-vous que tous les scores individuels sont sur la même échelle, par exemple de 0 à 1.
4. Définissez les seuils d'action. Par exemple, les scores supérieurs à 0,7 déclenchent des alertes.

Sur la base des analyses ci-dessus et du score de propension à la réadmission d'un patient, les cliniciens ou les responsables de soins peuvent configurer des alertes pour surveiller leurs patients

individuels sur la base du score calculé. S'il est supérieur à un seuil prédéfini, ils sont avertis lorsque ce seuil est atteint. Cela permet aux responsables des soins d'être proactifs plutôt que réactifs lorsqu'ils élaborent des plans de soins pour la sortie de leurs patients. Enregistrez les résultats, le comportement et les scores de propension des patients sous forme indexée dans une base de données vectorielle Amazon OpenSearch Service afin que les responsables de soins puissent les récupérer facilement à l'aide d'un agent d'intelligence artificielle conversationnel.

Le schéma suivant montre le flux de travail d'un agent d'IA conversationnel qu'un clinicien ou un responsable de soins peut utiliser pour obtenir des informations sur les résultats des patients, le comportement attendu et la propension à la réadmission. Les utilisateurs peuvent récupérer des informations au niveau du patient, du département ou de l'hôpital. L'agent AI récupère ces informations, qui sont stockées sous forme indexée dans une base de données vectorielle Amazon OpenSearch Service. L'agent utilise la requête pour récupérer les données pertinentes et fournit des réponses personnalisées, y compris des suggestions d'actions pour les patients présentant un risque élevé de réadmission. En fonction du niveau de risque, l'agent peut également configurer des rappels pour les patients et les soignants.



Le schéma suivant illustre le flux de travail suivant :

1. Le clinicien pose une question à un agent d'intelligence artificielle conversationnel qui héberge une AWS Lambda fonction.
2. La fonction Lambda lance un LangChain agent.
3. Le LangChain l'agent envoie la question de l'utilisateur à un point de terminaison d'intégration de texte Amazon SageMaker AI. Le point de terminaison intègre la question.
4. Le LangChain l'agent transmet la question intégrée à une base de connaissances médicales sur Amazon OpenSearch Service.
5. Amazon OpenSearch Service renvoie les informations spécifiques les plus pertinentes pour la requête de l'utilisateur au LangChain agent.
6. Le LangChain les agents envoient la requête et le contexte extrait de la base de connaissances à un modèle de base Amazon Bedrock.

7. Le modèle Amazon Bedrock Foundation génère une réponse et l'envoie à la fonction Lambda.
8. La fonction Lambda renvoie la réponse au clinicien.
9. Le clinicien lance une fonction Lambda qui envoie un e-mail de suivi à un patient présentant un risque élevé de réadmission.

Alignement sur le framework AWS Well-Architected

L'architecture permettant de suivre le comportement des patients et de prévoir les taux de réadmission à l'hôpital intègre Services AWS des graphiques de connaissances médicales et LLMs permet d'améliorer les résultats des soins de santé tout en s'alignant sur les six piliers du Well-Architected AWS Framework :

- Excellence opérationnelle — La solution est un système découplé et automatisé qui utilise Amazon Bedrock et AWS Lambda fournit des alertes en temps réel.
- Sécurité — Cette solution est conçue pour se conformer aux réglementations sanitaires, telles que la loi HIPAA. Vous pouvez également implémenter le chiffrement, un contrôle d'accès précis et des garde-corps Amazon Bedrock pour protéger les données des patients.
- Fiabilité — L'architecture utilise un système sans serveur tolérant aux pannes. Services AWS
- Efficacité des performances — Amazon OpenSearch Service et le service affiné LLMs peuvent fournir des prévisions rapides et précises.
- Optimisation des coûts — Les technologies et les pay-per-inference modèles sans serveur permettent de minimiser les coûts. Bien qu'un LLM affiné puisse entraîner des frais supplémentaires, le modèle utilise une approche RAG qui réduit les données et le temps de calcul nécessaires au processus de réglage précis.
- Durabilité — L'architecture minimise la consommation de ressources grâce à l'utilisation d'une infrastructure sans serveur. Il soutient également des opérations de santé efficaces et évolutives.

Cas d'utilisation : gestion et renforcement des compétences de votre personnel de santé

La mise en œuvre de stratégies de transformation des talents et de renforcement des compétences aide les travailleurs à rester aptes à utiliser les nouvelles technologies et pratiques dans les services médicaux et de santé. Les initiatives proactives de renforcement des compétences permettent aux professionnels de santé de fournir des soins de haute qualité aux patients, d'optimiser l'efficacité opérationnelle et de respecter les normes réglementaires. De plus, la transformation des talents favorise une culture d'apprentissage continu. Cela est essentiel pour s'adapter à l'évolution du paysage des soins de santé et relever les nouveaux défis de santé publique. Les approches de formation traditionnelles, telles que la formation en classe et les modules d'apprentissage statiques, offrent un contenu uniforme à un large public. Ils ne disposent souvent pas de parcours d'apprentissage personnalisés, essentiels pour répondre aux besoins spécifiques et aux niveaux de compétence de chaque praticien. Cette one-size-fits-all stratégie peut entraîner un désengagement et une rétention des connaissances sous-optimale.

Par conséquent, les établissements de santé doivent adopter des solutions innovantes, évolutives et axées sur la technologie qui peuvent déterminer l'écart entre la situation actuelle et la situation future de chacun de leurs employés. Ces solutions devraient recommander des parcours d'apprentissage hyperpersonnalisés et le bon ensemble de contenus d'apprentissage. Cela prépare efficacement le personnel au futur des soins de santé.

Dans le secteur de la santé, vous pouvez appliquer l'IA générative pour vous aider à comprendre et à améliorer les compétences de votre personnel. En connectant de grands modèles linguistiques (LLMs) et des outils de recherche avancés, les organisations peuvent comprendre les compétences dont elles disposent actuellement et identifier les compétences clés qui pourraient être nécessaires à l'avenir. Ces informations vous aident à combler le fossé en recrutant de nouveaux travailleurs et en améliorant les compétences de la main-d'œuvre actuelle. À l'aide d'Amazon Bedrock et des graphes de connaissances, les établissements de santé peuvent développer des applications spécifiques à un domaine qui facilitent l'apprentissage continu et le développement des compétences.

Les connaissances fournies par cette solution vous aident à gérer efficacement les talents, à optimiser les performances du personnel, à favoriser le succès de l'organisation, à identifier les compétences existantes et à élaborer une stratégie de gestion des talents. Cette solution peut vous aider à effectuer ces tâches en quelques semaines au lieu de plusieurs mois.

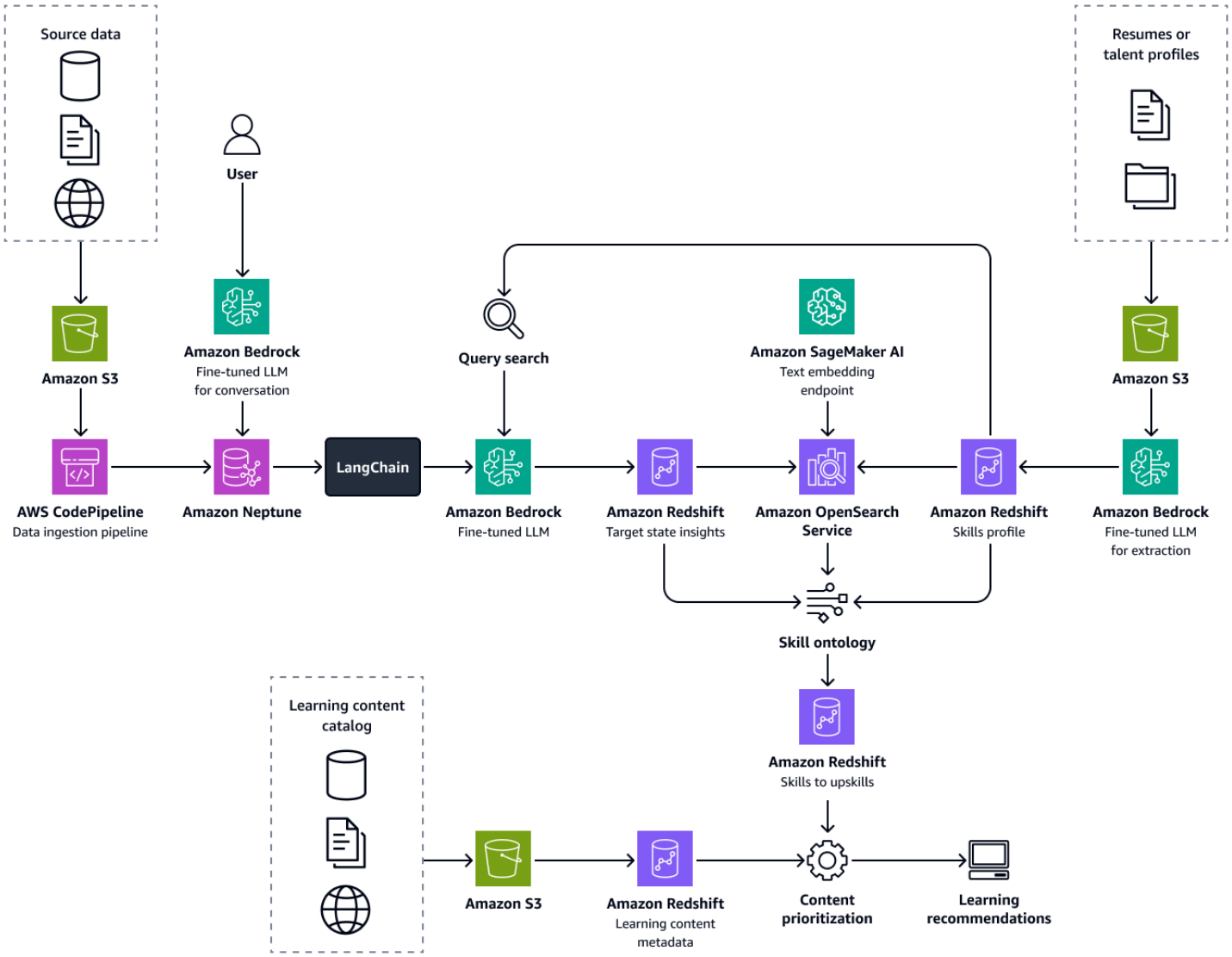
Présentation de la solution

Cette solution est un cadre de transformation des talents du secteur de la santé qui comprend les éléments suivants :

- Analyseur de CV intelligent — Ce composant peut lire le CV d'un candidat et extraire avec précision les informations relatives au candidat, y compris ses compétences. Solution intelligente d'extraction d'informations conçue à l'aide du modèle Llama 2 affiné d'Amazon Bedrock sur un ensemble de données de formation exclusif couvrant les CV et les profils de talents de plus de 19 secteurs d'activité. Ce processus basé sur le LLM permet d'économiser des centaines d'heures en automatisant le processus de révision manuelle des CV et en faisant correspondre les meilleurs candidats aux postes vacants.
- Graphe de connaissances : graphe de connaissances basé sur Amazon Neptune, un référentiel unifié d'informations sur les talents, y compris la taxonomie des rôles et des compétences de l'organisation et du secteur, qui capture la sémantique des talents du secteur de la santé à l'aide de définitions des compétences, des rôles et de leurs propriétés, des relations et des contraintes logiques.
- Ontologie des compétences — La découverte de la proximité des compétences entre les compétences des candidats et les compétences de l'état actuel ou futur idéales (récupérées à l'aide d'un graphe de connaissances) est réalisée grâce à des algorithmes d'ontologie qui mesurent la similitude sémantique entre les compétences des candidats et les compétences de l'état cible.
- Parcours et contenu d'apprentissage — Ce composant est un moteur de recommandation pédagogique capable de recommander le bon contenu d'apprentissage à partir d'un catalogue de supports d'apprentissage de n'importe quel fournisseur en fonction des lacunes de compétences identifiées. Identifier les parcours de perfectionnement les plus optimaux pour chaque candidat en analysant les lacunes en matière de compétences et en recommandant des contenus d'apprentissage prioritaires, afin de permettre un développement professionnel continu et fluide pour chaque candidat pendant la transition vers un nouveau poste.

Cette solution automatisée basée sur le cloud est alimentée par des services d'apprentissage automatique LLMs, des graphes de connaissances et la génération augmentée de récupération (RAG). Il peut évoluer pour traiter des dizaines ou des milliers de CV en un minimum de temps, créer des profils de candidats instantanés, identifier les lacunes dans leur état actuel ou futur potentiel, puis recommander efficacement le contenu de formation approprié pour combler ces lacunes.

L'image suivante montre le end-to-end flux du framework. La solution est basée sur Amazon Bedrock affiné LLMs . Ils LLMs extraient des données de la base de connaissances sur les talents du secteur de la santé d'Amazon Neptune. Les algorithmes basés sur les données fournissent des recommandations pour des parcours d'apprentissage optimaux pour chaque candidat.



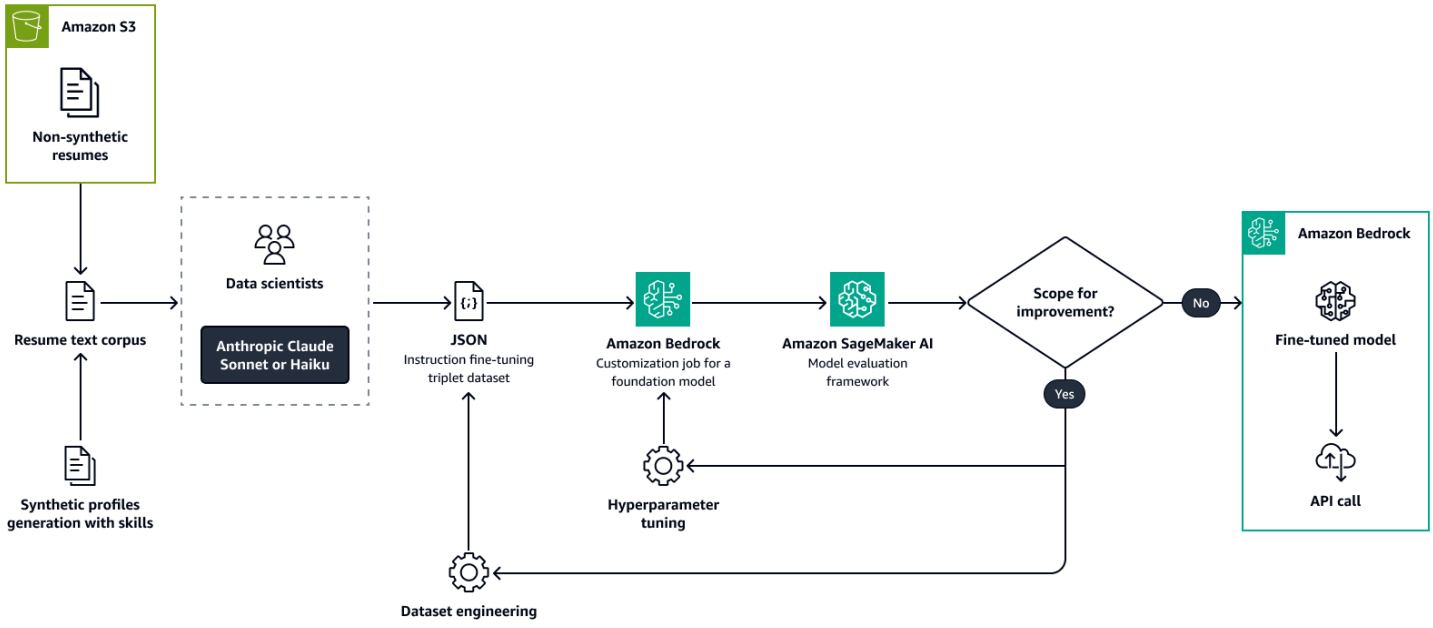
La création de cette solution comprend les étapes suivantes :

- Étape 1 : Extraction des informations sur les talents et établissement d'un profil de compétences
- Étape 2 : Découverte de la role-to-skill pertinence à partir d'un graphe de connaissances
- Étape 3 : Identifier les lacunes en matière de compétences et recommander une formation

Étape 1 : Extraction des informations sur les talents et établissement d'un profil de compétences

Tout d'abord, vous peaufinez un grand modèle de langage, tel que Llama 2, dans Amazon Bedrock à l'aide d'un ensemble de données personnalisé. Cela adapte le LLM au cas d'utilisation. Au cours de la formation, vous extrayez de manière précise et cohérente les principaux attributs des talents des CV des candidats ou de profils de talents similaires. Ces attributs de talent incluent les compétences, le titre du poste actuel, les titres d'expérience avec dates, la formation et les certifications. Pour plus d'informations, consultez la section Personnaliser votre modèle afin d'améliorer ses performances pour votre cas d'utilisation dans la documentation Amazon Bedrock.

L'image suivante montre le processus permettant de peaufiner un modèle d'analyse de CV à l'aide d'Amazon Bedrock. Les CV réels et synthétiques sont transmis à un LLM afin d'en extraire des informations clés. Un groupe de data scientists valide les informations extraites par rapport au texte brut d'origine. Les informations extraites sont ensuite concaténées à l'aide d'chain-of-thoughtinstructions et du texte original pour obtenir un ensemble de données d'apprentissage à affiner. Ce jeu de données est ensuite transmis à une tâche de personnalisation Amazon Bedrock, qui affine le modèle. Une tâche par lots Amazon SageMaker AI exécute un cadre d'évaluation de modèle qui évalue le modèle affiné. Si le modèle doit être amélioré, la tâche est réexécutée avec davantage de données ou des hyperparamètres différents. Une fois que l'évaluation répond aux normes, vous hébergez le modèle personnalisé via le débit provisionné d'Amazon Bedrock.



Étape 2 : Découverte de la role-to-skill pertinence à partir d'un graphe de connaissances

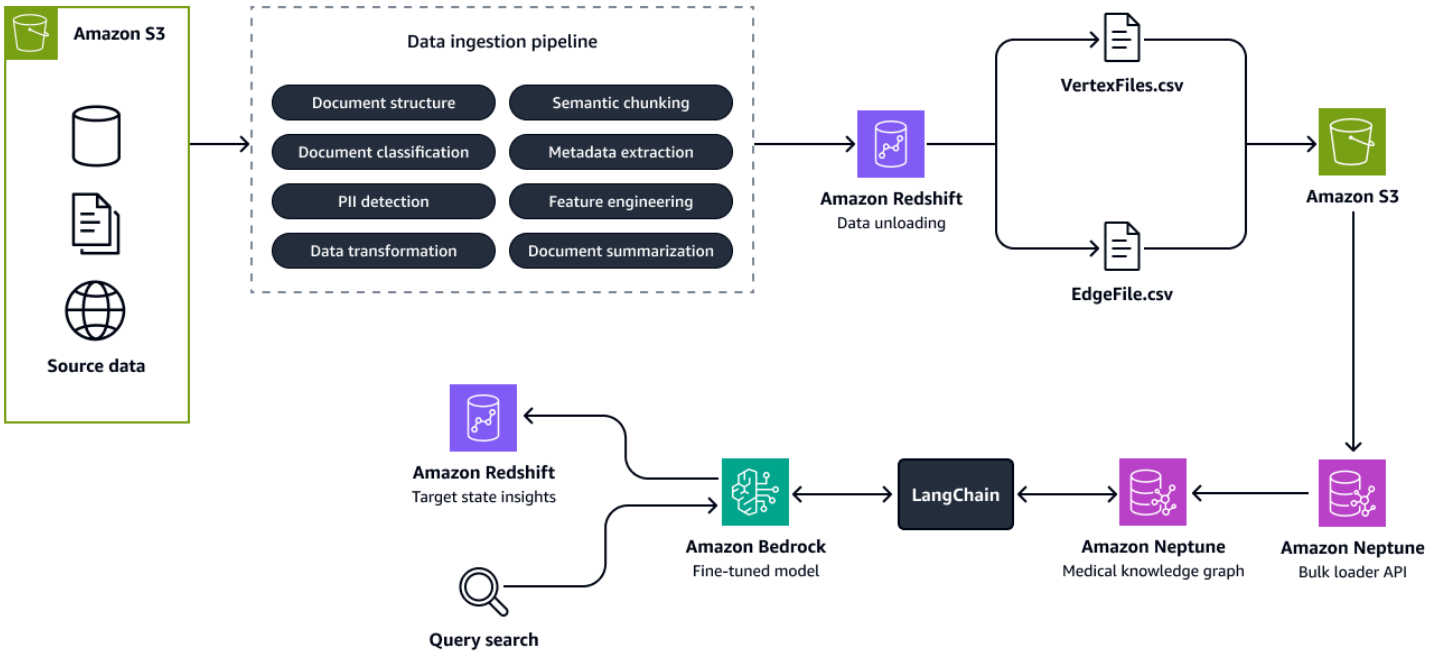
Ensuite, vous créez un graphe de connaissances qui résume les compétences et la taxonomie des rôles de votre organisation et des autres organisations du secteur de la santé. Cette base de connaissances enrichie provient de données agrégées sur les talents et les organisations dans Amazon Redshift. Vous pouvez recueillir des données sur les talents auprès de divers fournisseurs de données sur le marché du travail et de sources de données structurées et non structurées spécifiques à l'organisation, telles que les systèmes de planification des ressources d'entreprise (ERP), les systèmes d'information sur les ressources humaines (HRIS), les CV des employés, les descriptions de poste et les documents relatifs à l'architecture des talents.

Créez le graphe de connaissances sur Amazon Neptune. Les nœuds représentent les compétences et les rôles, tandis que les arêtes représentent les relations entre eux. Enrichissez ce graphique avec des métadonnées pour inclure des détails tels que le nom de l'organisation, le secteur d'activité, la famille d'emplois, le type de compétence, le type de rôle et les tags sectoriels.

Ensuite, vous développez une application Graph Retrieval Augmented Generation (Graph RAG). Graph RAG est une approche RAG qui récupère les données d'une base de données de graphes. Les composants de l'application Graph RAG sont les suivants :

- Intégration avec un LLM dans Amazon Bedrock — L'application utilise un LLM dans Amazon Bedrock pour comprendre le langage naturel et générer des requêtes. Les utilisateurs peuvent interagir avec le système en utilisant le langage naturel. Cela le rend accessible aux parties prenantes non techniques.
- Orchestration et récupération d'informations — Utilisation ou LlamaIndexLangChaindes orchestrateurs pour faciliter l'intégration entre le LLM et le graphe de connaissances Neptune. Ils gèrent le processus de conversion des requêtes en langage naturel en requêtes OpenCypher. Ils exécutent ensuite les requêtes sur le graphe de connaissances. Utilisez une ingénierie rapide pour informer le LLM des meilleures pratiques pour créer des requêtes OpenCypher. Cela permet d'optimiser les requêtes pour récupérer le sous-graphe pertinent, qui contient toutes les entités et relations pertinentes concernant les rôles et les compétences demandés.
- Génération d'informations — Le LLM d'Amazon Bedrock traite les données graphiques récupérées. Il génère des informations détaillées sur l'état actuel et projette les états futurs pour le rôle demandé et les compétences associées.

L'image suivante montre les étapes à suivre pour créer un graphe de connaissances à partir des données sources. Vous transmettez les données sources structurées et non structurées au pipeline d'ingestion de données. Le pipeline extrait et transforme les informations en une formation de chargement groupé CSV compatible avec Amazon Neptune. L'API Bulk Loader télécharge les fichiers CSV stockés dans un compartiment Amazon S3 vers le graphe de connaissances Neptune. Pour les requêtes des utilisateurs relatives aux talents, à l'état futur, aux rôles pertinents ou aux compétences, le LLM affiné d'Amazon Bedrock interagit avec le graphe de connaissances via un LangChain orchestrateur. L'orchestrateur extrait le contexte pertinent à partir du graphe de connaissances et transmet les réponses au tableau d'informations d'Amazon Redshift. Le LangChain Un orchestrateur, comme Graph QAChain, convertit la requête en langage naturel de l'utilisateur en une requête OpenCypher afin d'interroger le graphe de connaissances. Le modèle affiné d'Amazon Bedrock génère une réponse basée sur le contexte extrait.



Étape 3 : Identifier les lacunes en matière de compétences et recommander une formation

Au cours de cette étape, vous calculez avec précision la proximité entre l'état actuel d'un professionnel de santé et les futurs rôles gouvernementaux potentiels. Pour ce faire, vous effectuez une analyse d'affinité des compétences en comparant les compétences de l'individu avec le rôle professionnel. Dans une base de données vectorielle Amazon OpenSearch Service, vous stockez les informations de taxonomie des compétences et les métadonnées des compétences, telles que

la description des compétences, le type de compétence et les groupes de compétences. Utilisez un modèle d'intégration Amazon Bedrock, tel que les modèles Amazon Titan Text Embeddings, pour intégrer la compétence clé identifiée dans des vecteurs. Grâce à une recherche vectorielle, vous pouvez récupérer les descriptions des compétences de l'état actuel et des compétences de l'état cible et effectuer une analyse ontologique. L'analyse fournit des scores de proximité entre les paires de compétences de l'État actuel et de l'État cible. Pour chaque paire, vous utilisez les scores ontologiques calculés pour identifier les écarts dans les affinités de compétences. Ensuite, vous recommandez la voie optimale pour le renforcement des compétences, que le candidat peut envisager lors des transitions de rôle.

Pour chaque rôle, la recommandation du contenu d'apprentissage approprié à des fins de perfectionnement ou de requalification implique une approche systématique qui commence par la création d'un catalogue complet de contenus d'apprentissage. Ce catalogue, que vous stockez dans une base de données Amazon Redshift, regroupe le contenu de différents fournisseurs et inclut des métadonnées, telles que la durée du contenu, le niveau de difficulté et le mode d'apprentissage. L'étape suivante consiste à extraire les compétences clés proposées par chaque élément de contenu, puis à les associer aux compétences individuelles requises pour le rôle cible. Vous réalisez cette cartographie en analysant la couverture fournie par le contenu grâce à une analyse de proximité des compétences. Cette analyse évalue dans quelle mesure les compétences enseignées par le contenu correspondent aux compétences souhaitées pour le poste. Les métadonnées jouent un rôle essentiel dans la sélection du contenu le plus approprié pour chaque compétence, en veillant à ce que les apprenants reçoivent des recommandations personnalisées adaptées à leurs besoins d'apprentissage. LLMs Utilisez-le dans Amazon Bedrock pour extraire des compétences des métadonnées du contenu, effectuer l'ingénierie des fonctionnalités et valider les recommandations de contenu. Cela améliore la précision et la pertinence du processus de mise à niveau ou de requalification.

Alignement sur le framework AWS Well-Architected

La solution s'aligne sur les six piliers du AWS Well-Architected Framework :

- Excellence opérationnelle — Un pipeline modulaire et automatisé améliore l'excellence opérationnelle. Les composants clés du pipeline sont découplés et automatisés, ce qui permet d'accélérer les mises à jour des modèles et de faciliter la surveillance. De plus, les pipelines de formation automatisés permettent de publier plus rapidement des modèles affinés.
- Sécurité — Cette solution traite les informations sensibles et personnellement identifiables (PII), telles que les données figurant dans les CV et les profils de talents. Dans AWS Identity and Access

Management (IAM), mettez en œuvre des politiques de contrôle d'accès précises et assurez-vous que seul le personnel autorisé a accès à ces données.

- Fiabilité — La solution utilise des Services AWS outils tels que Neptune, Amazon Bedrock et OpenSearch Service, qui offrent une tolérance aux pannes, une haute disponibilité et un accès ininterrompu aux informations, même en cas de forte demande.
- Efficacité des performances : optimisées LLMs dans Amazon Bedrock et OpenSearch Service, les bases de données vectorielles sont conçues pour traiter rapidement et avec précision de grands ensembles de données afin de fournir des recommandations d'apprentissage personnalisées en temps opportun.
- Optimisation des coûts — Cette solution utilise une approche RAG, qui réduit le besoin de pré-entraînement continu des modèles. Au lieu de peaufiner l'ensemble du modèle à plusieurs reprises, le système ne peaufine que des processus spécifiques, tels que l'extraction d'informations à partir de CV et la structuration des résultats. Cela se traduit par des économies de coûts importantes. En minimisant la fréquence et l'ampleur de la formation des modèles gourmands en ressources et en utilisant les services pay-per-use cloud, les établissements de santé peuvent optimiser leurs coûts opérationnels tout en maintenant des performances élevées.
- Durabilité — Cette solution utilise des services évolutifs et natifs du cloud qui allouent les ressources informatiques de manière dynamique. Cela permet de réduire la consommation d'énergie et l'impact environnemental tout en soutenant les initiatives de transformation des talents à grande échelle et gourmandes en données.

Développement et orchestration de solutions d'IA générative pour le secteur de la santé

Pour créer les solutions présentées dans ce guide, vous devez créer une architecture RAG qui utilise des outils optimisés LLMs pour fournir aux prestataires de soins de santé des données augmentées sur les patients, des informations cliniques et diagnostiques et des résultats prédits pour les patients. Cela nécessite l'intégration de plusieurs Services AWS outils pour créer un flux de travail cohérent et efficace. Cette section aborde les points suivants :

- Amazon Q Developer— Utilisez Amazon Q Developer pour résoudre les questions d'ingénierie et les erreurs de code pendant le processus de développement.
- Design RAG à plusieurs récupérateurs— Concevez et implémentez des solutions RAG qui utilisent plusieurs récupérateurs pour trouver le contexte médical approprié à la question de l'utilisateur.
- ReAct agents— Implémentez des agents qui combinent raisonnement et action dynamique.

Amazon Q Developer

Lors de la création d'une solution d'IA générative, il peut être difficile de créer des agents d'IA et de connecter des services clés. Cependant, Amazon Q Developer aide les scientifiques des données et les ingénieurs en IA en leur donnant accès à un assistant d'IA générative avancé. Amazon Q peut répondre rapidement et précisément aux questions des utilisateurs et aux erreurs de code, ce qui peut vous aider à optimiser le processus de développement du LLM. Amazon Q offre des avantages considérables aux développeurs qui créent des applications utilisant les modèles de base d'Amazon Bedrock. Il permet de rationaliser les flux de travail et d'améliorer la qualité du code. Il automatise la génération de scripts Python et de configurations d'infrastructure sous forme de code (IaC), réduisant ainsi considérablement le temps et les efforts de développement. Grâce à des fonctionnalités de refactoring avancées, Amazon Q peut améliorer les performances du code, identifier les failles de sécurité et s'assurer que les développeurs respectent les meilleures pratiques. En outre, il facilite l'apprentissage et l'adoption par les débutants en fournissant des suggestions et des explications contextuelles, ce qui rend les tâches de codage complexes plus accessibles et plus efficaces.

Design RAG à plusieurs récupérateurs

Dans une application d'IA générative, un pipeline RAG multi-récupérateurs peut récupérer efficacement des informations à partir de plusieurs sources de données pour aider les prestataires de soins de santé et les cliniciens à répondre aux questions médicales. Ce pipeline utilise différents types de récupérateurs pour extraire les données pertinentes de différentes bases de connaissances. Chaque récupérateur est spécialisé dans la récupération d'un type particulier d'informations, telles que les antécédents des patients, les informations diagnostiques, les notes cliniques ou le contenu de recherches médicales et de textes universitaires.

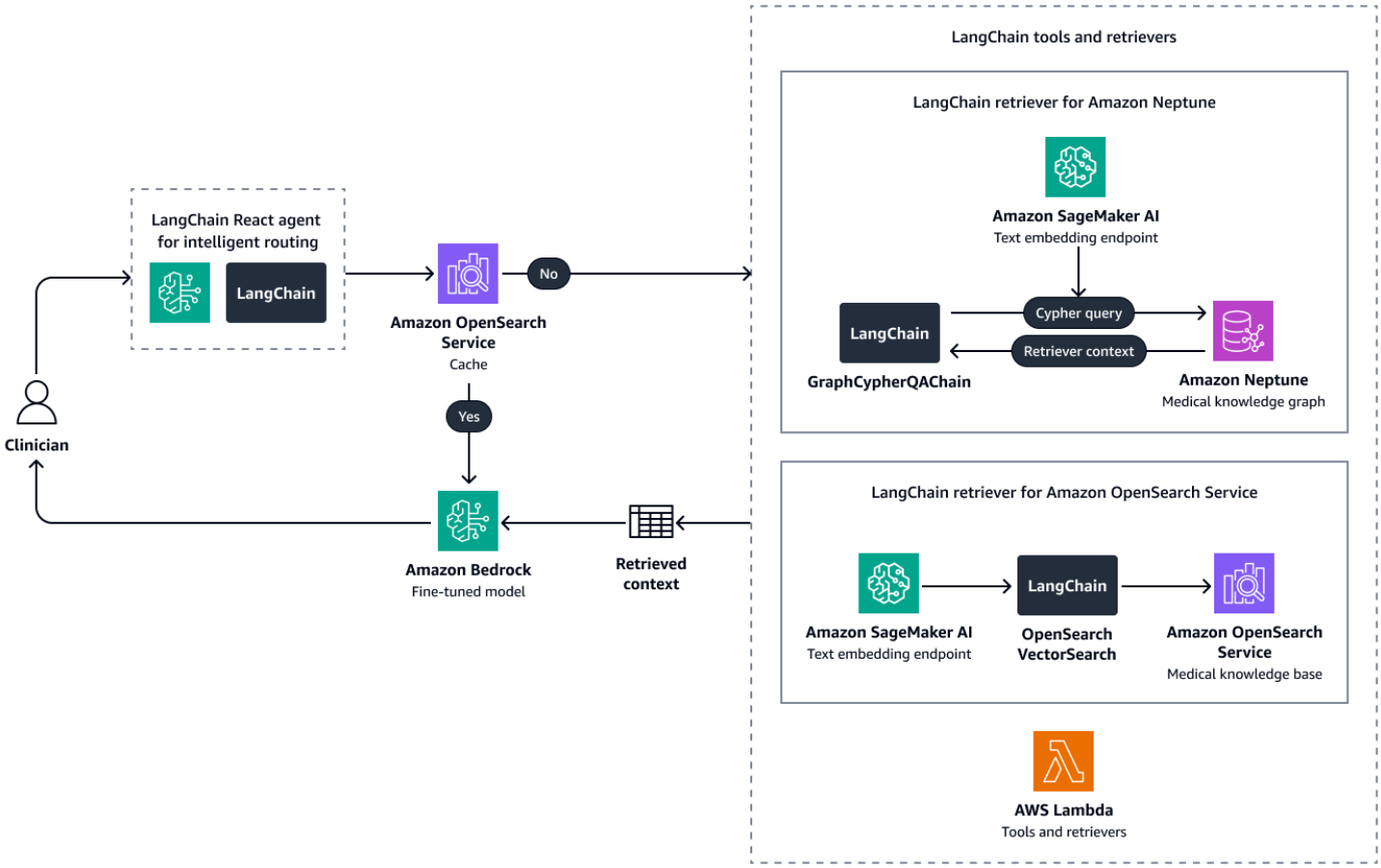
Utilisez la nature des données et les exigences spécifiques de l'application pour déterminer quelle est la base de connaissances backend adaptée à votre cas d'utilisation. Une base de données vectorielle Amazon OpenSearch Service convient parfaitement à de grands volumes de données de santé non structurées ou semi-structurées, notamment des résumés d'évaluation diagnostique par image, des résumés de sortie, des rapports cliniques, des recherches médicales et des textes universitaires. D'autre part, un service de base de données graphique, tel qu'Amazon Neptune, peut être idéal pour les cas d'utilisation dans le secteur de la santé qui nécessitent une exploration approfondie des relations temporelles entre les entités, telles que le patient, les antécédents du patient, le prestataire de soins de santé, les médicaments, les symptômes et les traitements.

Un élément essentiel de ce pipeline est la prédiction de l'intention des requêtes des utilisateurs. Cela permet de s'assurer que le système achemine la requête vers la bonne chaîne de récupération. Par exemple, si un clinicien pose des questions sur les antécédents thérapeutiques d'un patient, ses symptômes, son interaction avec l'hôpital, la probabilité d'une réadmission à l'hôpital ou les résultats potentiels du patient, le module de prédiction de l'intention des requêtes identifie cette intention. Il dirige la demande vers la chaîne de recherche qui peut récupérer les dossiers des patients ou les données chronologiques des traitements à partir du graphe des connaissances médicales. Par ailleurs, si la question porte sur la découverte d'une maladie, des évaluations diagnostiques spécifiques ou des détails de procédures cliniques spécifiques tirés de manuels universitaires, la requête est acheminée vers la chaîne de recherche qui peut récupérer ces informations dans la OpenSearch base de données des vecteurs de service. Vous pouvez utiliser la fonctionnalité d'appel d'outils de LangChain pour lier un outil personnalisé à l'Amazon Bedrock LLM qui permet de classer une question d'utilisateur selon des intentions prédéfinies.

Ce système RAG multi-retriever comprend LangChain agents conçus pour gérer l'accès à la base de connaissances spécifique. Vous pouvez utiliser … LangChain pour orchestrer l'interaction entre le LLM Amazon Bedrock, les différents retrievers et les outils. LangChain inclut une classe

d'appel d'outils qui vous aide à créer des outils personnalisés, tels qu'un classificateur d'intention, un récupérateur pour Neptune, un récupérateur OpenSearch pour Service ou tout autre outil pouvant être développé pour classer les intentions de l'utilisateur et accéder aux données d'une base de connaissances spécifique dans un format structuré. Vous transmettez ensuite ces outils à la classe pour créer un agent Reasoning and Acting (ReAct). L' ReAct agent traite la question de l'utilisateur, planifie les étapes séquentielles pour y répondre, puis exécute de manière itérative les outils disponibles et traite les réponses des outils pour enfin répondre à la requête de l'utilisateur.

L'image suivante montre le fonctionnement d'un système RAG à plusieurs récupérateurs conçu pour une récupération efficace des connaissances et une résolution intelligente des requêtes. A LangChain ReAct l'agent analyse l'intention de l'utilisateur, formule un plan d'exécution structuré et sélectionne les outils de récupération les plus pertinents. Le système interroge un cache de questions précédentes et recherche des requêtes similaires en fonction d'attributs clés, tels que l'identifiant du patient, son état de santé et la date de la visite. Si une question très similaire est trouvée, la réponse correspondante est récupérée directement. Dans le cas contraire, l'agent exécute le récupérateur approprié. Pour récupérer des informations centrées sur le patient, telles que l'historique du traitement, les symptômes, les interactions avec l'hôpital ou la probabilité de réadmission, le système utilise un extracteur de graphes. Pour les évaluations diagnostiques, les procédures cliniques et les résultats médicaux structurés, l'agent utilise un récupérateur de base de données vectorielles. Dans les scénarios qui nécessitent une combinaison de connaissances contextuelles provenant des deux magasins de données, pour générer une réponse complète, le système utilise une stratégie de récupération hybride qui intègre les résultats du graphe de connaissances et de la base de données vectorielle.



ReAct agents

Les agents Reasoning and Acting (ReAct) sont conçus pour les applications RAG à multiples facettes. Ces agents fournissent une puissante combinaison de raisonnement et d'action dynamique, en particulier pour les applications complexes qui impliquent des flux de step-by-step travail logiques de récupération d'informations. Pour plus d'informations, voir [ReAct: Synergie du raisonnement et de l'action dans les modèles linguistiques].

Dans les contextes médicaux et de soins de santé, les demandes d'un clinicien ou d'un médecin comportent souvent de multiples facettes. Par exemple, un clinicien peut demander « Quels traitements ont été administrés à des patients similaires souffrant à la fois d'hypertension et de diabète de type 2 ? » Après avoir identifié l'intention de l'utilisateur, qui est de récupérer les traitements pour l'hypertension et le diabète de type 2, l'agent d'intelligence artificielle doit diviser cette requête en sous-tâches, puis choisir la stratégie de récupération la plus efficace. Dans ce cas, l'agent d'IA doit identifier les nœuds les plus pertinents (tels que l'âge, le sexe, les affections, les traitements et les médicaments du patient), puis interroger le graphique pour ces entités,

leurs attributs et leurs relations. ReAct les agents sont très utiles car ils combinent la capacité de raisonnement (inférence logique) d'un LLM avec une action (interrogation ou interaction avec des ressources externes ou des bases de connaissances).

Pour répondre à la question de l'utilisateur « Quels traitements ont été administrés à des patients similaires souffrant à la fois d'hypertension et de diabète de type 2 ? » , l'exemple suivant illustre le fonctionnement d'un ReAct agent :

1. Raisonnement de l' ReAct agent — L'agent en déduit que la question implique de récupérer des informations sur des affections (diabète et hypertension). Il tient compte de l'âge du patient, des traitements, des médicaments et de la période à analyser.
2. Action de l'agent — L'agent utilise OpenCypher pour interroger le graphe de connaissances sur les traitements spécifiques au diabète de type 2 et à l'hypertension. Il permet également de récupérer les médicaments administrés, les dates des visites à l'hôpital, les effets secondaires des médicaments, les résultats connus des patients et les données de référence croisées pour des patients similaires (tels que des patients du même sexe et du même âge).
3. Observation de l'agent — À partir du graphe de connaissances, l'agent extrait les données tabulaires les plus récentes des six derniers mois sur les traitements administrés à des patients atteints à la fois d'hypertension et de diabète de type 2.
4. Raisonnement de l'agent — Pour classer les résultats à partir des dossiers récupérés, l'agent identifie des attributs importants, tels que la récence, les effets secondaires des médicaments ou les résultats connus pour les patients.
5. Action de l'agent : l'agent reclasse les enregistrements en fonction des attributs identifiés et de la logique prédéfinie transmise par le biais de l'invite du système.
6. Génération de réponses — Le LLM d'Amazon Bedrock génère une réponse basée sur le contexte préparé par l' ReAct agent.

Évaluation des solutions d'IA générative pour le secteur de la santé

L'évaluation des solutions d'IA pour le secteur de la santé que vous créez est essentielle pour garantir leur efficacité, leur fiabilité et leur évolutivité dans des environnements médicaux réels. Utilisez une approche systématique pour évaluer les performances de chaque composant de la solution. Vous trouverez ci-dessous un résumé des méthodologies et des mesures que vous pouvez utiliser pour évaluer votre solution.

Rubriques

- Évaluation de l'extraction d'informations
- Évaluation des solutions RAG avec plusieurs récupérateurs
- Évaluation d'une solution à l'aide d'un LLM

Évaluation de l'extraction d'informations

Évaluez les performances des solutions d'extraction d'informations, telles que l'analyseur de CV intelligent et l'extracteur d'entités personnalisé. Vous pouvez mesurer l'alignement des réponses de ces solutions à l'aide d'un jeu de données de test. Si vous ne disposez pas d'un ensemble de données couvrant les profils de talents polyvalents du secteur de la santé et les dossiers médicaux des patients, vous pouvez créer un ensemble de données de tests personnalisé en utilisant la capacité de raisonnement d'un LLM. Par exemple, vous pouvez utiliser un modèle à grands paramètres, tel que Anthropic Claude modèles, pour générer un ensemble de données de test.

Voici trois indicateurs clés que vous pouvez utiliser pour évaluer les modèles d'extraction d'informations :

- Exactitude et exhaustivité — Ces mesures évaluent la mesure dans laquelle le résultat a capturé les informations correctes et complètes présentes dans les données de vérité sur le terrain. Cela implique de vérifier à la fois l'exactitude des informations extraites et la présence de tous les détails pertinents dans les informations extraites.
- Similarité et pertinence — Ces mesures évaluent les similitudes sémantiques, structurelles et contextuelles entre les données de sortie et les données de base (la similitude) et la mesure dans laquelle la sortie s'aligne sur le contenu, le contexte et l'intention des données de base (la pertinence) et les prend en compte.

- Taux de rappel ou de capture ajusté — Ces taux déterminent de manière empirique le nombre de valeurs actuelles des données de vérité sur le terrain qui ont été correctement identifiées par le modèle. Le taux doit inclure une pénalisation pour toutes les fausses valeurs extraites par le modèle.
- Score de précision : le score de précision vous aide à déterminer le nombre de faux positifs présents dans les prédictions, par rapport aux vrais positifs. Par exemple, vous pouvez utiliser des mesures de précision pour mesurer l'exactitude de la compétence extraite.

Évaluation des solutions RAG avec plusieurs récupérateurs

Pour évaluer dans quelle mesure le système récupère les informations pertinentes et dans quelle mesure il utilise ces informations pour générer des réponses précises et adaptées au contexte, vous pouvez utiliser les métriques suivantes :

- Pertinence de la réponse — Mesurez la pertinence de la réponse générée, qui utilise le contexte extrait, par rapport à la requête d'origine.
- Précision du contexte — Sur le total des résultats récupérés, évaluez la proportion de documents ou d'extraits extraits pertinents pour la requête. Une précision contextuelle plus élevée indique que le mécanisme de récupération est efficace pour sélectionner les informations pertinentes.
- Fidélité — Évalue avec quelle précision la réponse générée reflète les informations contenues dans le contexte extrait. En d'autres termes, mesurez si la réponse reste fidèle à l'information source.

Évaluation d'une solution à l'aide d'un LLM

Vous pouvez utiliser une technique appelée LLM- as-a-judge pour évaluer les réponses textuelles de votre solution d'IA générative. Cela implique d'utiliser LLMs pour évaluer et évaluer les performances des résultats du modèle. Cette technique utilise les fonctionnalités d'Amazon Bedrock pour évaluer divers attributs, tels que la qualité des réponses, la cohérence, l'adhésion, la précision et l'exhaustivité par rapport aux préférences humaines ou aux données de base. Vous utilisez des techniques chain-of-thought (CoT) et quelques techniques d'incitation pour une évaluation complète. L'invite demande au LLM d'évaluer la réponse générée à l'aide d'une grille de notation, et les quelques échantillons de l'invite illustrent le processus d'évaluation réel. L'invite comprend également des directives à suivre par l'évaluateur LLM. Par exemple, vous pouvez envisager d'utiliser une ou plusieurs des techniques d'évaluation suivantes qui utilisent un LLM pour évaluer les réponses générées :

- Comparaison par paires — Donnez à l'évaluateur LLM une question médicale et plusieurs réponses générées par différentes versions itératives des systèmes RAG que vous avez créés. Demandez à l'évaluateur LLM de déterminer la meilleure réponse en fonction de la qualité de la réponse, de la cohérence et du respect de la question initiale.
- Notation à réponse unique : cette technique convient parfaitement aux cas d'utilisation dans lesquels vous devez évaluer l'exactitude de la catégorisation, comme la classification des résultats des patients, la catégorisation du comportement des patients, la probabilité de réadmission des patients et la catégorisation des risques. Utilisez l'évaluateur LLM pour analyser la catégorisation ou la classification individuelle de manière isolée, et évaluez le raisonnement qu'il a fourni par rapport aux données de base.
- Notation guidée par référence — Fournissez à l'évaluateur du LLM une série de questions médicales nécessitant des réponses descriptives. Créez des exemples de réponses à ces questions, tels que des réponses de référence ou des réponses idéales. Demandez à l'évaluateur LLM de comparer la réponse générée par le LLM aux réponses de référence ou aux réponses idéales, et demandez-lui d'évaluer la réponse générée en termes d'exactitude, d'exhaustivité, de similitude, de pertinence ou d'autres attributs. Cette technique vous permet d'évaluer si les réponses générées correspondent à une réponse standard ou exemplaire bien définie.

Ressources

AWS documentation

- Documentation Amazon Bedrock
- Documentation Amazon Neptune
- Documentation Amazon OpenSearch Service
- Appliquer le framework AWS Well-Architected pour Amazon Neptune (directives prescriptives)AWS
- Bonnes pratiques opérationnelles pour Amazon OpenSearch Service (documentation OpenSearch du service)
- Utilisation d'Amazon Comprehend Medical LLMs et pour les soins de santé et les sciences de la vieAWS (directives prescriptives)

AWS articles de blog

- Créez des applications RAG et d'IA générative basées sur des agents avec le nouveau modèle Amazon Titan Text Premier, disponible sur Amazon Bedrock
- Complétez l'intelligence commerciale en créant un graphe de connaissances à partir d'un entrepôt de données avec Amazon Neptune
- Utilisation de graphes de connaissances pour créer des applications GraphRag avec Amazon Bedrock et Amazon Neptune

Autres ressources

- Intégrer la génération augmentée par extraction à de grands modèles linguistiques en néphrologie : faire progresser les applications pratiques (PubMed Central, National Library of Medicine)
- Présentation de LangChain (LangChain documentation)

Collaborateurs

Conception

- Nitu Nivedita, directeur général, responsable de l'intelligence artificielle, des données et de l'IA, Accenture
- Manoj Appully, fondateur et directeur technique, Cadiem
- Conor Folan, consultant en données et intelligence artificielle, Accenture
- Deepak Krishna AR, consultant en données et intelligence artificielle, Accenture
- Almore Cato, responsable des données et de l'IA, Accenture
- Soonam Kurian, architecte de solutions principal, AWS

Révision

- Sally Lin, directrice principale de la science des données, données et intelligence artificielle, Accenture
- Terry Huang, responsable de la science des données — Données et intelligence artificielle, Accenture
- William Lorenz, architecte des solutions pour les partenaires, AWS

Rédaction technique

- Lilly AbouHarb, rédactrice technique senior, AWS

Historique du document

Le tableau suivant décrit les modifications importantes apportées à ce guide. Pour être averti des mises à jour à venir, abonnez-vous à un fil RSS.

Modification	Description	Date
Publication initiale	—	14 mars 2025

AWS Glossaire des directives prescriptives

Les termes suivants sont couramment utilisés dans les stratégies, les guides et les modèles fournis par les directives AWS prescriptives. Pour suggérer des entrées, veuillez utiliser le lien Faire un commentaire à la fin du glossaire.

Nombres

7 R

Sept politiques de migration courantes pour transférer des applications vers le cloud. Ces politiques s'appuient sur les 5 R identifiés par Gartner en 2011 et sont les suivantes :

- Refactorisation/réarchitecture : transférez une application et modifiez son architecture en tirant pleinement parti des fonctionnalités natives cloud pour améliorer l'agilité, les performances et la capacité de mise à l'échelle. Cela implique généralement le transfert du système d'exploitation et de la base de données. Exemple : migrez votre base de données Oracle sur site vers l'édition compatible avec Amazon Aurora PostgreSQL.
- Replateformer (déplacer et remodeler) : transférez une application vers le cloud et introduisez un certain niveau d'optimisation pour tirer parti des fonctionnalités du cloud. Exemple : migrez votre base de données Oracle sur site vers Amazon Relational Database Service (Amazon RDS) pour Oracle dans le. AWS Cloud
- Racheter (rachat) : optez pour un autre produit, généralement en passant d'une licence traditionnelle à un modèle SaaS. Exemple : migrez votre système de gestion de la relation client (CRM) vers Salesforce.com.
- Réhéberger (lift and shift) : transférez une application vers le cloud sans apporter de modifications pour tirer parti des fonctionnalités du cloud. Exemple : migrez votre base de données Oracle locale vers Oracle sur une EC2 instance du AWS Cloud.
- Relocaliser (lift and shift au niveau de l'hyperviseur) : transférez l'infrastructure vers le cloud sans acheter de nouveau matériel, réécrire des applications ou modifier vos opérations existantes. Vous migrez des serveurs d'une plateforme sur site vers un service cloud pour la même plateforme. Exemple : migrer une Microsoft Hyper-V application vers AWS.
- Retenir : conservez les applications dans votre environnement source. Il peut s'agir d'applications nécessitant une refactorisation majeure, que vous souhaitez retarder, et d'applications existantes que vous souhaitez retenir, car rien ne justifie leur migration sur le plan commercial.

- Retirer : mettez hors service ou supprimez les applications dont vous n'avez plus besoin dans votre environnement source.

A

ABAC

Voir contrôle d'accès basé sur les attributs.

services abstraits

Consultez la section Services gérés.

ACIDE

Voir atomicité, consistance, isolation, durabilité.

migration active-active

Méthode de migration de base de données dans laquelle la synchronisation des bases de données source et cible est maintenue (à l'aide d'un outil de réplication bidirectionnelle ou d'opérations d'écriture double), tandis que les deux bases de données gèrent les transactions provenant de la connexion d'applications pendant la migration. Cette méthode prend en charge la migration par petits lots contrôlés au lieu d'exiger un basculement ponctuel. Elle est plus flexible mais demande plus de travail qu'une migration active-passive.

migration active-passive

Méthode de migration de base de données dans laquelle la synchronisation des bases de données source et cible est maintenue, mais seule la base de données source gère les transactions provenant de la connexion d'applications pendant que les données sont répliquées vers la base de données cible. La base de données cible n'accepte aucune transaction pendant la migration.

fonction d'agrégation

Fonction SQL qui agit sur un groupe de lignes et calcule une valeur de retour unique pour le groupe. Des exemples de fonctions d'agrégation incluent `SUM` et`MAX`.

AI

Voir intelligence artificielle.

AIOps

Voir les opérations d'intelligence artificielle.

anonymisation

Processus de suppression définitive d'informations personnelles dans un ensemble de données. L'anonymisation peut contribuer à protéger la vie privée. Les données anonymisées ne sont plus considérées comme des données personnelles.

anti-motif

Solution fréquemment utilisée pour un problème récurrent lorsque la solution est contre-productive, inefficace ou moins efficace qu'une alternative.

contrôle des applications

Une approche de sécurité qui permet d'utiliser uniquement des applications approuvées afin de protéger un système contre les logiciels malveillants.

portefeuille d'applications

Ensemble d'informations détaillées sur chaque application utilisée par une organisation, y compris le coût de génération et de maintenance de l'application, ainsi que sa valeur métier. Ces informations sont essentielles pour le processus de découverte et d'analyse du portefeuille et permettent d'identifier et de prioriser les applications à migrer, à moderniser et à optimiser.

intelligence artificielle (IA)

Domaine de l'informatique consacré à l'utilisation des technologies de calcul pour exécuter des fonctions cognitives généralement associées aux humains, telles que l'apprentissage, la résolution de problèmes et la reconnaissance de modèles. Pour plus d'informations, veuillez consulter Qu'est-ce que l'intelligence artificielle ?

opérations d'intelligence artificielle (AIOps)

Processus consistant à utiliser des techniques de machine learning pour résoudre les problèmes opérationnels, réduire les incidents opérationnels et les interventions humaines, mais aussi améliorer la qualité du service. Pour plus d'informations sur son AIOps utilisation dans la stratégie de AWS migration, consultez le guide d'intégration des opérations.

chiffrement asymétrique

Algorithme de chiffrement qui utilise une paire de clés, une clé publique pour le chiffrement et une clé privée pour le déchiffrement. Vous pouvez partager la clé publique, car elle n'est pas utilisée pour le déchiffrement, mais l'accès à la clé privée doit être très restreint.

atomicité, cohérence, isolement, durabilité (ACID)

Ensemble de propriétés logicielles garantissant la validité des données et la fiabilité opérationnelle d'une base de données, même en cas d'erreur, de panne de courant ou d'autres problèmes.

contrôle d'accès par attributs (ABAC)

Pratique qui consiste à créer des autorisations détaillées en fonction des attributs de l'utilisateur, tels que le service, le poste et le nom de l'équipe. Pour plus d'informations, consultez ABAC pour AWS dans la documentation AWS Identity and Access Management (IAM).

source de données faisant autorité

Emplacement où vous stockez la version principale des données, considérée comme la source d'information la plus fiable. Vous pouvez copier les données de la source de données officielle vers d'autres emplacements à des fins de traitement ou de modification des données, par exemple en les anonymisant, en les expurgant ou en les pseudonymisant.

Zone de disponibilité

Un emplacement distinct au sein d'un Région AWS réseau isolé des défaillances dans d'autres zones de disponibilité et fournissant une connectivité réseau peu coûteuse et à faible latence aux autres zones de disponibilité de la même région.

AWS Cadre d'adoption du cloud (AWS CAF)

Un cadre de directives et de meilleures pratiques visant AWS à aider les entreprises à élaborer un plan efficace pour réussir leur migration vers le cloud. AWS La CAF organise ses conseils en six domaines prioritaires appelés perspectives : les affaires, les personnes, la gouvernance, les plateformes, la sécurité et les opérations. Les perspectives d'entreprise, de personnes et de gouvernance mettent l'accent sur les compétences et les processus métier, tandis que les perspectives relatives à la plateforme, à la sécurité et aux opérations se concentrent sur les compétences et les processus techniques. Par exemple, la perspective liée aux personnes cible les parties prenantes qui s'occupent des ressources humaines (RH), des fonctions de dotation en personnel et de la gestion des personnes. Dans cette perspective, la AWS CAF fournit des conseils pour le développement du personnel, la formation et les communications afin de préparer

l'organisation à une adoption réussie du cloud. Pour plus d'informations, veuillez consulter le site Web AWS CAF et le livre blanc AWS CAF.

AWS Cadre de qualification de la charge de travail (AWS WQF)

Outil qui évalue les charges de travail liées à la migration des bases de données, recommande des stratégies de migration et fournit des estimations de travail. AWS Le WQF est inclus avec AWS Schema Conversion Tool (AWS SCT). Il analyse les schémas de base de données et les objets de code, le code d'application, les dépendances et les caractéristiques de performance, et fournit des rapports d'évaluation.

B

mauvais bot

Un bot destiné à perturber ou à nuire à des individus ou à des organisations.

BCP

Consultez la section Planification de la continuité des activités.

graphique de comportement

Vue unifiée et interactive des comportements des ressources et des interactions au fil du temps. Vous pouvez utiliser un graphique de comportement avec Amazon Detective pour examiner les tentatives de connexion infructueuses, les appels d'API suspects et les actions similaires. Pour plus d'informations, veuillez consulter Data in a behavior graph dans la documentation Detective.

système de poids fort

Système qui stocke d'abord l'octet le plus significatif. Voir aussi endianité.

classification binaire

Processus qui prédit un résultat binaire (l'une des deux classes possibles). Par exemple, votre modèle de machine learning peut avoir besoin de prévoir des problèmes tels que « Cet e-mail est-il du spam ou non ? » ou « Ce produit est-il un livre ou une voiture ? ».

filtre de Bloom

Structure de données probabiliste et efficace en termes de mémoire qui est utilisée pour tester si un élément fait partie d'un ensemble.

déploiement bleu/vert

Stratégie de déploiement dans laquelle vous créez deux environnements distincts mais identiques. Vous exécutez la version actuelle de l'application dans un environnement (bleu) et la nouvelle version de l'application dans l'autre environnement (vert). Cette stratégie vous permet de revenir rapidement en arrière avec un impact minimal.

bot

Application logicielle qui exécute des tâches automatisées sur Internet et simule l'activité ou l'interaction humaine. Certains robots sont utiles ou bénéfiques, comme les robots d'exploration Web qui indexent des informations sur Internet. D'autres robots, appelés « bots malveillants », sont destinés à perturber ou à nuire à des individus ou à des organisations.

botnet

Réseaux de robots infectés par des logiciels malveillants et contrôlés par une seule entité, connue sous le nom d'herder ou d'opérateur de bots. Les botnets sont le mécanisme le plus connu pour faire évoluer les bots et leur impact.

branche

Zone contenue d'un référentiel de code. La première branche créée dans un référentiel est la branche principale. Vous pouvez créer une branche à partir d'une branche existante, puis développer des fonctionnalités ou corriger des bogues dans la nouvelle branche. Une branche que vous créez pour générer une fonctionnalité est communément appelée branche de fonctionnalités. Lorsque la fonctionnalité est prête à être publiée, vous fusionnez à nouveau la branche de fonctionnalités dans la branche principale. Pour plus d'informations, consultez À propos des branches (GitHub documentation).

accès par brise-vitre

Dans des circonstances exceptionnelles et par le biais d'un processus approuvé, c'est un moyen rapide pour un utilisateur d'accéder à un accès auquel Compte AWS il n'est généralement pas autorisé. Pour plus d'informations, consultez l'indicateur Implementation break-glass procedures dans le guide Well-Architected AWS .

stratégie existante (brownfield)

L'infrastructure existante de votre environnement. Lorsque vous adoptez une stratégie existante pour une architecture système, vous concevez l'architecture en fonction des contraintes des systèmes et de l'infrastructure actuels. Si vous étendez l'infrastructure existante, vous pouvez combiner des politiques brownfield (existantes) et greenfield (inédites).

cache de tampon

Zone de mémoire dans laquelle sont stockées les données les plus fréquemment consultées.

capacité métier

Ce que fait une entreprise pour générer de la valeur (par exemple, les ventes, le service client ou le marketing). Les architectures de microservices et les décisions de développement peuvent être dictées par les capacités métier. Pour plus d'informations, veuillez consulter la section Organisation en fonction des capacités métier du livre blanc Exécution de microservices conteneurisés sur AWS.

planification de la continuité des activités (BCP)

Plan qui tient compte de l'impact potentiel d'un événement perturbateur, tel qu'une migration à grande échelle, sur les opérations, et qui permet à une entreprise de reprendre ses activités rapidement.

C

CAF

Voir le cadre d'adoption du AWS cloud.

déploiement de Canary

Diffusion lente et progressive d'une version pour les utilisateurs finaux. Lorsque vous êtes sûr, vous déployez la nouvelle version et remplacez la version actuelle dans son intégralité.

CCo E

Voir le Centre d'excellence du cloud.

CDC

Voir capture des données de modification.

capture des données de modification (CDC)

Processus de suivi des modifications apportées à une source de données, telle qu'une table de base de données, et d'enregistrement des métadonnées relatives à ces modifications. Vous pouvez utiliser la CDC à diverses fins, telles que l'audit ou la réplication des modifications dans un système cible afin de maintenir la synchronisation.

ingénierie du chaos

Introduire intentionnellement des défaillances ou des événements perturbateurs pour tester la résilience d'un système. Vous pouvez utiliser AWS Fault Injection Service (AWS FIS) pour effectuer des expériences qui stressent vos AWS charges de travail et évaluer leur réponse.

CI/CD

Découvrez l'intégration continue et la livraison continue.

classification

Processus de catégorisation qui permet de générer des prédictions. Les modèles de ML pour les problèmes de classification prédisent une valeur discrète. Les valeurs discrètes se distinguent toujours les unes des autres. Par exemple, un modèle peut avoir besoin d'évaluer la présence ou non d'une voiture sur une image.

chiffrement côté client

Chiffrement des données localement, avant que la cible ne les Service AWS reçoive.

Centre d'excellence du cloud (CCoE)

Une équipe multidisciplinaire qui dirige les efforts d'adoption du cloud au sein d'une organisation, notamment en développant les bonnes pratiques en matière de cloud, en mobilisant des ressources, en établissant des délais de migration et en guidant l'organisation dans le cadre de transformations à grande échelle. Pour plus d'informations, consultez les CCoarticles électroniques du blog sur la stratégie AWS Cloud d'entreprise.

cloud computing

Technologie cloud généralement utilisée pour le stockage de données à distance et la gestion des appareils IoT. Le cloud computing est généralement associé à la technologie informatique de pointe.

modèle d'exploitation du cloud

Dans une organisation informatique, modèle d'exploitation utilisé pour créer, faire évoluer et optimiser un ou plusieurs environnements cloud. Pour plus d'informations, consultez la section Création de votre modèle d'exploitation cloud.

étapes d'adoption du cloud

Les quatre phases que les entreprises traversent généralement lorsqu'elles migrent vers AWS Cloud :

- Projet : exécution de quelques projets liés au cloud à des fins de preuve de concept et d'apprentissage
- Base : réaliser des investissements fondamentaux pour accélérer votre adoption du cloud (par exemple, créer une zone de landing zone, définir un CCo E, établir un modèle opérationnel)
- Migration : migration d'applications individuelles
- Réinvention : optimisation des produits et services et innovation dans le cloud

Ces étapes ont été définies par Stephen Orban dans le billet de blog The Journey Toward Cloud-First & the Stages of Adoption publié sur le blog AWS Cloud Enterprise Strategy. Pour plus d'informations sur leur lien avec la stratégie de AWS migration, consultez le guide de préparation à la migration.

CMDB

Voir base de données de gestion de configuration.

référentiel de code

Emplacement où le code source et d'autres ressources, comme la documentation, les exemples et les scripts, sont stockés et mis à jour par le biais de processus de contrôle de version. Les référentiels cloud courants incluent GitHub ouBitbucket Cloud. Chaque version du code est appelée branche. Dans une structure de microservice, chaque référentiel est consacré à une seule fonctionnalité. Un seul pipeline CI/CD peut utiliser plusieurs référentiels.

cache passif

Cache tampon vide, mal rempli ou contenant des données obsolètes ou non pertinentes. Cela affecte les performances, car l'instance de base de données doit lire à partir de la mémoire principale ou du disque, ce qui est plus lent que la lecture à partir du cache tampon.

données gelées

Données rarement consultées et généralement historiques. Lorsque vous interrogez ce type de données, les requêtes lentes sont généralement acceptables. Le transfert de ces données vers des niveaux ou classes de stockage moins performants et moins coûteux peut réduire les coûts.

vision par ordinateur (CV)

Domaine de l'IA qui utilise l'apprentissage automatique pour analyser et extraire des informations à partir de formats visuels tels que des images numériques et des vidéos. Par exemple, Amazon SageMaker AI fournit des algorithmes de traitement d'image pour les CV.

dérive de configuration

Pour une charge de travail, une modification de configuration par rapport à l'état attendu. Cela peut entraîner une non-conformité de la charge de travail, et cela est généralement progressif et involontaire.

base de données de gestion des configurations (CMDB)

Référentiel qui stocke et gère les informations relatives à une base de données et à son environnement informatique, y compris les composants matériels et logiciels ainsi que leurs configurations. Vous utilisez généralement les données d'une CMDB lors de la phase de découverte et d'analyse du portefeuille de la migration.

pack de conformité

Ensemble de AWS Config règles et d'actions correctives que vous pouvez assembler pour personnaliser vos contrôles de conformité et de sécurité. Vous pouvez déployer un pack de conformité en tant qu'entité unique dans une région Compte AWS et, ou au sein d'une organisation, à l'aide d'un modèle YAML. Pour plus d'informations, consultez la section Packs de conformité dans la AWS Config documentation.

intégration continue et livraison continue (CI/CD)

Processus d'automatisation des étapes de source, de construction, de test, de préparation et de production du processus de publication du logiciel. CI/CD is commonly described as a pipeline. CI/CDpeut vous aider à automatiser les processus, à améliorer la productivité, à améliorer la qualité du code et à accélérer les livraisons. Pour plus d'informations, veuillez consulter Avantages de la livraison continue. CD peut également signifier déploiement continu. Pour plus d'informations, veuillez consulter Livraison continue et déploiement continu.

CV

Voir vision par ordinateur.

D

données au repos

Données stationnaires dans votre réseau, telles que les données stockées.

classification des données

Processus permettant d'identifier et de catégoriser les données de votre réseau en fonction de leur sévérité et de leur sensibilité. Il s'agit d'un élément essentiel de toute stratégie de gestion des risques de cybersécurité, car il vous aide à déterminer les contrôles de protection et de conservation appropriés pour les données. La classification des données est une composante du pilier de sécurité du AWS Well-Architected Framework. Pour plus d'informations, veuillez consulter Classification des données.

dérive des données

Une variation significative entre les données de production et les données utilisées pour entraîner un modèle ML, ou une modification significative des données d'entrée au fil du temps. La dérive des données peut réduire la qualité, la précision et l'équité globales des prédictions des modèles ML.

données en transit

Données qui circulent activement sur votre réseau, par exemple entre les ressources du réseau.

maillage de données

Un cadre architectural qui fournit une propriété des données distribuée et décentralisée avec une gestion et une gouvernance centralisées.

minimisation des données

Le principe de collecte et de traitement des seules données strictement nécessaires. La pratique de la minimisation des données AWS Cloud peut réduire les risques liés à la confidentialité, les coûts et l'empreinte carbone de vos analyses.

périmètre de données

Ensemble de garde-fous préventifs dans votre AWS environnement qui permettent de garantir que seules les identités fiables accèdent aux ressources fiables des réseaux attendus. Pour plus d'informations, voir Création d'un périmètre de données sur AWS.

prétraitement des données

Pour transformer les données brutes en un format facile à analyser par votre modèle de ML. Le prétraitement des données peut impliquer la suppression de certaines colonnes ou lignes et le traitement des valeurs manquantes, incohérentes ou en double.

provenance des données

Le processus de suivi de l'origine et de l'historique des données tout au long de leur cycle de vie, par exemple la manière dont les données ont été générées, transmises et stockées.

sujet des données

Personne dont les données sont collectées et traitées.

entrepôt des données

Un système de gestion des données qui prend en charge les informations commerciales, telles que les analyses. Les entrepôts de données contiennent généralement de grandes quantités de données historiques et sont généralement utilisés pour les requêtes et les analyses.

langage de définition de base de données (DDL)

Instructions ou commandes permettant de créer ou de modifier la structure des tables et des objets dans une base de données.

langage de manipulation de base de données (DML)

Instructions ou commandes permettant de modifier (insérer, mettre à jour et supprimer) des informations dans une base de données.

DDL

Voir langage de définition de base de données.

ensemble profond

Sert à combiner plusieurs modèles de deep learning à des fins de prédiction. Vous pouvez utiliser des ensembles profonds pour obtenir une prévision plus précise ou pour estimer l'incertitude des prédictions.

deep learning

Un sous-champ de ML qui utilise plusieurs couches de réseaux neuronaux artificiels pour identifier le mappage entre les données d'entrée et les variables cibles d'intérêt.

defense-in-depth

Approche de la sécurité de l'information dans laquelle une série de mécanismes et de contrôles de sécurité sont judicieusement répartis sur l'ensemble d'un réseau informatique afin de protéger la confidentialité, l'intégrité et la disponibilité du réseau et des données qu'il contient. Lorsque vous adoptez cette stratégie AWS, vous ajoutez plusieurs contrôles à différentes couches de

la AWS Organizations structure afin de sécuriser les ressources. Par exemple, une defense-in-depth approche peut combiner l'authentification multifactorielle, la segmentation du réseau et le chiffrement.

administrateur délégué

Dans AWS Organizations, un service compatible peut enregistrer un compte AWS membre pour administrer les comptes de l'organisation et gérer les autorisations pour ce service. Ce compte est appelé administrateur délégué pour ce service. Pour plus d'informations et une liste des services compatibles, veuillez consulter la rubrique Services qui fonctionnent avec AWS Organizations dans la documentation AWS Organizations .

déploiement

Processus de mise à disposition d'une application, de nouvelles fonctionnalités ou de corrections de code dans l'environnement cible. Le déploiement implique la mise en œuvre de modifications dans une base de code, puis la génération et l'exécution de cette base de code dans les environnements de l'application.

environnement de développement

Voir environnement.

contrôle de détection

Contrôle de sécurité conçu pour détecter, journaliser et alerter après la survenue d'un événement. Ces contrôles constituent une deuxième ligne de défense et vous alertent en cas d'événements de sécurité qui ont contourné les contrôles préventifs en place. Pour plus d'informations, veuillez consulter la rubrique Contrôles de détection dans Implementing security controls on AWS.

cartographie de la chaîne de valeur du développement (DVSM)

Processus utilisé pour identifier et hiérarchiser les contraintes qui nuisent à la rapidité et à la qualité du cycle de vie du développement logiciel. DVSM étend le processus de cartographie de la chaîne de valeur initialement conçu pour les pratiques de production allégée. Il met l'accent sur les étapes et les équipes nécessaires pour créer et transférer de la valeur tout au long du processus de développement logiciel.

jumeau numérique

Représentation virtuelle d'un système réel, tel qu'un bâtiment, une usine, un équipement industriel ou une ligne de production. Les jumeaux numériques prennent en charge la maintenance prédictive, la surveillance à distance et l'optimisation de la production.

tableau des dimensions

Dans un schéma en étoile, table plus petite contenant les attributs de données relatifs aux données quantitatives d'une table de faits. Les attributs des tables de dimensions sont généralement des champs de texte ou des nombres discrets qui se comportent comme du texte. Ces attributs sont couramment utilisés pour la contrainte des requêtes, le filtrage et l'étiquetage des ensembles de résultats.

catastrophe

Un événement qui empêche une charge de travail ou un système d'atteindre ses objectifs commerciaux sur son site de déploiement principal. Ces événements peuvent être des catastrophes naturelles, des défaillances techniques ou le résultat d'actions humaines, telles qu'une mauvaise configuration involontaire ou une attaque de logiciel malveillant.

reprise après sinistre (DR)

La stratégie et le processus que vous utilisez pour minimiser les temps d'arrêt et les pertes de données causés par un sinistre. Pour plus d'informations, consultez Disaster Recovery of Workloads on AWS : Recovery in the Cloud in the AWS Well-Architected Framework.

DML

Voir langage de manipulation de base de données.

conception axée sur le domaine

Approche visant à développer un système logiciel complexe en connectant ses composants à des domaines évolutifs, ou objectifs métier essentiels, que sert chaque composant. Ce concept a été introduit par Eric Evans dans son ouvrage Domain-Driven Design: Tackling Complexity in the Heart of Software (Boston : Addison-Wesley Professional, 2003). Pour plus d'informations sur l'utilisation du design piloté par domaine avec le modèle de figuier étrangleur, veuillez consulter Modernizing legacy Microsoft ASP.NET (ASMX) web services incrementally by using containers and Amazon API Gateway.

DR

Voir reprise après sinistre.

détection de dérive

Suivi des écarts par rapport à une configuration de référence. Par exemple, vous pouvez l'utiliser AWS CloudFormation pour détecter la dérive des ressources du système ou AWS Control Tower

pour détecter les modifications de votre zone d'atterrissage susceptibles d'affecter le respect des exigences de gouvernance.

DVSM

Voir la cartographie de la chaîne de valeur du développement.

E

EDA

Voir analyse exploratoire des données.

EDI

Voir échange de données informatisé.

informatique de périphérie

Technologie qui augmente la puissance de calcul des appareils intelligents en périphérie d'un réseau IoT. Comparé au cloud computing, l'informatique de pointe peut réduire la latence des communications et améliorer le temps de réponse.

échange de données informatisé (EDI)

L'échange automatique de documents commerciaux entre les organisations. Pour plus d'informations, voir Qu'est-ce que l'échange de données informatisé ?

chiffrement

Processus informatique qui transforme des données en texte clair, lisibles par l'homme, en texte chiffré.

clé de chiffrement

Chaîne cryptographique de bits aléatoires générée par un algorithme cryptographique. La longueur des clés peut varier, et chaque clé est conçue pour être imprévisible et unique.

endianisme

Ordre selon lequel les octets sont stockés dans la mémoire de l'ordinateur. Les systèmes de poids fort stockent d'abord l'octet le plus significatif. Les systèmes de poids faible stockent d'abord l'octet le moins significatif.

point de terminaison

Voir point de terminaison de service.

service de point de terminaison

Service que vous pouvez héberger sur un cloud privé virtuel (VPC) pour le partager avec d'autres utilisateurs. Vous pouvez créer un service de point de terminaison avec AWS PrivateLink et accorder des autorisations à d'autres Comptes AWS ou à AWS Identity and Access Management (IAM) principaux. Ces comptes ou principaux peuvent se connecter à votre service de point de terminaison de manière privée en créant des points de terminaison d'un VPC d'interface. Pour plus d'informations, veuillez consulter Création d'un service de point de terminaison dans la documentation Amazon Virtual Private Cloud (Amazon VPC).

planification des ressources d'entreprise (ERP)

Système qui automatise et gère les principaux processus métier (tels que la comptabilité, le MES et la gestion de projet) pour une entreprise.

chiffrement d'enveloppe

Processus de chiffrement d'une clé de chiffrement à l'aide d'une autre clé de chiffrement. Pour plus d'informations, consultez la section Chiffrement des enveloppes dans la documentation AWS Key Management Service (AWS KMS).

environment

Instance d'une application en cours d'exécution. Les types d'environnement les plus courants dans le cloud computing sont les suivants :

- Environnement de développement : instance d'une application en cours d'exécution à laquelle seule l'équipe principale chargée de la maintenance de l'application peut accéder. Les environnements de développement sont utilisés pour tester les modifications avant de les promouvoir dans les environnements supérieurs. Ce type d'environnement est parfois appelé environnement de test.
- Environnements inférieurs : tous les environnements de développement d'une application, tels que ceux utilisés pour les générations et les tests initiaux.
- Environnement de production : instance d'une application en cours d'exécution à laquelle les utilisateurs finaux peuvent accéder. Dans un pipeline CI/CD, l'environnement de production est le dernier environnement de déploiement.
- Environnements supérieurs : tous les environnements accessibles aux utilisateurs autres que l'équipe de développement principale. Ils peuvent inclure un environnement de production, des

environnements de préproduction et des environnements pour les tests d'acceptation par les utilisateurs.

épopée

Dans les méthodologies agiles, catégories fonctionnelles qui aident à organiser et à prioriser votre travail. Les épopées fournissent une description détaillée des exigences et des tâches d'implémentation. Par exemple, les points forts de la AWS CAF en matière de sécurité incluent la gestion des identités et des accès, les contrôles de détection, la sécurité des infrastructures, la protection des données et la réponse aux incidents. Pour plus d'informations sur les épopées dans la stratégie de migration AWS , veuillez consulter le guide d'implémentation du programme.

ERP

Voir Planification des ressources d'entreprise.

analyse exploratoire des données (EDA)

Processus d'analyse d'un jeu de données pour comprendre ses principales caractéristiques. Vous collectez ou agrégez des données, puis vous effectuez des enquêtes initiales pour trouver des modèles, détecter des anomalies et vérifier les hypothèses. L'EDA est réalisée en calculant des statistiques récapitulatives et en créant des visualisations de données.

F

tableau des faits

La table centrale dans un schéma en étoile. Il stocke des données quantitatives sur les opérations commerciales. Généralement, une table de faits contient deux types de colonnes : celles qui contiennent des mesures et celles qui contiennent une clé étrangère pour une table de dimensions.

échouer rapidement

Une philosophie qui utilise des tests fréquents et progressifs pour réduire le cycle de vie du développement. C'est un élément essentiel d'une approche agile.

limite d'isolation des défauts

Dans le AWS Cloud, une limite telle qu'une zone de disponibilité Région AWS, un plan de contrôle ou un plan de données qui limite l'effet d'une panne et contribue à améliorer la résilience des

charges de travail. Pour plus d'informations, consultez la section Limites d'isolation des AWS pannes.

branche de fonctionnalités

Voir succursale.

fonctionnalités

Les données d'entrée que vous utilisez pour faire une prédiction. Par exemple, dans un contexte de fabrication, les fonctionnalités peuvent être des images capturées périodiquement à partir de la ligne de fabrication.

importance des fonctionnalités

Le niveau d'importance d'une fonctionnalité pour les prédictions d'un modèle. Il s'exprime généralement sous la forme d'un score numérique qui peut être calculé à l'aide de différentes techniques, telles que la méthode Shapley Additive Explanations (SHAP) et les gradients intégrés. Pour plus d'informations, voir Interprétabilité du modèle d'apprentissage automatique avec AWS.

transformation de fonctionnalité

Optimiser les données pour le processus de ML, notamment en enrichissant les données avec des sources supplémentaires, en mettant à l'échelle les valeurs ou en extrayant plusieurs ensembles d'informations à partir d'un seul champ de données. Cela permet au modèle de ML de tirer parti des données. Par exemple, si vous décomposez la date « 2021-05-27 00:15:37 » en « 2021 », « mai », « jeudi » et « 15 », vous pouvez aider l'algorithme d'apprentissage à apprendre des modèles nuancés associés à différents composants de données.

invitation en quelques coups

Fournir à un LLM un petit nombre d'exemples illustrant la tâche et le résultat souhaité avant de lui demander d'effectuer une tâche similaire. Cette technique est une application de l'apprentissage contextuel, dans le cadre de laquelle les modèles apprennent à partir d'exemples (prises de vue) intégrés dans des instructions. Les instructions en quelques étapes peuvent être efficaces pour les tâches qui nécessitent un formatage, un raisonnement ou des connaissances de domaine spécifiques. Voir également l'invite Zero-Shot.

FGAC

Découvrez le contrôle d'accès détaillé.

contrôle d'accès détaillé (FGAC)

Utilisation de plusieurs conditions pour autoriser ou refuser une demande d'accès.

migration instantanée (flash-cut)

Méthode de migration de base de données qui utilise la réplication continue des données par le biais de la capture des données de modification afin de migrer les données dans les plus brefs délais, au lieu d'utiliser une approche progressive. L'objectif est de réduire au maximum les temps d'arrêt.

FM

Voir le modèle de fondation.

modèle de fondation (FM)

Un vaste réseau neuronal d'apprentissage profond qui s'est entraîné sur d'énormes ensembles de données généralisées et non étiquetées. FMs sont capables d'effectuer une grande variété de tâches générales, telles que comprendre le langage, générer du texte et des images et converser en langage naturel. Pour plus d'informations, voir Que sont les modèles de base ?

G

IA générative

Sous-ensemble de modèles d'IA qui ont été entraînés sur de grandes quantités de données et qui peuvent utiliser une simple invite textuelle pour créer de nouveaux contenus et artefacts, tels que des images, des vidéos, du texte et du son. Pour plus d'informations, consultez Qu'est-ce que l'IA générative.

blocage géographique

Voir les restrictions géographiques.

restrictions géographiques (blocage géographique)

Sur Amazon CloudFront, option permettant d'empêcher les utilisateurs de certains pays d'accéder aux distributions de contenu. Vous pouvez utiliser une liste d'autorisation ou une liste de blocage pour spécifier les pays approuvés et interdits. Pour plus d'informations, consultez la section Restreindre la distribution géographique de votre contenu dans la CloudFront documentation.

Flux de travail Gitflow

Approche dans laquelle les environnements inférieurs et supérieurs utilisent différentes branches dans un référentiel de code source. Le flux de travail Gitflow est considéré comme existant, et le flux de travail basé sur les troncs est l'approche moderne préférée.

image dorée

Un instantané d'un système ou d'un logiciel utilisé comme modèle pour déployer de nouvelles instances de ce système ou logiciel. Par exemple, dans le secteur de la fabrication, une image dorée peut être utilisée pour fournir des logiciels sur plusieurs appareils et contribue à améliorer la vitesse, l'évolutivité et la productivité des opérations de fabrication des appareils.

stratégie inédite

L'absence d'infrastructures existantes dans un nouvel environnement. Lorsque vous adoptez une stratégie inédite pour une architecture système, vous pouvez sélectionner toutes les nouvelles technologies sans restriction de compatibilité avec l'infrastructure existante, également appelée brownfield. Si vous étendez l'infrastructure existante, vous pouvez combiner des politiques brownfield (existantes) et greenfield (inédites).

barrière de protection

Règle de haut niveau qui permet de régir les ressources, les politiques et la conformité au sein des unités organisationnelles (OUs). Les barrières de protection préventives appliquent des politiques pour garantir l'alignement sur les normes de conformité. Elles sont mises en œuvre à l'aide de politiques de contrôle des services et de limites des autorisations IAM. Les barrières de protection de détection détectent les violations des politiques et les problèmes de conformité, et génèrent des alertes pour y remédier. Ils sont implémentés à l'aide d'Amazon AWS Config AWS Security Hub GuardDuty AWS Trusted Advisor, d'Amazon Inspector et de AWS Lambda contrôles personnalisés.

H

HA

Découvrez la haute disponibilité.

migration de base de données hétérogène

Migration de votre base de données source vers une base de données cible qui utilise un moteur de base de données différent (par exemple, Oracle vers Amazon Aurora). La migration hétérogène fait généralement partie d'un effort de réarchitecture, et la conversion du schéma peut s'avérer une tâche complexe. AWS propose AWS SCT qui facilite les conversions de schémas.

haute disponibilité (HA)

Capacité d'une charge de travail à fonctionner en continu, sans intervention, en cas de difficultés ou de catastrophes. Les systèmes HA sont conçus pour basculer automatiquement, fournir constamment des performances de haute qualité et gérer différentes charges et défaillances avec un impact minimal sur les performances.

modernisation des historiens

Approche utilisée pour moderniser et mettre à niveau les systèmes de technologie opérationnelle (OT) afin de mieux répondre aux besoins de l'industrie manufacturière. Un historien est un type de base de données utilisé pour collecter et stocker des données provenant de diverses sources dans une usine.

données de rétention

Partie de données historiques étiquetées qui n'est pas divulguée dans un ensemble de données utilisé pour entraîner un modèle d'apprentissage automatique. Vous pouvez utiliser les données de blocage pour évaluer les performances du modèle en comparant les prévisions du modèle aux données de blocage.

migration de base de données homogène

Migration de votre base de données source vers une base de données cible qui partage le même moteur de base de données (par exemple, Microsoft SQL Server vers Amazon RDS for SQL Server). La migration homogène s'inscrit généralement dans le cadre d'un effort de réhébergement ou de replateforme. Vous pouvez utiliser les utilitaires de base de données natifs pour migrer le schéma.

données chaudes

Données fréquemment consultées, telles que les données en temps réel ou les données translationnelles récentes. Ces données nécessitent généralement un niveau ou une classe de stockage à hautes performances pour fournir des réponses rapides aux requêtes.

correctif

Solution d'urgence à un problème critique dans un environnement de production. En raison de son urgence, un correctif est généralement créé en dehors du flux de travail de DevOps publication habituel.

période de soins intensifs

Immédiatement après le basculement, période pendant laquelle une équipe de migration gère et surveille les applications migrées dans le cloud afin de résoudre les problèmes éventuels. En règle générale, cette période dure de 1 à 4 jours. À la fin de la période de soins intensifs, l'équipe de migration transfère généralement la responsabilité des applications à l'équipe des opérations cloud.

I

IaC

Considérez l'infrastructure comme un code.

politique basée sur l’identité

Politique attachée à un ou plusieurs principaux IAM qui définit leurs autorisations au sein de l' AWS Cloud environnement.

application inactive

Application dont l'utilisation moyenne du processeur et de la mémoire se situe entre 5 et 20 % sur une période de 90 jours. Dans un projet de migration, il est courant de retirer ces applications ou de les retenir sur site.

IIo T

Voir Internet industriel des objets.

infrastructure immuable

Modèle qui déploie une nouvelle infrastructure pour les charges de travail de production au lieu de mettre à jour, d'appliquer des correctifs ou de modifier l'infrastructure existante. Les infrastructures immuables sont intrinsèquement plus cohérentes, fiables et prévisibles que les infrastructures mutables. Pour plus d'informations, consultez les meilleures pratiques de déploiement à l'aide d'une infrastructure immuable dans le AWS Well-Architected Framework.

VPC entrant (d'entrée)

Dans une architecture AWS multi-comptes, un VPC qui accepte, inspecte et achemine les connexions réseau depuis l'extérieur d'une application. L'architecture AWS de référence de sécurité recommande de configurer votre compte réseau avec les fonctions entrantes, sortantes

et d'inspection VPCs afin de protéger l'interface bidirectionnelle entre votre application et l'Internet en général.

migration incrémentielle

Stratégie de basculement dans le cadre de laquelle vous migrez votre application par petites parties au lieu d'effectuer un basculement complet unique. Par exemple, il se peut que vous ne transfériez que quelques microservices ou utilisateurs vers le nouveau système dans un premier temps. Après avoir vérifié que tout fonctionne correctement, vous pouvez transférer progressivement des microservices ou des utilisateurs supplémentaires jusqu'à ce que vous puissiez mettre hors service votre système hérité. Cette stratégie réduit les risques associés aux migrations de grande ampleur.

Industry 4.0

Terme introduit par Klaus Schwab en 2016 pour désigner la modernisation des processus de fabrication grâce aux avancées en matière de connectivité, de données en temps réel, d'automatisation, d'analyse et d'IA/ML.

infrastructure

Ensemble des ressources et des actifs contenus dans l'environnement d'une application.

infrastructure en tant que code (IaC)

Processus de mise en service et de gestion de l'infrastructure d'une application via un ensemble de fichiers de configuration. IaC est conçue pour vous aider à centraliser la gestion de l'infrastructure, à normaliser les ressources et à mettre à l'échelle rapidement afin que les nouveaux environnements soient reproductibles, fiables et cohérents.

Internet industriel des objets (IIoT)

L'utilisation de capteurs et d'appareils connectés à Internet dans les secteurs industriels tels que la fabrication, l'énergie, l'automobile, les soins de santé, les sciences de la vie et l'agriculture. Pour plus d'informations, voir Élaboration d'une stratégie de transformation numérique de l'Internet des objets (IIoT) industriel.

VPC d'inspection

Dans une architecture AWS multi-comptes, un VPC centralisé qui gère les inspections du trafic réseau VPCs entre (identique ou Régions AWS différent), Internet et les réseaux locaux. L'architecture AWS de référence de sécurité recommande de configurer votre compte réseau

avec les fonctions entrantes, sortantes et d'inspection VPCs afin de protéger l'interface bidirectionnelle entre votre application et l'Internet en général.

Internet des objets (IoT)

Réseau d'objets physiques connectés dotés de capteurs ou de processeurs intégrés qui communiquent avec d'autres appareils et systèmes via Internet ou via un réseau de communication local. Pour plus d'informations, veuillez consulter la section Qu'est-ce que l'IoT ?.

interprétabilité

Caractéristique d'un modèle de machine learning qui décrit dans quelle mesure un être humain peut comprendre comment les prédictions du modèle dépendent de ses entrées. Pour plus d'informations, voir Interprétabilité du modèle d'apprentissage automatique avec AWS.

IoT

Voir Internet des objets.

Bibliothèque d'informations informatiques (ITIL)

Ensemble de bonnes pratiques pour proposer des services informatiques et les aligner sur les exigences métier. L'ITIL constitue la base de l'ITSM.

gestion des services informatiques (ITSM)

Activités associées à la conception, à la mise en œuvre, à la gestion et à la prise en charge de services informatiques d'une organisation. Pour plus d'informations sur l'intégration des opérations cloud aux outils ITSM, veuillez consulter le guide d'intégration des opérations.

ITIL

Consultez la bibliothèque d'informations informatiques.

ITSM

Voir Gestion des services informatiques.

L

contrôle d'accès basé sur des étiquettes (LBAC)

Une implémentation du contrôle d'accès obligatoire (MAC) dans laquelle une valeur d'étiquette de sécurité est explicitement attribuée aux utilisateurs et aux données elles-mêmes. L'intersection

entre l'étiquette de sécurité utilisateur et l'étiquette de sécurité des données détermine les lignes et les colonnes visibles par l'utilisateur.

zone de destination

Une zone d'atterrissage est un AWS environnement multi-comptes bien conçu, évolutif et sécurisé. Il s'agit d'un point de départ à partir duquel vos entreprises peuvent rapidement lancer et déployer des charges de travail et des applications en toute confiance dans leur environnement de sécurité et d'infrastructure. Pour plus d'informations sur les zones de destination, veuillez consulter Setting up a secure and scalable multi-account AWS environment.

grand modèle de langage (LLM)

Un modèle d'intelligence artificielle basé sur le deep learning qui est préentraîné sur une grande quantité de données. Un LLM peut effectuer plusieurs tâches, telles que répondre à des questions, résumer des documents, traduire du texte dans d'autres langues et compléter des phrases. Pour plus d'informations, voir Que sont LLMs.

migration de grande envergure

Migration de 300 serveurs ou plus.

LBAC

Voir contrôle d'accès basé sur des étiquettes.

principe de moindre privilège

Bonne pratique de sécurité qui consiste à accorder les autorisations minimales nécessaires à l'exécution d'une tâche. Pour plus d'informations, veuillez consulter la rubrique Accorder les autorisations de moindre privilège dans la documentation IAM.

lift and shift

Voir 7 Rs.

système de poids faible

Système qui stocke d'abord l'octet le moins significatif. Voir aussi endianité.

LLM

Voir le grand modèle de langage.

environnements inférieurs

Voir environnement.

M

machine learning (ML)

Type d'intelligence artificielle qui utilise des algorithmes et des techniques pour la reconnaissance et l'apprentissage de modèles. Le ML analyse et apprend à partir de données enregistrées, telles que les données de l'Internet des objets (IoT), pour générer un modèle statistique basé sur des modèles. Pour plus d'informations, veuillez consulter Machine Learning.

branche principale

Voir succursale.

malware

Logiciel conçu pour compromettre la sécurité ou la confidentialité de l'ordinateur. Les logiciels malveillants peuvent perturber les systèmes informatiques, divulguer des informations sensibles ou obtenir un accès non autorisé. Parmi les malwares, on peut citer les virus, les vers, les rançongiciels, les chevaux de Troie, les logiciels espions et les enregistreurs de frappe.

services gérés

Services AWS pour lequel AWS fonctionnent la couche d'infrastructure, le système d'exploitation et les plateformes, et vous accédez aux points de terminaison pour stocker et récupérer des données. Amazon Simple Storage Service (Amazon S3) et Amazon DynamoDB sont des exemples de services gérés. Ils sont également connus sous le nom de services abstraits.

système d'exécution de la fabrication (MES)

Un système logiciel pour le suivi, la surveillance, la documentation et le contrôle des processus de production qui convertissent les matières premières en produits finis dans l'atelier.

MAP

Voir Migration Acceleration Program.

mécanisme

Processus complet au cours duquel vous créez un outil, favorisez son adoption, puis inspectez les résultats afin de procéder aux ajustements nécessaires. Un mécanisme est un cycle qui se renforce et s'améliore au fur et à mesure de son fonctionnement. Pour plus d'informations, voir Création de mécanismes dans le cadre AWS Well-Architected.

compte membre

Tous, à l' Comptes AWS exception du compte de gestion, qui font partie d'une organisation dans AWS Organizations. Un compte ne peut être membre que d'une seule organisation à la fois.

MAILLES

Voir le système d'exécution de la fabrication.

Transport télémétrique en file d'attente de messages (MQTT)

Protocole de communication léger machine-to-machine (M2M), basé sur le modèle de publication/d'abonnement, pour les appareils IoT aux ressources limitées.

microservice

Un petit service indépendant qui communique via un réseau bien défini APIs et qui est généralement détenu par de petites équipes autonomes. Par exemple, un système d'assurance peut inclure des microservices qui mappent à des capacités métier, telles que les ventes ou le marketing, ou à des sous-domaines, tels que les achats, les réclamations ou l'analytique. Les avantages des microservices incluent l'agilité, la flexibilité de la mise à l'échelle, la facilité de déploiement, la réutilisation du code et la résilience. Pour plus d'informations, consultez la section Intégration de microservices à l'aide de services AWS sans serveur.

architecture de microservices

Approche de création d'une application avec des composants indépendants qui exécutent chaque processus d'application en tant que microservice. Ces microservices communiquent via une interface bien définie en utilisant Lightweight. APIs Chaque microservice de cette architecture peut être mis à jour, déployé et mis à l'échelle pour répondre à la demande de fonctions spécifiques d'une application. Pour plus d'informations, consultez la section Implémentation de microservices sur AWS.

Programme d'accélération des migrations (MAP)

Un AWS programme qui fournit un support de conseil, des formations et des services pour aider les entreprises à établir une base opérationnelle solide pour passer au cloud, et pour aider à compenser le coût initial des migrations. MAP inclut une méthodologie de migration pour exécuter les migrations héritées de manière méthodique, ainsi qu'un ensemble d'outils pour automatiser et accélérer les scénarios de migration courants.

migration à grande échelle

Processus consistant à transférer la majeure partie du portefeuille d'applications vers le cloud par vagues, un plus grand nombre d'applications étant déplacées plus rapidement à chaque vague. Cette phase utilise les bonnes pratiques et les enseignements tirés des phases précédentes pour implémenter une usine de migration d'équipes, d'outils et de processus en vue de rationaliser la migration des charges de travail grâce à l'automatisation et à la livraison agile. Il s'agit de la troisième phase de la stratégie de migration AWS.

usine de migration

Équipes interfonctionnelles qui rationalisent la migration des charges de travail grâce à des approches automatisées et agiles. Les équipes de Migration Factory comprennent généralement des responsables des opérations, des analystes commerciaux et des propriétaires, des ingénieurs de migration, des développeurs et DevOps des professionnels travaillant dans le cadre de sprints. Entre 20 et 50 % du portefeuille d'applications d'entreprise est constitué de modèles répétés qui peuvent être optimisés par une approche d'usine. Pour plus d'informations, veuillez consulter la rubrique discussion of migration factories et le guide Cloud Migration Factory dans cet ensemble de contenus.

métadonnées de migration

Informations relatives à l'application et au serveur nécessaires pour finaliser la migration. Chaque modèle de migration nécessite un ensemble de métadonnées de migration différent. Les exemples de métadonnées de migration incluent le sous-réseau cible, le groupe de sécurité et le AWS compte.

modèle de migration

Tâche de migration reproductible qui détaille la stratégie de migration, la destination de la migration et l'application ou le service de migration utilisé. Exemple : réorganisez la migration vers Amazon EC2 avec le service de migration AWS d'applications.

Évaluation du portefeuille de migration (MPA)

Outil en ligne qui fournit des informations pour valider l'analyse de rentabilisation en faveur de la migration vers le. AWS Cloud La MPA propose une évaluation détaillée du portefeuille (dimensionnement approprié des serveurs, tarification, comparaison du coût total de possession, analyse des coûts de migration), ainsi que la planification de la migration (analyse et collecte des données d'applications, regroupement des applications, priorisation des migrations et planification des vagues). L'outil MPA (connexion requise) est disponible gratuitement pour tous les AWS consultants et consultants APN Partner.

Évaluation de la préparation à la migration (MRA)

Processus qui consiste à obtenir des informations sur l'état de préparation d'une organisation au cloud, à identifier les forces et les faiblesses et à élaborer un plan d'action pour combler les lacunes identifiées, à l'aide du AWS CAF. Pour plus d'informations, veuillez consulter le guide de préparation à la migration. La MRA est la première phase de la stratégie de migration AWS.

stratégie de migration

L'approche utilisée pour migrer une charge de travail vers le AWS Cloud. Pour plus d'informations, reportez-vous aux 7 R de ce glossaire et à Mobiliser votre organisation pour accélérer les migrations à grande échelle.

ML

Voir apprentissage automatique.

modernisation

Transformation d'une application obsolète (héritée ou monolithique) et de son infrastructure en un système agile, élastique et hautement disponible dans le cloud afin de réduire les coûts, de gagner en efficacité et de tirer parti des innovations. Pour plus d'informations, consultez la section Stratégie de modernisation des applications dans le AWS Cloud.

évaluation de la préparation à la modernisation

Évaluation qui permet de déterminer si les applications d'une organisation sont prêtes à être modernisées, d'identifier les avantages, les risques et les dépendances, et qui détermine dans quelle mesure l'organisation peut prendre en charge l'état futur de ces applications. Le résultat de l'évaluation est un plan de l'architecture cible, une feuille de route détaillant les phases de développement et les étapes du processus de modernisation, ainsi qu'un plan d'action pour combler les lacunes identifiées. Pour plus d'informations, consultez la section Évaluation de l'état de préparation à la modernisation des applications dans le AWS Cloud.

applications monolithiques (monolithes)

Applications qui s'exécutent en tant que service unique avec des processus étroitement couplés. Les applications monolithiques ont plusieurs inconvénients. Si une fonctionnalité de l'application connaît un pic de demande, l'architecture entière doit être mise à l'échelle. L'ajout ou l'amélioration des fonctionnalités d'une application monolithique devient également plus complexe lorsque la base de code s'élargit. Pour résoudre ces problèmes, vous pouvez utiliser une architecture de microservices. Pour plus d'informations, veuillez consulter Decomposing monoliths into microservices.

MPA

Voir Évaluation du portefeuille de migration.

MQTT

Voir Message Queuing Telemetry Transport.

classification multi-classes

Processus qui permet de générer des prédictions pour plusieurs classes (prédiction d'un résultat parmi plus de deux). Par exemple, un modèle de ML peut demander « Ce produit est-il un livre, une voiture ou un téléphone ? » ou « Quelle catégorie de produits intéresse le plus ce client ? ».

infrastructure mutable

Modèle qui met à jour et modifie l'infrastructure existante pour les charges de travail de production. Pour améliorer la cohérence, la fiabilité et la prévisibilité, le AWS Well-Architected Framework recommande l'utilisation d'une infrastructure immuable comme meilleure pratique.

O

OAC

Voir Contrôle d'accès à l'origine.

OAI

Voir l'identité d'accès à l'origine.

OCM

Voir gestion du changement organisationnel.

migration hors ligne

Méthode de migration dans laquelle la charge de travail source est supprimée au cours du processus de migration. Cette méthode implique un temps d'arrêt prolongé et est généralement utilisée pour de petites charges de travail non critiques.

OI

Voir Intégration des opérations.

OLA

Voir l'accord au niveau opérationnel.

migration en ligne

Méthode de migration dans laquelle la charge de travail source est copiée sur le système cible sans être mise hors ligne. Les applications connectées à la charge de travail peuvent continuer à fonctionner pendant la migration. Cette méthode implique un temps d'arrêt nul ou minimal et est généralement utilisée pour les charges de travail de production critiques.

OPC-UA

Voir Open Process Communications - Architecture unifiée.

Communications par processus ouvert - Architecture unifiée (OPC-UA)

Un protocole de communication machine-to-machine (M2M) pour l'automatisation industrielle. L'OPC-UA fournit une norme d'interopérabilité avec des schémas de cryptage, d'authentification et d'autorisation des données.

accord au niveau opérationnel (OLA)

Accord qui précise ce que les groupes informatiques fonctionnels s'engagent à fournir les uns aux autres, afin de prendre en charge un contrat de niveau de service (SLA).

examen de l'état de préparation opérationnelle (ORR)

Une liste de questions et de bonnes pratiques associées qui vous aident à comprendre, évaluer, prévenir ou réduire l'ampleur des incidents et des défaillances possibles. Pour plus d'informations, voir Operational Readiness Reviews (ORR) dans le AWS Well-Architected Framework.

technologie opérationnelle (OT)

Systèmes matériels et logiciels qui fonctionnent avec l'environnement physique pour contrôler les opérations, les équipements et les infrastructures industriels. Dans le secteur manufacturier, l'intégration des systèmes OT et des technologies de l'information (IT) est au cœur des transformations de l'industrie 4.0.

intégration des opérations (OI)

Processus de modernisation des opérations dans le cloud, qui implique la planification de la préparation, l'automatisation et l'intégration. Pour en savoir plus, veuillez consulter le guide d'intégration des opérations.

journal de suivi d'organisation

Un parcours créé par AWS CloudTrail qui enregistre tous les événements pour tous les membres Comptes AWS d'une organisation dans AWS Organizations. Ce journal de suivi est créé dans

chaque Compte AWS qui fait partie de l'organisation et suit l'activité de chaque compte. Pour plus d'informations, consultez la section Création d'un suivi pour une organisation dans la CloudTrail documentation.

gestion du changement organisationnel (OCM)

Cadre pour gérer les transformations métier majeures et perturbatrices du point de vue des personnes, de la culture et du leadership. L'OCM aide les organisations à se préparer et à effectuer la transition vers de nouveaux systèmes et de nouvelles politiques en accélérant l'adoption des changements, en abordant les problèmes de transition et en favorisant des changements culturels et organisationnels. Dans la stratégie de AWS migration, ce cadre est appelé accélération du personnel, en raison de la rapidité du changement requise dans les projets d'adoption du cloud. Pour plus d'informations, veuillez consulter le guide OCM.

contrôle d'accès d'origine (OAC)

Dans CloudFront, une option améliorée pour restreindre l'accès afin de sécuriser votre contenu Amazon Simple Storage Service (Amazon S3). L'OAC prend en charge tous les compartiments S3 dans leur ensemble Régions AWS, le chiffrement côté serveur avec AWS KMS (SSE-KMS) et les requêtes dynamiques `PUT` adressées au compartiment S3. `DELETE`

identité d'accès d'origine (OAI)

Dans CloudFront, une option permettant de restreindre l'accès afin de sécuriser votre contenu Amazon S3. Lorsque vous utilisez OAI, il CloudFront crée un principal auprès duquel Amazon S3 peut s'authentifier. Les principaux authentifiés ne peuvent accéder au contenu d'un compartiment S3 que par le biais d'une distribution spécifique CloudFront . Voir également OAC, qui fournit un contrôle d'accès plus précis et amélioré.

ORR

Voir l'examen de l'état de préparation opérationnelle.

DE

Voir technologie opérationnelle.

VPC sortant (de sortie)

Dans une architecture AWS multi-comptes, un VPC qui gère les connexions réseau initiées depuis une application. L'architecture AWS de référence de sécurité recommande de configurer votre compte réseau avec les fonctions entrantes, sortantes et d'inspection VPCs afin de protéger l'interface bidirectionnelle entre votre application et l'Internet en général.

P

limite des autorisations

Politique de gestion IAM attachée aux principaux IAM pour définir les autorisations maximales que peut avoir l'utilisateur ou le rôle. Pour plus d'informations, veuillez consulter la rubrique Limites des autorisations dans la documentation IAM.

informations personnelles identifiables (PII)

Informations qui, lorsqu'elles sont consultées directement ou associées à d'autres données connexes, peuvent être utilisées pour déduire raisonnablement l'identité d'une personne. Les exemples d'informations personnelles incluent les noms, les adresses et les informations de contact.

PII

Voir les informations personnelles identifiables.

manuel stratégique

Ensemble d'étapes prédéfinies qui capturent le travail associé aux migrations, comme la fourniture de fonctions d'opérations de base dans le cloud. Un manuel stratégique peut revêtir la forme de scripts, de runbooks automatisés ou d'un résumé des processus ou des étapes nécessaires au fonctionnement de votre environnement modernisé.

PLC

Voir contrôleur logique programmable.

PLM

Consultez la section Gestion du cycle de vie des produits.

politique

Objet capable de définir les autorisations (voir la politique basée sur l'identité), de spécifier les conditions d'accès (voir la politique basée sur les ressources) ou de définir les autorisations maximales pour tous les comptes d'une organisation dans AWS Organizations (voir la politique de contrôle des services).

persistance polyglotte

Choix indépendant de la technologie de stockage de données d'un microservice en fonction des modèles d'accès aux données et d'autres exigences. Si vos microservices utilisent la même

technologie de stockage de données, ils peuvent rencontrer des difficultés d'implémentation ou présenter des performances médiocres. Les microservices sont plus faciles à mettre en œuvre, atteignent de meilleures performances, ainsi qu'une meilleure capacité de mise à l'échelle s'ils utilisent l'entrepôt de données le mieux adapté à leurs besoins. Pour plus d'informations, veuillez consulter Enabling data persistence in microservices.

évaluation du portefeuille

Processus de découverte, d'analyse et de priorisation du portefeuille d'applications afin de planifier la migration. Pour plus d'informations, veuillez consulter Evaluating migration readiness.

predicate

Une condition de requête qui renvoie `true` ou`false`, généralement située dans une `WHERE` clause.

prédicat pushdown

Technique d'optimisation des requêtes de base de données qui filtre les données de la requête avant le transfert. Cela réduit la quantité de données qui doivent être extraites et traitées à partir de la base de données relationnelle et améliore les performances des requêtes.

contrôle préventif

Contrôle de sécurité conçu pour empêcher qu'un événement ne se produise. Ces contrôles constituent une première ligne de défense pour empêcher tout accès non autorisé ou toute modification indésirable de votre réseau. Pour plus d'informations, veuillez consulter Preventative controls dans Implementing security controls on AWS.

principal

Entité capable d'effectuer AWS des actions et d'accéder à des ressources. Cette entité est généralement un utilisateur root pour un Compte AWS rôle IAM ou un utilisateur. Pour plus d'informations, veuillez consulter la rubrique Principal dans Termes et concepts relatifs aux rôles, dans la documentation IAM.

confidentialité dès la conception

Une approche d'ingénierie système qui prend en compte la confidentialité tout au long du processus de développement.

zones hébergées privées

Conteneur contenant des informations sur la manière dont vous souhaitez qu'Amazon Route 53 réponde aux requêtes DNS pour un domaine et ses sous-domaines au sein d'un ou de plusieurs

VPCs domaines. Pour plus d'informations, veuillez consulter Working with private hosted zones dans la documentation Route 53.

contrôle proactif

Contrôle de sécurité conçu pour empêcher le déploiement de ressources non conformes. Ces contrôles analysent les ressources avant qu'elles ne soient provisionnées. Si la ressource n'est pas conforme au contrôle, elle n'est pas provisionnée. Pour plus d'informations, consultez le guide de référence sur les contrôles dans la AWS Control Tower documentation et consultez la section Contrôles proactifs dans Implémentation des contrôles de sécurité sur AWS.

gestion du cycle de vie des produits (PLM)

Gestion des données et des processus d'un produit tout au long de son cycle de vie, depuis la conception, le développement et le lancement, en passant par la croissance et la maturité, jusqu'au déclin et au retrait.

environnement de production

Voir environnement.

contrôleur logique programmable (PLC)

Dans le secteur manufacturier, un ordinateur hautement fiable et adaptable qui surveille les machines et automatise les processus de fabrication.

chaînage rapide

Utiliser le résultat d'une invite LLM comme entrée pour l'invite suivante afin de générer de meilleures réponses. Cette technique est utilisée pour décomposer une tâche complexe en sous-tâches ou pour affiner ou développer de manière itérative une réponse préliminaire. Cela permet d'améliorer la précision et la pertinence des réponses d'un modèle et permet d'obtenir des résultats plus précis et personnalisés.

pseudonymisation

Processus de remplacement des identifiants personnels dans un ensemble de données par des valeurs fictives. La pseudonymisation peut contribuer à protéger la vie privée. Les données pseudonymisées sont toujours considérées comme des données personnelles.

publish/subscribe (pub/sub)

Modèle qui permet des communications asynchrones entre les microservices afin d'améliorer l'évolutivité et la réactivité. Par exemple, dans un MES basé sur des microservices, un microservice peut publier des messages d'événements sur un canal auquel d'autres microservices

peuvent s'abonner. Le système peut ajouter de nouveaux microservices sans modifier le service de publication.

Q

plan de requête

Série d'étapes, telles que des instructions, utilisées pour accéder aux données d'un système de base de données relationnelle SQL.

régression du plan de requêtes

Le cas où un optimiseur de service de base de données choisit un plan moins optimal qu'avant une modification donnée de l'environnement de base de données. Cela peut être dû à des changements en termes de statistiques, de contraintes, de paramètres d'environnement, de liaisons de paramètres de requêtes et de mises à jour du moteur de base de données.

R

Matrice RACI

Voir responsable, responsable, consulté, informé (RACI).

CHIFFON

Voir Retrieval Augmented Generation.

rançongiciel

Logiciel malveillant conçu pour bloquer l'accès à un système informatique ou à des données jusqu'à ce qu'un paiement soit effectué.

Matrice RASCI

Voir responsable, responsable, consulté, informé (RACI).

RCAC

Voir contrôle d'accès aux lignes et aux colonnes.

réplica en lecture

Copie d'une base de données utilisée en lecture seule. Vous pouvez acheminer les requêtes vers le réplica de lecture pour réduire la charge sur votre base de données principale.

réarchitecte

Voir 7 Rs.

objectif de point de récupération (RPO)

Durée maximale acceptable depuis le dernier point de récupération des données. Il détermine ce qui est considéré comme étant une perte de données acceptable entre le dernier point de reprise et l'interruption du service.

objectif de temps de récupération (RTO)

Le délai maximum acceptable entre l'interruption du service et le rétablissement du service.

refactoriser

Voir 7 Rs.

Région

Un ensemble de AWS ressources dans une zone géographique. Chacun Région AWS est isolé et indépendant des autres pour garantir tolérance aux pannes, stabilité et résilience. Pour plus d'informations, voir Spécifier ce que Régions AWS votre compte peut utiliser.

régression

Technique de ML qui prédit une valeur numérique. Par exemple, pour résoudre le problème « Quel sera le prix de vente de cette maison ? », un modèle de ML pourrait utiliser un modèle de régression linéaire pour prédire le prix de vente d'une maison sur la base de faits connus à son sujet (par exemple, la superficie en mètres carrés).

réhéberger

Voir 7 Rs.

version

Dans un processus de déploiement, action visant à promouvoir les modifications apportées à un environnement de production.

déplacer

Voir 7 Rs.

replateforme

Voir 7 Rs.

rachat

Voir 7 Rs.

résilience

La capacité d'une application à résister aux perturbations ou à s'en remettre. La haute disponibilité et la reprise après sinistre sont des considérations courantes lors de la planification de la résilience dans le AWS Cloud. Pour plus d'informations, consultez la section AWS Cloud Résilience.

politique basée sur les ressources

Politique attachée à une ressource, comme un compartiment Amazon S3, un point de terminaison ou une clé de chiffrement. Ce type de politique précise les principaux auxquels l'accès est autorisé, les actions prises en charge et toutes les autres conditions qui doivent être remplies.

matrice responsable, redevable, consulté et informé (RACI)

Une matrice qui définit les rôles et les responsabilités de toutes les parties impliquées dans les activités de migration et les opérations cloud. Le nom de la matrice est dérivé des types de responsabilité définis dans la matrice : responsable (R), responsable (A), consulté (C) et informé (I). Le type de support (S) est facultatif. Si vous incluez le support, la matrice est appelée matrice RASCI, et si vous l'excluez, elle est appelée matrice RACI.

contrôle réactif

Contrôle de sécurité conçu pour permettre de remédier aux événements indésirables ou aux écarts par rapport à votre référence de sécurité. Pour plus d'informations, veuillez consulter la rubrique Responsive controls dans Implementing security controls on AWS.

retain

Voir 7 Rs.

se retirer

Voir 7 Rs.

Génération augmentée de récupération (RAG)

Technologie d'IA générative dans laquelle un LLM fait référence à une source de données faisant autorité qui se trouve en dehors de ses sources de données de formation avant de générer une

réponse. Par exemple, un modèle RAG peut effectuer une recherche sémantique dans la base de connaissances ou dans les données personnalisées d'une organisation. Pour plus d'informations, voir Qu'est-ce que RAG ?

rotation

Processus de mise à jour périodique d'un secret pour empêcher un attaquant d'accéder aux informations d'identification.

contrôle d'accès aux lignes et aux colonnes (RCAC)

Utilisation d'expressions SQL simples et flexibles dotées de règles d'accès définies. Le RCAC comprend des autorisations de ligne et des masques de colonnes.

RPO

Voir l'objectif du point de récupération.

RTO

Voir l'objectif relatif au temps de rétablissement.

runbook

Ensemble de procédures manuelles ou automatisées nécessaires à l'exécution d'une tâche spécifique. Elles visent généralement à rationaliser les opérations ou les procédures répétitives présentant des taux d'erreur élevés.

S

SAML 2.0

Un standard ouvert utilisé par de nombreux fournisseurs d'identité (IdPs). Cette fonctionnalité permet l'authentification unique fédérée (SSO), afin que les utilisateurs puissent se connecter AWS Management Console ou appeler les opérations de l' AWS API sans que vous ayez à créer un utilisateur dans IAM pour tous les membres de votre organisation. Pour plus d'informations sur la fédération SAML 2.0, veuillez consulter À propos de la fédération SAML 2.0 dans la documentation IAM.

SCADA

Voir Contrôle de supervision et acquisition de données.

SCP

Voir la politique de contrôle des services.

secret

Dans AWS Secrets Manager des informations confidentielles ou restreintes, telles qu'un mot de passe ou des informations d'identification utilisateur, que vous stockez sous forme cryptée. Il comprend la valeur secrète et ses métadonnées. La valeur secrète peut être binaire, une chaîne unique ou plusieurs chaînes. Pour plus d'informations, voir Que contient le secret d'un Secrets Manager ? dans la documentation de Secrets Manager.

sécurité dès la conception

Une approche d'ingénierie système qui prend en compte la sécurité tout au long du processus de développement.

contrôle de sécurité

Barrière de protection technique ou administrative qui empêche, détecte ou réduit la capacité d'un assaillant d'exploiter une vulnérabilité de sécurité. Il existe quatre principaux types de contrôles de sécurité : préventifs, détectifs, réactifs et proactifs.

renforcement de la sécurité

Processus qui consiste à réduire la surface d'attaque pour la rendre plus résistante aux attaques. Cela peut inclure des actions telles que la suppression de ressources qui ne sont plus requises, la mise en œuvre des bonnes pratiques de sécurité consistant à accorder le moindre privilège ou la désactivation de fonctionnalités inutiles dans les fichiers de configuration.

système de gestion des informations et des événements de sécurité (SIEM)

Outils et services qui associent les systèmes de gestion des informations de sécurité (SIM) et de gestion des événements de sécurité (SEM). Un système SIEM collecte, surveille et analyse les données provenant de serveurs, de réseaux, d'appareils et d'autres sources afin de détecter les menaces et les failles de sécurité, mais aussi de générer des alertes.

automatisation des réponses de sécurité

Action prédéfinie et programmée conçue pour répondre automatiquement à un événement de sécurité ou y remédier. Ces automatisations servent de contrôles de sécurité détectifs ou réactifs qui vous aident à mettre en œuvre les meilleures pratiques AWS de sécurité. Parmi les actions de réponse automatique, citons la modification d'un groupe de sécurité VPC, l'application de correctifs à une EC2 instance Amazon ou la rotation des informations d'identification.

chiffrement côté serveur

Chiffrement des données à destination, par celui Service AWS qui les reçoit.

Politique de contrôle des services (SCP)

Politique qui fournit un contrôle centralisé des autorisations pour tous les comptes d'une organisation dans AWS Organizations. SCPs définissez des garde-fous ou des limites aux actions qu'un administrateur peut déléguer à des utilisateurs ou à des rôles. Vous pouvez les utiliser SCPs comme listes d'autorisation ou de refus pour spécifier les services ou les actions autorisés ou interdits. Pour plus d'informations, consultez la section Politiques de contrôle des services dans la AWS Organizations documentation.

point de terminaison du service

URL du point d'entrée pour un Service AWS. Pour vous connecter par programmation au service cible, vous pouvez utiliser un point de terminaison. Pour plus d'informations, veuillez consulter la rubrique Service AWS endpoints dans Références générales AWS.

contrat de niveau de service (SLA)

Accord qui précise ce qu'une équipe informatique promet de fournir à ses clients, comme le temps de disponibilité et les performances des services.

indicateur de niveau de service (SLI)

Mesure d'un aspect des performances d'un service, tel que son taux d'erreur, sa disponibilité ou son débit.

objectif de niveau de service (SLO)

Mesure cible qui représente l'état d'un service, tel que mesuré par un indicateur de niveau de service.

modèle de responsabilité partagée

Un modèle décrivant la responsabilité que vous partagez en matière AWS de sécurité et de conformité dans le cloud. AWS est responsable de la sécurité du cloud, alors que vous êtes responsable de la sécurité dans le cloud. Pour de plus amples informations, veuillez consulter Modèle de responsabilité partagée.

SIEM

Consultez les informations de sécurité et le système de gestion des événements.

point de défaillance unique (SPOF)

Défaillance d'un seul composant critique d'une application susceptible de perturber le système.

SLA

Voir le contrat de niveau de service.

SLI

Voir l'indicateur de niveau de service.

SLO

Voir l'objectif de niveau de service.

split-and-seed modèle

Modèle permettant de mettre à l'échelle et d'accélérer les projets de modernisation. Au fur et à mesure que les nouvelles fonctionnalités et les nouvelles versions de produits sont définies, l'équipe principale se divise pour créer des équipes de produit. Cela permet de mettre à l'échelle les capacités et les services de votre organisation, d'améliorer la productivité des développeurs et de favoriser une innovation rapide. Pour plus d'informations, consultez la section Approche progressive de la modernisation des applications dans le. AWS Cloud

SPOF

Voir point de défaillance unique.

schéma en étoile

Structure organisationnelle de base de données qui utilise une grande table de faits pour stocker les données transactionnelles ou mesurées et utilise une ou plusieurs tables dimensionnelles plus petites pour stocker les attributs des données. Cette structure est conçue pour être utilisée dans un entrepôt de données ou à des fins de business intelligence.

modèle de figuier étrangleur

Approche de modernisation des systèmes monolithiques en réécrivant et en remplaçant progressivement les fonctionnalités du système jusqu'à ce que le système hérité puisse être mis hors service. Ce modèle utilise l'analogie d'un figuier de vigne qui se développe dans un arbre existant et qui finit par supplanter son hôte. Le schéma a été présenté par Martin Fowler comme un moyen de gérer les risques lors de la réécriture de systèmes monolithiques. Pour obtenir un

exemple d'application de ce modèle, veuillez consulter Modernizing legacy Microsoft ASP.NET (ASMX) web services incrementally by using containers and Amazon API Gateway.

sous-réseau

Plage d'adresses IP dans votre VPC. Un sous-réseau doit se trouver dans une seule zone de disponibilité.

contrôle de supervision et acquisition de données (SCADA)

Dans le secteur manufacturier, un système qui utilise du matériel et des logiciels pour surveiller les actifs physiques et les opérations de production.

chiffrement symétrique

Algorithme de chiffrement qui utilise la même clé pour chiffrer et déchiffrer les données.

tests synthétiques

Tester un système de manière à simuler les interactions des utilisateurs afin de détecter les problèmes potentiels ou de surveiller les performances. Vous pouvez utiliser Amazon CloudWatch Synthetics pour créer ces tests.

invite du système

Technique permettant de fournir un contexte, des instructions ou des directives à un LLM afin d'orienter son comportement. Les instructions du système aident à définir le contexte et à établir des règles pour les interactions avec les utilisateurs.

T

balises

Des paires clé-valeur qui agissent comme des métadonnées pour organiser vos AWS ressources. Les balises peuvent vous aider à gérer, identifier, organiser, rechercher et filtrer des ressources. Pour plus d'informations, veuillez consulter la rubrique Balisage de vos AWS ressources.

variable cible

La valeur que vous essayez de prédire dans le cadre du ML supervisé. Elle est également qualifiée de variable de résultat. Par exemple, dans un environnement de fabrication, la variable cible peut être un défaut du produit.

liste de tâches

Outil utilisé pour suivre les progrès dans un runbook. Liste de tâches qui contient une vue d'ensemble du runbook et une liste des tâches générales à effectuer. Pour chaque tâche générale, elle inclut le temps estimé nécessaire, le propriétaire et l'avancement.

environnement de test

Voir environnement.

entraînement

Pour fournir des données à partir desquelles votre modèle de ML peut apprendre. Les données d'entraînement doivent contenir la bonne réponse. L'algorithme d'apprentissage identifie des modèles dans les données d'entraînement, qui mettent en correspondance les attributs des données d'entrée avec la cible (la réponse que vous souhaitez prédire). Il fournit un modèle de ML qui capture ces modèles. Vous pouvez alors utiliser le modèle de ML pour obtenir des prédictions sur de nouvelles données pour lesquelles vous ne connaissez pas la cible.

passerelle de transit

Un hub de transit réseau que vous pouvez utiliser pour interconnecter vos réseaux VPCs et ceux sur site. Pour plus d'informations, voir Qu'est-ce qu'une passerelle de transit dans la AWS Transit Gateway documentation.

flux de travail basé sur jonction

Approche selon laquelle les développeurs génèrent et testent des fonctionnalités localement dans une branche de fonctionnalités, puis fusionnent ces modifications dans la branche principale. La branche principale est ensuite intégrée aux environnements de développement, de préproduction et de production, de manière séquentielle.

accès sécurisé

Accorder des autorisations à un service que vous spécifiez pour effectuer des tâches au sein de votre organisation AWS Organizations et dans ses comptes en votre nom. Le service de confiance crée un rôle lié au service dans chaque compte, lorsque ce rôle est nécessaire, pour effectuer des tâches de gestion à votre place. Pour plus d'informations, consultez la section Utilisation AWS Organizations avec d'autres AWS services dans la AWS Organizations documentation.

réglage

Pour modifier certains aspects de votre processus d'entraînement afin d'améliorer la précision du modèle de ML. Par exemple, vous pouvez entraîner le modèle de ML en générant un ensemble d'étiquetage, en ajoutant des étiquettes, puis en répétant ces étapes plusieurs fois avec différents paramètres pour optimiser le modèle.

équipe de deux pizzas

Une petite DevOps équipe que vous pouvez nourrir avec deux pizzas. Une équipe de deux pizzas garantit les meilleures opportunités de collaboration possible dans le développement de logiciels.

U

incertitude

Un concept qui fait référence à des informations imprécises, incomplètes ou inconnues susceptibles de compromettre la fiabilité des modèles de ML prédictifs. Il existe deux types d'incertitude : l'incertitude épistémique est causée par des données limitées et incomplètes, alors que l'incertitude aléatoire est causée par le bruit et le caractère aléatoire inhérents aux données. Pour plus d'informations, veuillez consulter le guide Quantifying uncertainty in deep learning systems.

tâches indifférenciées

Également connu sous le nom de « levage de charges lourdes », ce travail est nécessaire pour créer et exploiter une application, mais qui n'apporte pas de valeur directe à l'utilisateur final ni d'avantage concurrentiel. Les exemples de tâches indifférenciées incluent l'approvisionnement, la maintenance et la planification des capacités.

environnements supérieurs

Voir environnement.

V

mise à vide

Opération de maintenance de base de données qui implique un nettoyage après des mises à jour incrémentielles afin de récupérer de l'espace de stockage et d'améliorer les performances.

contrôle de version

Processus et outils permettant de suivre les modifications, telles que les modifications apportées au code source dans un référentiel.

Appairage de VPC

Une connexion entre deux VPCs qui vous permet d'acheminer le trafic en utilisant des adresses IP privées. Pour plus d'informations, veuillez consulter la rubrique Qu'est-ce que l'appairage de VPC ? dans la documentation Amazon VPC.

vulnérabilités

Défaut logiciel ou matériel qui compromet la sécurité du système.

W

cache actif

Cache tampon qui contient les données actuelles et pertinentes fréquemment consultées. L'instance de base de données peut lire à partir du cache tampon, ce qui est plus rapide que la lecture à partir de la mémoire principale ou du disque.

données chaudes

Données rarement consultées. Lorsque vous interrogez ce type de données, des requêtes modérément lentes sont généralement acceptables.

fonction de fenêtre

Fonction SQL qui effectue un calcul sur un groupe de lignes liées d'une manière ou d'une autre à l'enregistrement en cours. Les fonctions de fenêtre sont utiles pour traiter des tâches, telles que le calcul d'une moyenne mobile ou l'accès à la valeur des lignes en fonction de la position relative de la ligne en cours.

charge de travail

Ensemble de ressources et de code qui fournit une valeur métier, par exemple une application destinée au client ou un processus de backend.

flux de travail

Groupes fonctionnels d'un projet de migration chargés d'un ensemble de tâches spécifique. Chaque flux de travail est indépendant, mais prend en charge les autres flux de travail du projet.

Par exemple, le flux de travail du portefeuille est chargé de prioriser les applications, de planifier les vagues et de collecter les métadonnées de migration. Le flux de travail du portefeuille fournit ces actifs au flux de travail de migration, qui migre ensuite les serveurs et les applications.

VER

Voir écrire une fois, lire plusieurs.

WQF

Voir le cadre AWS de qualification de la charge de travail.

écrire une fois, lire plusieurs (WORM)

Modèle de stockage qui écrit les données une seule fois et empêche leur suppression ou leur modification. Les utilisateurs autorisés peuvent lire les données autant de fois que nécessaire, mais ils ne peuvent pas les modifier. Cette infrastructure de stockage de données est considérée comme immuable.

Z

exploit Zero-Day

Une attaque, généralement un logiciel malveillant, qui tire parti d'une vulnérabilité de type « jour zéro ».

vulnérabilité « jour zéro »

Une faille ou une vulnérabilité non atténuée dans un système de production. Les acteurs malveillants peuvent utiliser ce type de vulnérabilité pour attaquer le système. Les développeurs prennent souvent conscience de la vulnérabilité à la suite de l'attaque.

invite Zero-Shot

Fournir à un LLM des instructions pour effectuer une tâche, mais aucun exemple (plans) pouvant aider à la guider. Le LLM doit utiliser ses connaissances pré-entraînées pour gérer la tâche. L'efficacité de l'invite zéro dépend de la complexité de la tâche et de la qualité de l'invite. Voir également les instructions en quelques clics.

application zombie

Application dont l'utilisation moyenne du processeur et de la mémoire est inférieure à 5 %. Dans un projet de migration, il est courant de retirer ces applications.

Les traductions sont fournies par des outils de traduction automatique. En cas de conflit entre le contenu d'une traduction et celui de la version originale en anglais, la version anglaise prévaudra.