Creazione di soluzioni Retrieval Augmented Generation per il settore sanitario AWS

AWS Guida prescrittiva

AWS Guida prescrittiva: Creazione di soluzioni Retrieval Augmented Generation per il settore sanitario AWS

Table of Contents

Creazione di soluzioni Retrieval Augmented Generation per il settore sanitario AWS

Amazon Web Services, Accenturee Cadiem (collaboratori)

marzo 2025 (cronologia del documento)

Prima dei modelli linguistici di grandi dimensioni (LLMs) e dell'intelligenza artificiale generativa, il compito di sviluppare applicazioni automatizzate e ad alta precisione nel settore sanitario era impegnativo. I metodi tradizionali si basavano in larga misura sull'immissione e l'analisi manuali dei dati. La complessità dell'analisi delle immagini mediche e delle cartelle cliniche dei pazienti richiedeva un ampio intervento umano, che spesso portava a flussi di lavoro frammentati e inefficienti. Il progresso delle tecnologie di intelligenza artificiale consente di creare applicazioni iperpersonalizzate su larga scala. Le applicazioni sanitarie possono ora integrarsi con le knowledge base mediche, interpretare le immagini diagnostiche con maggiore precisione e prevedere gli esiti dei pazienti utilizzando modelli predittivi.

Questa guida esplora come LLMs stiamo rivoluzionando l'assistenza sanitaria attraverso le applicazioni Retrieval Augmented Generation con cui è possibile creare. Servizi AWSRetrieval Augmented Generation (RAG) è una tecnologia di intelligenza artificiale generativa in cui un LLM fa riferimento a una fonte di dati autorevole esterna alle sue fonti di dati di addestramento prima di generare una risposta. Le applicazioni RAG basano i risultati del modello sulla conoscenza del mondo reale, che riduce le allucinazioni e aumenta la rilevanza della risposta. Nel settore sanitario, i RAG possono essere utilizzati per fornire informazioni up-to-date mediche accurate, garantendo che gli operatori sanitari abbiano accesso alle più recenti ricerche e linee guida cliniche. Trasformando i dati in informazioni fruibili e automatizzando processi complessi, queste tecnologie aiutano a migliorare l'assistenza ai pazienti, a semplificare le operazioni e a migliorare la produttività degli operatori sanitari.

In Amazon Bedrock, puoi ottimizzarli LLMs e integrarli con agenti intelligenti per creare soluzioni sanitarie avanzate. Evidenziando la sinergia tra Amazon OpenSearch Service e Amazon Neptune, la guida dimostra come questi servizi elevano le soluzioni RAG attraverso una maggiore rilevanza della ricerca e un recupero avanzato di dati da più fonti. Puoi orchestrare soluzioni Amazon Bedrock complete che utilizzano agenti Amazon Bedrock e LangChainper coordinare senza interruzioni le interazioni tra diversi repository di dati. Questa integrazione dimostra la potenza della combinazione di servizi specializzati per creare sistemi basati sull'intelligenza artificiale più efficaci ed efficienti.

Cura del paziente e produttività

Questa guida presenta due casi d'uso reali per l'assistenza e la produttività dei pazienti: l'aumento dei dati dei pazienti e la previsione dei rischi di riammissione. Fornisce modelli strategici per l'implementazione di queste soluzioni su larga scala, offrendo alle organizzazioni sanitarie un percorso chiaro verso l'industrializzazione dei processi basati sull'intelligenza artificiale. Grazie a queste informazioni, le istituzioni sanitarie possono utilizzare tecnologie di intelligenza artificiale avanzate per creare flussi di lavoro più efficienti e intelligenti.

Gestione dei talenti

Questa guida delinea anche le strategie per riqualificare e consentire agli operatori sanitari di integrare senza problemi l'intelligenza artificiale generativa nelle loro routine quotidiane. Ciò può migliorare sia la produttività che la qualità dell'assistenza ai pazienti. Dotando la propria forza lavoro delle competenze necessarie per utilizzare efficacemente strumenti avanzati di intelligenza artificiale, le organizzazioni sanitarie possono massimizzare il ritorno sull'investimento e promuovere l'innovazione nella cura dei pazienti.

Questa soluzione di gestione dei talenti basata sull'intelligenza artificiale include le seguenti funzionalità chiave:

- Analisi intelligente dei curriculum dei talenti: utilizzando la tecnologia avanzata LLMs disponibile in Amazon Bedrock, questo strumento estrae e analizza in modo efficiente le competenze e gli attributi critici dei talenti dai curriculum. Questo strumento può semplificare il processo di reclutamento.
- Talent Knowledge Base: basato su Amazon Neptune, questo database dinamico fornisce informazioni in tempo reale sui livelli di personale, sulla distribuzione delle competenze e sulle tendenze del settore. Questo ti aiuta a prendere decisioni basate sui dati sulla gestione della forza lavoro.
- Motore di raccomandazioni per l'apprendimento: questo strumento basato sull'intelligenza artificiale identifica le lacune di competenze all'interno dell'organizzazione e consiglia programmi di formazione personalizzati per il personale medico. Questo strumento promuove lo sviluppo professionale continuo e aiuta la forza lavoro ad adattarsi alle tecnologie sanitarie in evoluzione.

Insieme, queste funzionalità basate sull'intelligenza artificiale aiutano a ottimizzare le prestazioni della forza lavoro, rivoluzionando la gestione dei talenti con maggiore intelligenza ed efficienza.

Opportunità e sfide

Amazon Bedrock può fornire maggiore produttività, scalabilità, economicità e approfondimenti basati sui dati. Amazon Bedrock consente alle organizzazioni sanitarie di utilizzare in LLMs modo efficace diversi casi d'uso, dalla creazione di contenuti all'analisi dei dati fino al processo decisionale automatizzato. Questa guida fornisce approcci per superare le sfide più comuni dell'intelligenza artificiale generativa, come problemi di qualità dei dati, scalabilità dell'infrastruttura, mantenimento delle prestazioni del modello e requisiti di miglioramento continuo durante la transizione dalla dimostrazione di concetto alla produzione.

Opportunità per applicazioni di intelligenza artificiale generativa nel settore sanitario

Il settore sanitario è pronto per un cambiamento trasformativo, guidato dalle opportunità offerte dalle applicazioni di intelligenza artificiale generativa. L'intelligenza artificiale generativa ha il potenziale per migliorare l'assistenza ai pazienti, semplificare le operazioni e accelerare la ricerca medica. Utilizzando modelli di intelligenza artificiale avanzati, gli operatori sanitari possono automatizzare l'aumento delle cartelle cliniche. Le anamnesi complete e complete dei up-to-date pazienti facilitano diagnosi e piani di trattamento più accurati. L'analisi delle immagini basata sull'intelligenza artificiale, ad esempio l'interpretazione di ecografie e altre immagini mediche, può fornire informazioni rapide e precise, riducendo il carico di lavoro dei professionisti medici e minimizzando il rischio di errore umano.

Oltre alla diagnostica e al trattamento, l'intelligenza artificiale generativa può svolgere un ruolo fondamentale nell'analisi predittiva. L'analisi predittiva aiuta le organizzazioni sanitarie ad anticipare gli esiti dei pazienti e a personalizzare di conseguenza i piani di assistenza. Questa tecnologia può anche ottimizzare i processi amministrativi, dalla gestione dei dati dei pazienti alla semplificazione della comunicazione tra fornitori e pazienti. Integrando soluzioni di intelligenza artificiale generativa con i sistemi sanitari esistenti, le istituzioni mediche possono raggiungere una maggiore efficienza, ridurre i costi e, in ultima analisi, fornire cure di qualità superiore. L'integrazione dell'IA con l'assistenza sanitaria non è solo un miglioramento, ma un cambiamento fondamentale verso un'assistenza più intelligente, reattiva e incentrata sul paziente.

Analisi avanzata delle immagini

La combinazione di Amazon Bedrock con archivi di dati, come Amazon Neptune OpenSearch e Amazon Service, può aiutarti ad affrontare le complessità dell'analisi avanzata delle immagini nel settore sanitario. Le soluzioni di recupero delle informazioni possono aumentare il processo di scoperta della malattia e migliorare l'accuratezza dell'interpretazione valutando le immagini diagnostiche e interpretando le ecografie. La soluzione può integrare i dati di valutazione visiva e testuale con la revisione manuale della valutazione del paziente da parte dei medici.

Sfide legate all'industrializzazione delle soluzioni

Gli ostacoli principali da affrontare nell'industrializzazione delle soluzioni di intelligenza artificiale nel settore sanitario sono la qualità e la disponibilità dei dati. I dati sanitari spesso esistono in formati frammentati e incoerenti. Garantire che i modelli di intelligenza artificiale abbiano accesso a dati puliti, strutturati e rappresentativi è fondamentale per mantenere le prestazioni in scenari reali. La scalabilità dell'infrastruttura può diventare una sfida a causa degli ambienti di produzione. Questi ambienti devono gestire grandi volumi di dati dei pazienti in tempo reale, fornendo tempi di risposta rapidi e mantenendo la conformità alle normative sulla privacy dei dati, come l'Health Insurance Portability and Accountability Act (HIPAA). Inoltre, con le informazioni mediche emergenti e i dati dei pazienti che si evolvono nel tempo, i modelli di intelligenza artificiale devono essere riqualificati e aggiornati per rimanere pertinenti e fornire raccomandazioni accurate. Infine, l'integrazione di queste soluzioni di intelligenza artificiale nei sistemi sanitari esistenti può essere complessa a causa di problemi di interoperabilità e della necessità di allineamento con gli attuali flussi di lavoro clinici. Questa integrazione richiede modifiche sia tecniche che operative.

Caso d'uso: creazione di un'applicazione di intelligence medica con dati aumentati sui pazienti

L'intelligenza artificiale generativa può aiutare ad aumentare l'assistenza ai pazienti e la produttività del personale migliorando le funzioni cliniche e amministrative. L'analisi delle immagini basata sull'intelligenza artificiale, come l'interpretazione delle ecografie, accelera i processi diagnostici e migliora la precisione. Può fornire informazioni critiche a supporto di interventi medici tempestivi.

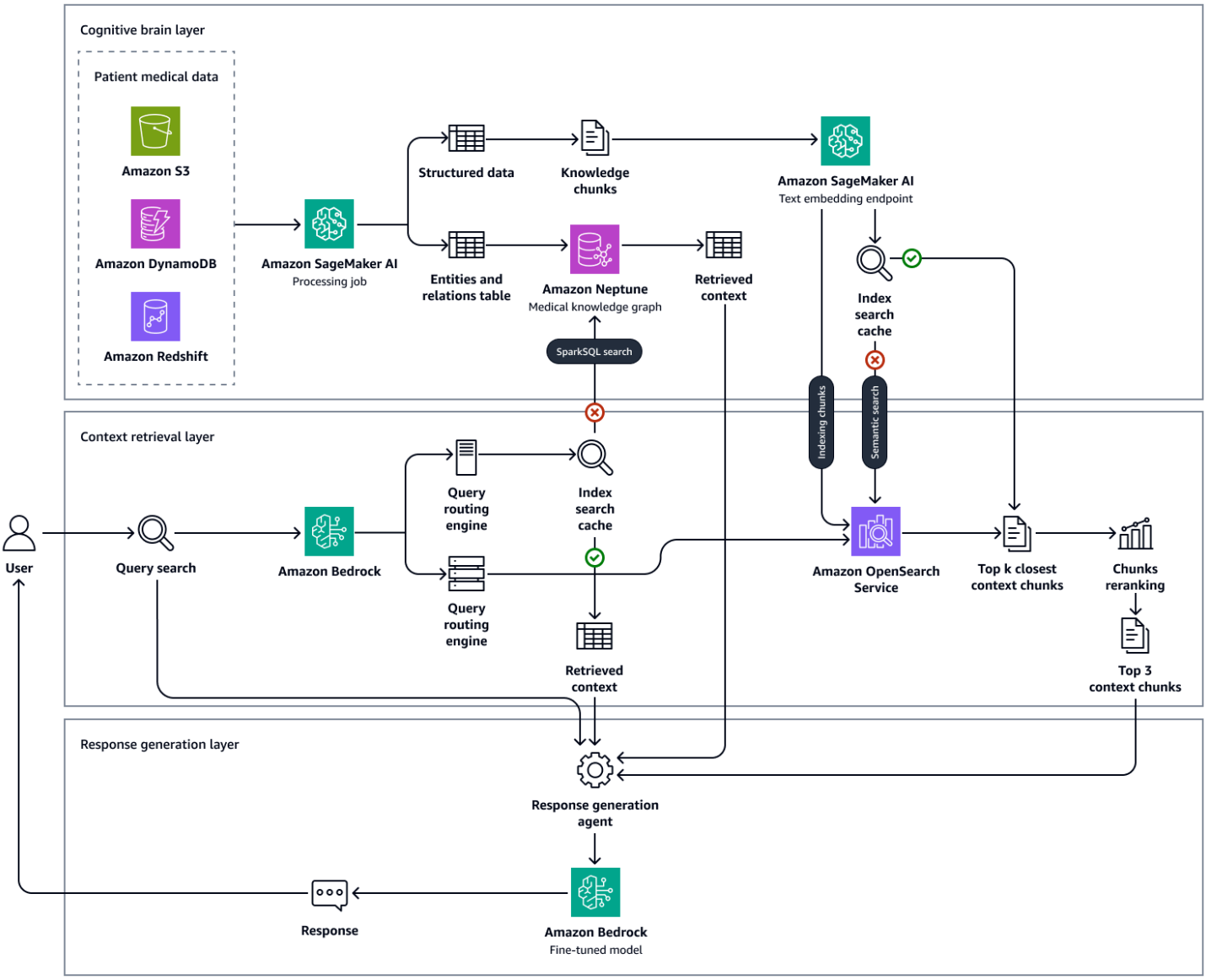
Quando si combinano modelli di intelligenza artificiale generativi con grafici di conoscenza, è possibile automatizzare l'organizzazione cronologica delle cartelle cliniche elettroniche dei pazienti. Ciò consente di integrare i dati in tempo reale provenienti dalle interazioni medico-paziente, dai sintomi, dalle diagnosi, dai risultati di laboratorio e dall'analisi delle immagini. Ciò fornisce al medico dati completi sui pazienti. Questi dati aiutano il medico a prendere decisioni mediche più accurate e tempestive, migliorando sia gli esiti dei pazienti che la produttività degli operatori sanitari.

Panoramica della soluzione

L'intelligenza artificiale può potenziare medici e medici sintetizzando i dati dei pazienti e le conoscenze mediche per fornire informazioni preziose. Questa soluzione Retrieval Augmented Generation (RAG) è un motore di intelligence medica che utilizza un set completo di dati e conoscenze dei pazienti derivanti da milioni di interazioni cliniche. Sfrutta la potenza dell'intelligenza artificiale generativa per creare informazioni basate sull'evidenza per migliorare l'assistenza ai pazienti. È progettato per migliorare i flussi di lavoro clinici, ridurre gli errori e migliorare gli esiti dei pazienti.

La soluzione include una funzionalità di elaborazione delle immagini automatizzata basata su. LLMs Questa funzionalità riduce il tempo che il personale medico deve dedicare manualmente alla ricerca di immagini diagnostiche simili e all'analisi dei risultati diagnostici.

L'immagine seguente mostra i vantaggi end-to-end-workflow di questa soluzione. Utilizza Amazon Neptune, Amazon AI, SageMaker OpenSearch Amazon Service e un modello base in Amazon Bedrock. Per l'agente di recupero del contesto che interagisce con il Medical Knowledge Graph di Neptune, puoi scegliere tra un agente Amazon Bedrock e un LangChain agente.



Nei nostri esperimenti con esempi di domande mediche, abbiamo osservato che le risposte finali generate dal nostro approccio utilizzando il Knowledge Graph gestito in Neptune OpenSearch, il database vettoriale che ospita la knowledge base clinica e Amazon LLMs Bedrock erano fondate sulla fattualità e sono molto più accurate riducendo i falsi positivi e aumentando i veri positivi. Questa soluzione può generare informazioni basate sull'evidenza sullo stato di salute del paziente e mira a migliorare i flussi di lavoro clinici, ridurre gli errori e migliorare gli esiti dei pazienti.

La creazione di questa soluzione prevede i seguenti passaggi:

- Fase 1: Scoperta dei dati
- Fase 2: Creazione di un grafico delle conoscenze mediche

- Fase 3: Creazione di agenti di recupero del contesto per interrogare il grafico delle conoscenze mediche
- Fase 4: creazione di una knowledge base di dati descrittivi in tempo reale
- Fase 5: Utilizzo LLMs per rispondere a domande mediche

Fase 1: Scoperta dei dati

Esistono molti set di dati medici open source che puoi utilizzare per supportare lo sviluppo di una soluzione sanitaria basata sull'intelligenza artificiale. Uno di questi set di dati è il set di dati MIMIC-IV, un set di dati di cartelle cliniche elettroniche (EHR) disponibile al pubblico ampiamente utilizzato nella comunità di ricerca sanitaria. MIMIC-IV contiene informazioni cliniche dettagliate, incluse note di dimissione a testo libero tratte dalle cartelle cliniche dei pazienti. È possibile utilizzare questi record per sperimentare tecniche di sommatoria del testo e di estrazione di entità. Queste tecniche consentono di estrarre informazioni mediche (come i sintomi dei pazienti, i farmaci somministrati e i trattamenti prescritti) da testi non strutturati.

È inoltre possibile utilizzare un set di dati che fornisca riepiloghi delle dimissioni dei pazienti annotati e anonimi, appositamente curati per scopi di ricerca. Un set di dati riepilogativi sulle dimissioni può aiutarvi a sperimentare l'estrazione delle entità, consentendovi di identificare le entità mediche chiave (come condizioni, procedure e farmaci) dal testo. Fase 2: Creazione di un grafico delle conoscenze mediche in questa guida viene descritto come utilizzare i dati strutturati estratti dai set di dati riassuntivi MIMIC-IV e sulle dimissioni per creare un grafico delle conoscenze mediche. Questo grafico della conoscenza medica funge da spina dorsale per sistemi avanzati di interrogazione e supporto decisionale per gli operatori sanitari.

Oltre ai set di dati basati su testo, puoi utilizzare set di dati di immagini. Ad esempio, il set di dati Musculoskeletal Radiographs (MURA), che è un database completo di immagini radiografiche a più viste delle ossa. Utilizzate tali set di dati di immagini per sperimentare la valutazione diagnostica mediante tecniche di decodifica delle immagini mediche. Queste tecniche di decodifica sono fondamentali per la diagnosi precoce di malattie, come le malattie muscoloscheletriche, le malattie cardiovascolari e l'osteoporosi. Ottimizzando i modelli di base della visione e del linguaggio sul set di dati di immagini mediche, è possibile rilevare anomalie nelle immagini diagnostiche. Questo aiuta il sistema a fornire ai medici informazioni diagnostiche tempestive e accurate. Utilizzando set di dati di immagini e testo, è possibile creare un'applicazione sanitaria basata sull'intelligenza artificiale in grado di elaborare dati di testo e immagini per migliorare l'assistenza ai pazienti.

Fase 2: Creazione di un grafico delle conoscenze mediche

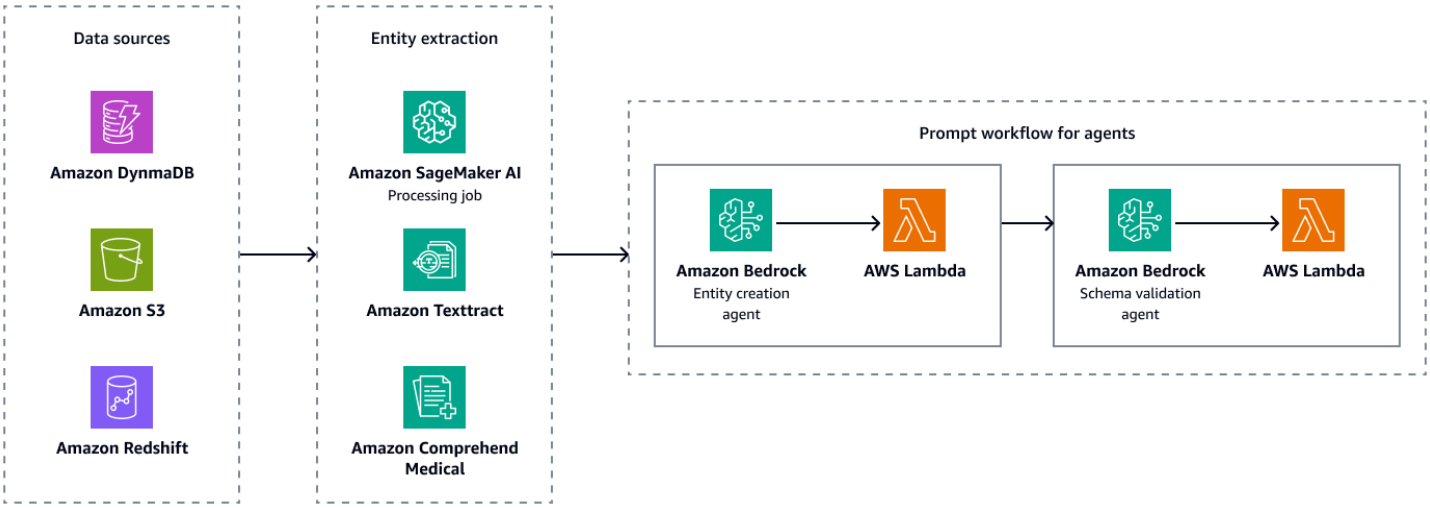
Per qualsiasi organizzazione sanitaria che desideri creare un sistema di supporto decisionale basato su un'enorme base di conoscenze, una sfida fondamentale è individuare ed estrarre le entità mediche presenti nelle note cliniche, nelle riviste mediche, nei riepiloghi delle dimissioni e in altre fonti di dati. È inoltre necessario acquisire le relazioni temporali, i soggetti e le valutazioni di certezza da queste cartelle cliniche per utilizzare efficacemente le entità, gli attributi e le relazioni estratti.

Il primo passo consiste nell'estrarre concetti medici dal testo medico non strutturato utilizzando un prompt in pochi passaggi per un modello base, come Llama 3 in Amazon Bedrock. Il prompt Few-shot si verifica quando si fornisce a un LLM un numero limitato di esempi che dimostrano l'attività e l'output desiderato prima di chiedergli di eseguire un'attività simile. Utilizzando un estrattore di entità mediche basato su LLM, è possibile analizzare il testo medico non strutturato e quindi generare una rappresentazione strutturata dei dati delle entità con conoscenze mediche. È inoltre possibile memorizzare gli attributi del paziente per l'analisi e l'automazione a valle. Il processo di estrazione dell'entità include le seguenti azioni:

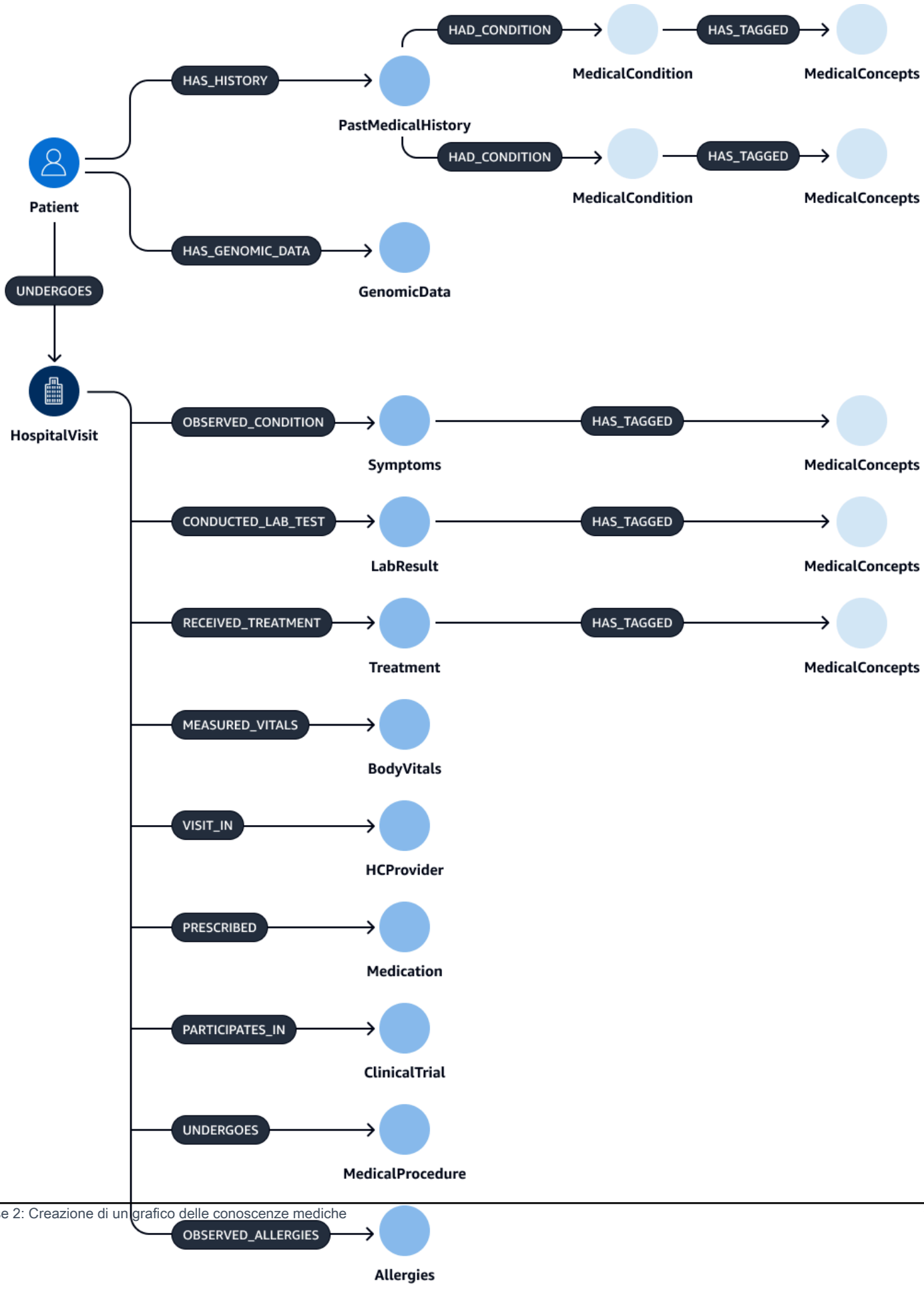
- Estrai informazioni su concetti medici, come malattie, farmaci, dispositivi medici, dosaggio, frequenza dei farmaci, durata del farmaco, sintomi, procedure mediche e relative caratteristiche clinicamente rilevanti.
- Acquisisci le caratteristiche funzionali, come le relazioni temporali tra le entità estratte, i soggetti e le valutazioni di certezza.
- Espandi i vocabolari medici standard, come i seguenti:
  - Identificatori concettuali (rxCUI) dal database RxNorm
  - Codici della classificazione internazionale delle malattie, decima revisione, modifica clinica (ICD-10-CM)
  - Termini tratti da Medical Subject Headings (MeSH)
  - Concetti tratti dalla nomenclatura sistematizzata della medicina, termini clinici (SNOMED CT)
  - Codici del sistema linguistico medico unificato (UMLS)
- Riassumi le note di dimissione e ricava informazioni mediche dalle trascrizioni.

La figura seguente mostra le fasi di estrazione delle entità e di convalida dello schema per creare combinazioni accoppiate valide di entità, attributi e relazioni. Puoi archiviare dati non strutturati, come riepiloghi delle dimissioni o note sui pazienti, in Amazon Simple Storage Service (Amazon S3). Puoi archiviare dati strutturati, come dati di pianificazione delle risorse aziendali (ERP), cartelle

cliniche elettroniche dei pazienti e sistemi informativi di laboratorio, in Amazon Redshift e Amazon DynamoDB. Puoi creare un agente per la creazione di entità Amazon Bedrock. Questo agente può integrare servizi, come le pipeline di estrazione dei dati Amazon SageMaker AI, Amazon Textract e Amazon Comprehend Medical, per estrarre entità, relazioni e attributi dalle fonti di dati strutturate e non strutturate. Infine, utilizzi un agente di convalida dello schema Amazon Bedrock per assicurarti che le entità e le relazioni estratte siano conformi allo schema grafico predefinito e mantengano l'integrità delle connessioni nodo-edge e delle proprietà associate.



Dopo l'estrazione e la convalida delle entità, delle relazioni e degli attributi, puoi collegarli per creare una subject-object-predicate tripletta. Questi dati vengono importati in un database grafico di Amazon Neptune, come illustrato nella figura seguente. I database grafici sono ottimizzati per archiviare e interrogare le relazioni tra gli elementi di dati.


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

È possibile creare un grafico della conoscenza completo con questi dati. Un Knowledge Graph consente di organizzare e interrogare tutti i tipi di informazioni connesse.
Ad esempio, è possibile creare un Knowledge Graph con i seguenti nodi principali: `HospitalVisitPastMedicalHistory`,`Symptoms`,`Medication`,`MedicalProcedures`, `eTreatment`.

Nelle tabelle seguenti sono elencate le entità e i relativi attributi che è possibile estrarre dalle note di discarica.

Entità	Attributes
`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`

Entità	Attributes
`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`

La tabella seguente elenca le relazioni che le entità possono avere e gli attributi corrispondenti. Ad esempio, l'`Patient`entità potrebbe connettersi all'`HospitalVisit`entità con la `[UNDERGOES]` relazione. L'attributo per questa relazione è`VisitDate`.

Entità oggetto	Relazione	Entità oggetto	Attributes
`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`

Entità oggetto	Relazione	Entità oggetto	Attributes
HospitalVisit	[PRESCRIBED]	Medication	Duration, Dosage, Adherence , VisitDate
Patient	[HAS_HISTORY]	PastMedicalHistory	Nessuno
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	Nessuna
LabResult	[HAS_TAGGED]	MedicalConcepts	Nessuna
Treatment	[HAS_TAGGED]	MedicalConcepts	Nessuna
Symptoms	[HAS_TAGGED]	MedicalConcepts	Nessuno

Fase 3: Creazione di agenti di recupero del contesto per interrogare il grafico delle conoscenze mediche

Dopo aver creato il database dei grafici medici, il passaggio successivo consiste nella creazione di agenti per l'interazione con i grafici. Gli agenti recuperano il contesto corretto e richiesto per la richiesta inserita da un medico o da un medico. Esistono diverse opzioni per configurare questi agenti che recuperano il contesto dal Knowledge Graph:

- Agenti Amazon Bedrock
- LangChain agenti

Agenti Amazon Bedrock per l'interazione con i grafici

Gli agenti Amazon Bedrock funzionano perfettamente con i database grafici di Amazon Neptune. Puoi eseguire interazioni avanzate tramite i gruppi di azioni Amazon Bedrock. Il gruppo di azioni avvia il processo chiamando una AWS Lambda funzione, che esegue le query OpenCypher di Neptune.

Per interrogare un Knowledge Graph, è possibile utilizzare due approcci distinti: esecuzione diretta delle query o interrogazione con incorporamento del contesto. Questi approcci possono essere applicati indipendentemente o combinati, a seconda del caso d'uso specifico e dei criteri di classificazione. Combinando entrambi gli approcci, è possibile fornire un contesto più completo all'LLM, che può migliorare i risultati. Di seguito sono riportati i due approcci di esecuzione delle query:

- Esecuzione diretta delle query Cypher senza incorporamenti: la funzione Lambda esegue le query direttamente su Neptune senza alcuna ricerca basata sugli incorporamenti. Di seguito è riportato un esempio di questo approccio:

```
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
```

- Esecuzione di query Direct Cypher utilizzando la ricerca di incorporamento: la funzione Lambda utilizza la ricerca di incorporamento per migliorare i risultati delle query. Questo approccio migliora l'esecuzione delle query incorporando gli incorporamenti, che sono rappresentazioni vettoriali dense di dati. Gli incorporamenti sono particolarmente utili quando la query richiede una somiglianza semantica o una comprensione più ampia che vada oltre le corrispondenze esatte.

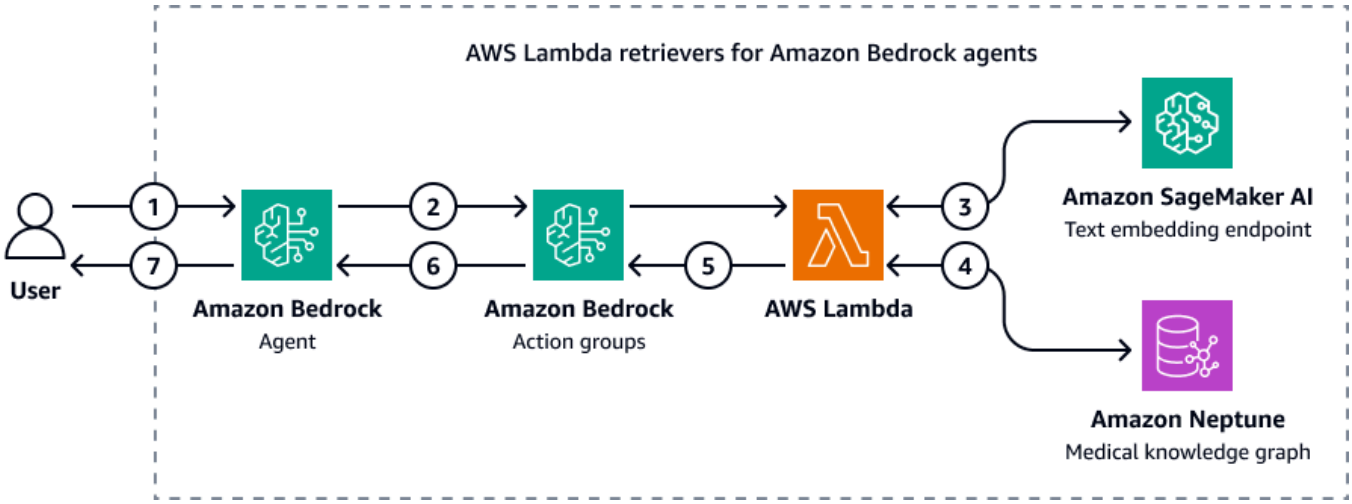
È possibile utilizzare modelli preformati o personalizzati per generare incorporamenti per ogni condizione medica. Di seguito è riportato un esempio di questo approccio:

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

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

In questo esempio, la `search_embedding("Acute Diabetes")` funzione recupera condizioni semanticamente simili al «diabete acuto». Ciò aiuta la query a trovare anche pazienti affetti da condizioni come il prediabete o la sindrome metabolica.

L'immagine seguente mostra come gli agenti di Amazon Bedrock interagiscono con Amazon Neptune per eseguire una query Cypher su un grafico delle conoscenze mediche.



Il diagramma mostra il flusso di lavoro seguente:

1. L'utente invia una domanda all'agente Amazon Bedrock.
2. L'agente Amazon Bedrock trasmette le variabili del filtro di domanda e input ai gruppi di azione Amazon Bedrock. Questi gruppi di azioni contengono una AWS Lambda funzione che interagisce con l'endpoint di incorporamento di testi di Amazon SageMaker AI e il grafico delle conoscenze mediche di Amazon Neptune.
3. La funzione Lambda si integra con l'endpoint di incorporamento del testo SageMaker AI per eseguire una ricerca semantica all'interno della query OpenCypher. Converte la query in

linguaggio naturale in una query OpenCypher utilizzando la funzione sottostante LangChain agenti.

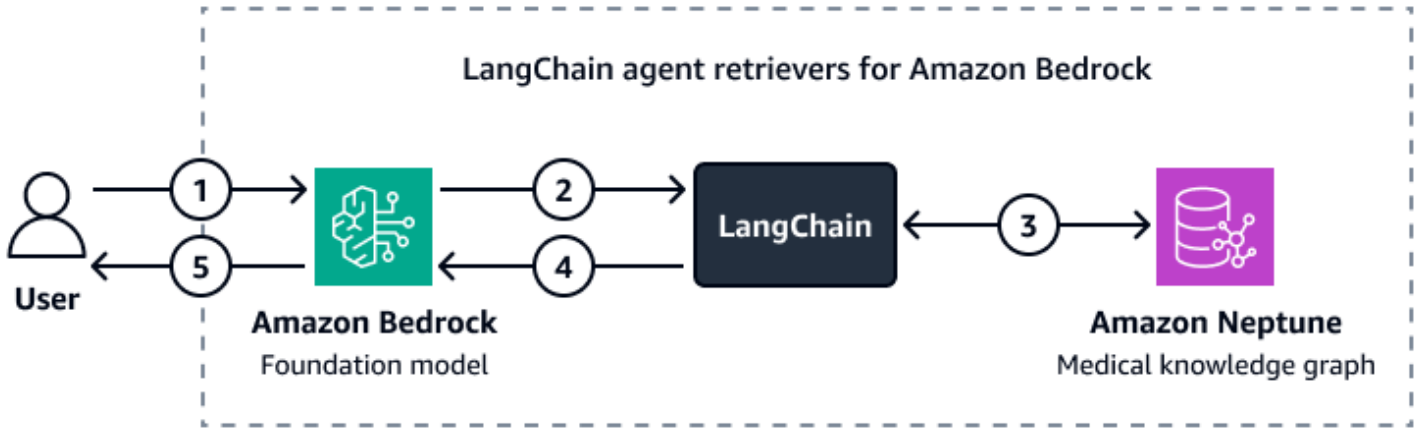
4. La funzione Lambda interroga il Neptune Medical Knowledge Graph per trovare il set di dati corretto e riceve l'output dal Neptune Medical Knowledge Graph.
5. La funzione Lambda restituisce i risultati da Neptune ai gruppi di azioni Amazon Bedrock.
6. I gruppi di azione Amazon Bedrock inviano il contesto recuperato all'agente Amazon Bedrock.
7. L'agente Amazon Bedrock genera la risposta utilizzando la query originale dell'utente e il contesto recuperato dal knowledge graph.

LangChain agenti per l'interazione con i grafici

È possibile integrare LangChain con Neptune per consentire interrogazioni e recuperi basati su grafici. Questo approccio può migliorare i flussi di lavoro basati sull'intelligenza artificiale utilizzando le funzionalità del database grafico di Neptune. La personalizzazione LangChain il retriever funge da intermediario. Il modello base di Amazon Bedrock può interagire con Neptune utilizzando sia query Cypher dirette che algoritmi grafici più complessi.

Puoi utilizzare il retriever personalizzato per perfezionare il modo in cui LangChain l'agente interagisce con gli algoritmi del grafico di Neptune. Ad esempio, è possibile utilizzare few-shot prompting, che consente di personalizzare le risposte del modello di base in base a modelli o esempi specifici. Puoi anche applicare filtri identificati da LLM per rifinire il contesto e migliorare la precisione delle risposte. Ciò può migliorare l'efficienza e la precisione dell'intero processo di recupero quando si interagisce con dati grafici complessi.

L'immagine seguente mostra come una personalizzazione LangChain agent orchestra l'interazione tra un modello di base Amazon Bedrock e un grafico della conoscenza medica di Amazon Neptune.



Il diagramma mostra il flusso di lavoro seguente:

1. Un utente invia una domanda ad Amazon Bedrock e al LangChain agente.
2. Il modello di base Amazon Bedrock utilizza lo schema Neptune, fornito da LangChain agente, per generare una query per la domanda dell'utente.
3. Il LangChain l'agente esegue la query sul grafico delle conoscenze mediche di Amazon Neptune.
4. Il LangChain agent invia il contesto recuperato al modello di base Amazon Bedrock.
5. Il modello di base Amazon Bedrock utilizza il contesto recuperato per generare una risposta alla domanda dell'utente.

Fase 4: creazione di una knowledge base di dati descrittivi in tempo reale

Successivamente, si crea una knowledge base di note descrittive sulle interazioni medico-paziente in tempo reale, valutazioni di immagini diagnostiche e rapporti di analisi di laboratorio. Questa knowledge base è un database vettoriale. Utilizzando un database vettoriale, in grado di archiviare conoscenze mediche descrittive in forma indicizzata e vettoriale, gli operatori sanitari possono interrogare e accedere in modo efficiente alle informazioni pertinenti da un vasto archivio. Queste rappresentazioni vettoriali consentono di recuperare dati semanticamente simili. Gli operatori sanitari possono navigare rapidamente tra note cliniche, immagini mediche e risultati di laboratorio. Ciò accelera il processo decisionale informato offrendo l'accesso immediato a informazioni contestualmente rilevanti, migliorando l'accuratezza e la velocità delle diagnosi e dei piani di trattamento.

Utilizzo di una base di conoscenze mediche del Service OpenSearch

Amazon OpenSearch Service è in grado di gestire grandi volumi di dati medici ad alta dimensione. È un servizio gestito che facilita la ricerca ad alte prestazioni e l'analisi in tempo reale. È adatto come database vettoriale per applicazioni RAG. OpenSearch Il servizio funge da strumento di backend per gestire grandi quantità di dati non strutturati o semistrutturati, come cartelle cliniche, articoli di ricerca e note cliniche. Le sue funzionalità avanzate di ricerca semantica consentono di recuperare informazioni contestualmente rilevanti. Ciò lo rende particolarmente utile in applicazioni come i sistemi di supporto alle decisioni cliniche, gli strumenti per la risoluzione delle domande dei pazienti e i sistemi di gestione delle conoscenze sanitarie. Ad esempio, un medico può trovare rapidamente i dati pertinenti sui pazienti o gli studi di ricerca che corrispondono a sintomi o protocolli di trattamento

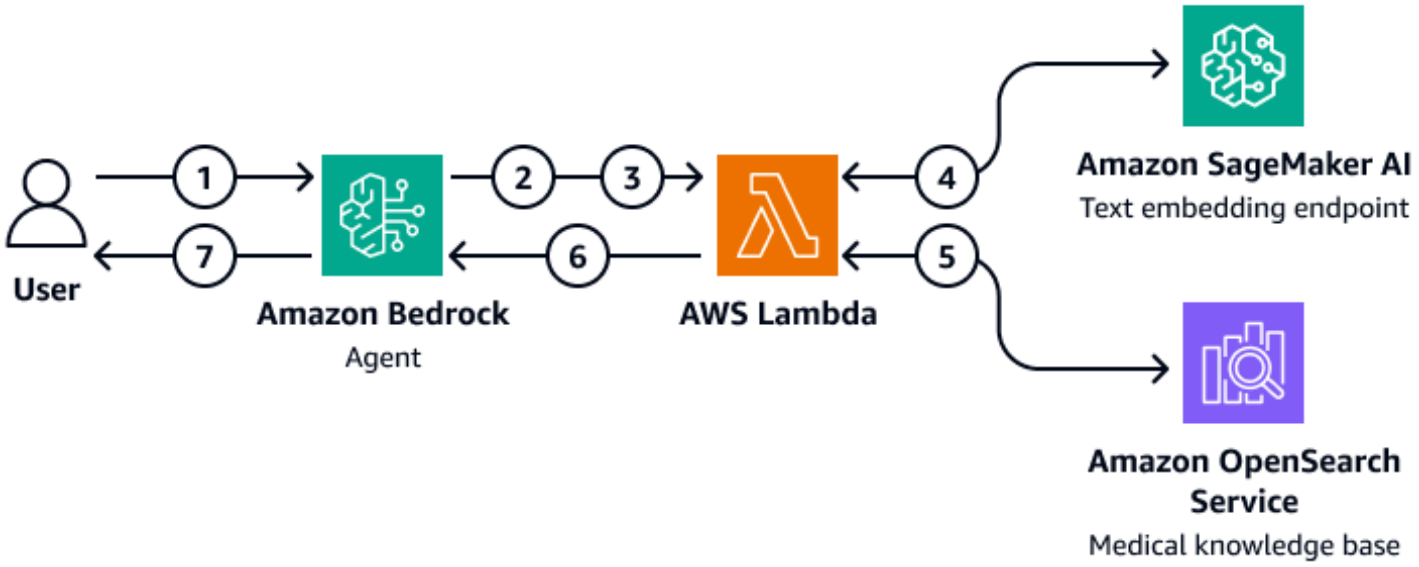
specifici. Questo aiuta i medici a prendere decisioni basate sulla maggior parte delle informazioni up-to-date pertinenti.

OpenSearch Il servizio può scalare e gestire l'indicizzazione e l'interrogazione dei dati in tempo reale. Ciò lo rende ideale per ambienti sanitari dinamici in cui l'accesso tempestivo a informazioni accurate è fondamentale. Inoltre, dispone di funzionalità di ricerca multimodali ottimali per le ricerche che richiedono input multipli, come immagini mediche e note mediche. Quando si implementano le applicazioni OpenSearch Service for Healthcare, è fondamentale definire campi e mappature precisi per ottimizzare l'indicizzazione e il recupero dei dati. I campi rappresentano i singoli dati, come le cartelle cliniche dei pazienti, le anamnesi e i codici diagnostici. Le mappature definiscono il modo in cui questi campi vengono archiviati (nel modulo di incorporamento o nel modulo originale) e come vengono interrogati. Per le applicazioni sanitarie, è essenziale stabilire mappature che contengano vari tipi di dati, inclusi dati strutturati (come i risultati dei test numerici), dati semistrutturati (come le note dei pazienti) e dati non strutturati (come le immagini mediche)

In OpenSearch Service, è possibile eseguire query di ricerca neurale complete tramite istruzioni curate per cercare nelle cartelle cliniche, nelle note cliniche o nei documenti di ricerca per trovare rapidamente informazioni pertinenti su sintomi, trattamenti o storie di pazienti specifici. Le query di ricerca neurali gestiscono automaticamente l'incorporamento del prompt di input e delle immagini utilizzando modelli di rete neurale integrati. Ciò consente di comprendere e acquisire le relazioni semantiche più profonde nei dati multimodali, offrendo risultati di ricerca più attenti al contesto e precisi rispetto ad altri algoritmi di query di ricerca, come la ricerca K-Nearest Neighbor (k-NN).

Creazione di un'architettura RAG

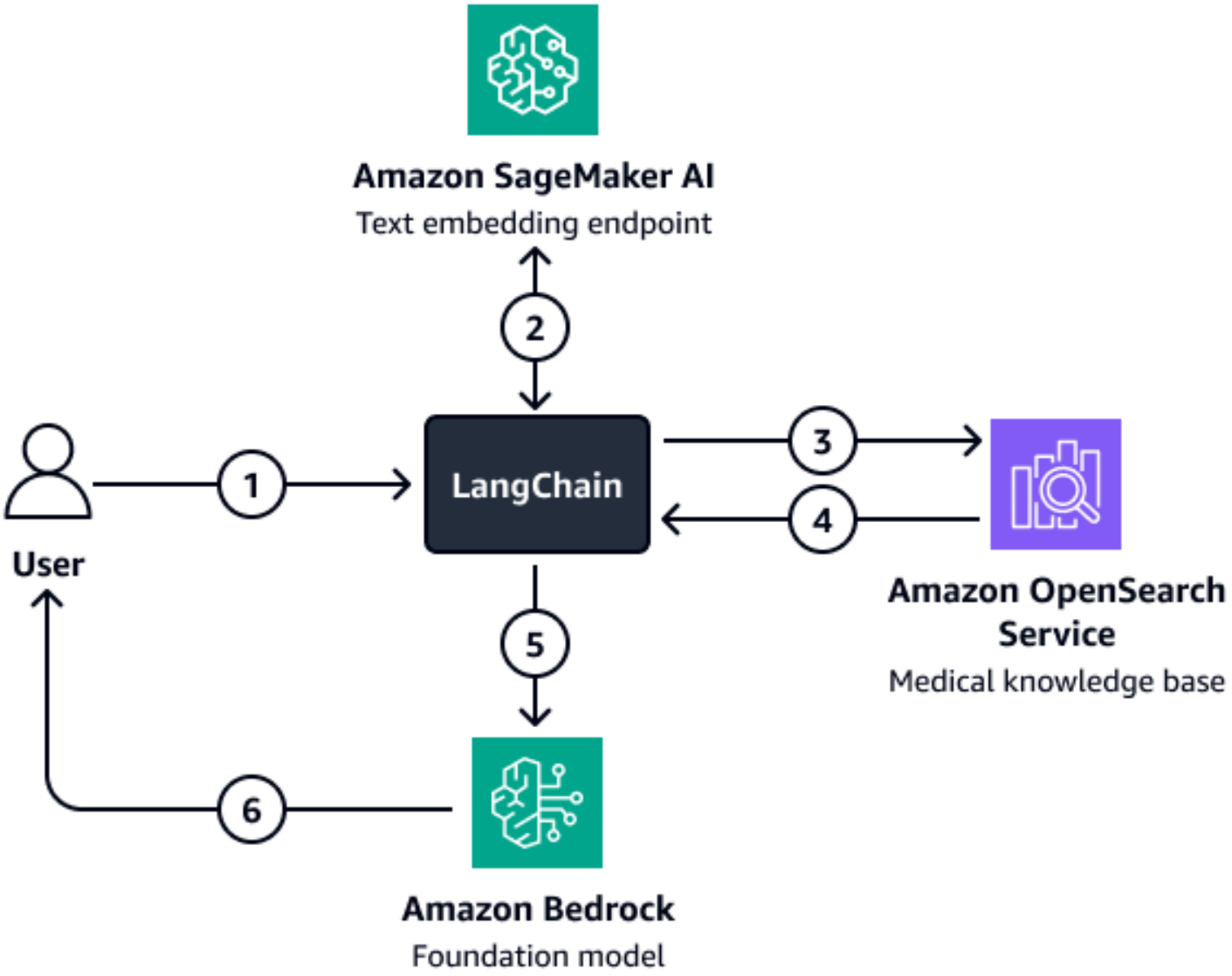
Puoi implementare una soluzione RAG personalizzata che utilizza agenti Amazon Bedrock per interrogare una knowledge base medica in Service. OpenSearch A tale scopo, crei una AWS Lambda funzione in grado di interagire e interrogare Service. OpenSearch La funzione Lambda incorpora la domanda di input dell'utente accedendo a un endpoint di incorporamento di testo SageMaker AI. L'agente Amazon Bedrock trasmette parametri di query aggiuntivi come input alla funzione Lambda. La funzione interroga la knowledge base medica di OpenSearch Service, che restituisce i contenuti medici pertinenti. Dopo aver configurato la funzione Lambda, aggiungila come gruppo di azioni all'interno dell'agente Amazon Bedrock. L'agente Amazon Bedrock prende l'input dell'utente, identifica le variabili necessarie, passa le variabili e la domanda alla funzione Lambda e quindi avvia la funzione. La funzione restituisce un contesto che aiuta il modello di base a fornire una risposta più accurata alla domanda dell'utente.



Il diagramma mostra il flusso di lavoro seguente:

1. Un utente invia una domanda all'agente Amazon Bedrock.
2. L'agente Amazon Bedrock seleziona il gruppo di azioni da avviare.
3. L'agente Amazon Bedrock avvia una AWS Lambda funzione e le trasmette i parametri.
4. La funzione Lambda avvia il modello di inserimento del testo di Amazon SageMaker AI per incorporare la domanda dell'utente.
5. La funzione Lambda passa il testo incorporato e parametri e filtri aggiuntivi ad Amazon OpenSearch Service. Amazon OpenSearch Service interroga la knowledge base medica e restituisce i risultati alla funzione Lambda.
6. La funzione Lambda restituisce i risultati all'agente Amazon Bedrock.
7. Il modello base dell'agente Amazon Bedrock genera una risposta basata sui risultati e restituisce la risposta all'utente.

Per le situazioni in cui sono necessari filtri più complessi, puoi utilizzare un filtro personalizzato LangChain recuperatore. Crea questo retriever configurando un client di ricerca vettoriale OpenSearch Service che viene caricato direttamente in LangChain. Questa architettura consente di passare più variabili per creare i parametri del filtro. Dopo aver configurato il retriever, usa il modello Amazon Bedrock e il retriever per configurare una catena di domande di recupero e risposta. Questa catena orchestra l'interazione tra il modello e il retriever passando l'input dell'utente e i potenziali filtri al retriever. Il retriever restituisce il contesto pertinente che aiuta il modello di base a rispondere alla domanda dell'utente.



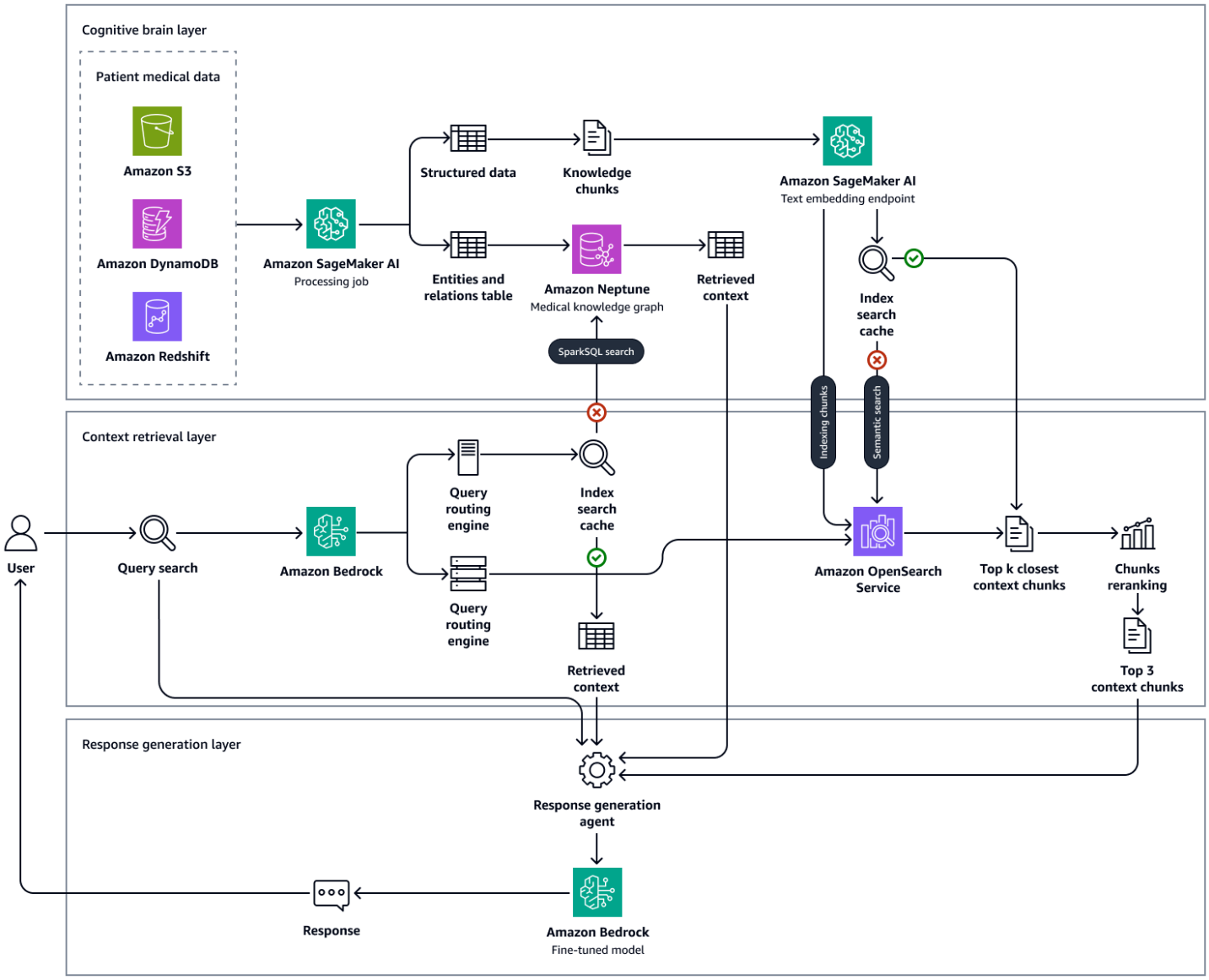
Il diagramma mostra il flusso di lavoro seguente:

1. Un utente invia una domanda al LangChain agente retriever.
2. Il LangChain retriever agent invia la domanda all'endpoint di incorporamento del testo Amazon SageMaker AI per incorporare la domanda.
3. Il LangChain retriever agent passa il testo incorporato ad Amazon OpenSearch Service.
4. Amazon OpenSearch Service restituisce i documenti recuperati al LangChain agente retriever.
5. Il LangChain retriever agent passa la domanda dell'utente e il contesto recuperato al modello Amazon Bedrock Foundation.
6. Il modello di base genera una risposta e la invia all'utente.

Fase 5: Utilizzo LLMs per rispondere a domande mediche

I passaggi precedenti consentono di creare un'applicazione di intelligence medica in grado di recuperare le cartelle cliniche di un paziente e riepilogare i farmaci pertinenti e le potenziali diagnosi. Ora crei il livello di generazione. Questo livello utilizza le funzionalità generative di un LLM in Amazon Bedrock, come Llama 3, per aumentare l'output dell'applicazione.

Quando un medico inserisce una query, il livello di recupero del contesto dell'applicazione esegue il processo di recupero dal Knowledge Graph e restituisce i record principali relativi all'anamnesi, ai dati demografici, ai sintomi, alla diagnosi e agli esiti del paziente. Dal database vettoriale, recupera anche note descrittive sull'interazione medico-paziente in tempo reale, approfondimenti sulla valutazione delle immagini diagnostiche, riassunti dei rapporti di analisi di laboratorio e approfondimenti da un enorme corpus di ricerche mediche e libri accademici. Questi principali risultati recuperati, la richiesta del medico e i prompt (che sono personalizzati per selezionare le risposte in base alla natura della domanda), vengono quindi passati al modello base di Amazon Bedrock. Questo è il livello di generazione della risposta. Il LLM utilizza il contesto recuperato per generare una risposta alla domanda del medico. La figura seguente mostra il end-to-end flusso di lavoro dei passaggi di questa soluzione.



Puoi utilizzare un modello di base pre-addestrato in Amazon Bedrock, come Llama 3, per una serie di casi d'uso che l'applicazione di intelligence medica deve gestire. Il LLM più efficace per una determinata attività varia a seconda del caso d'uso. Ad esempio, un modello pre-addestrato potrebbe essere sufficiente per riassumere le conversazioni tra paziente e medico, cercare tra i farmaci e le storie dei pazienti e recuperare informazioni dai set di dati medici interni e dalle conoscenze scientifiche. Tuttavia, potrebbe essere necessario un LLM ottimizzato per altri casi d'uso complessi, come valutazioni di laboratorio in tempo reale, raccomandazioni sulle procedure mediche e previsioni degli esiti dei pazienti. È possibile perfezionare un LLM addestrandolo su set di dati di dominio medico. Requisiti sanitari e delle scienze della vita specifici o complessi guidano lo sviluppo di questi modelli perfezionati.

Per ulteriori informazioni sulla messa a punto di un LLM o sulla scelta di un LLM esistente che sia stato formato sui dati del dominio medico, vedi Utilizzo di modelli linguistici di grandi dimensioni per i casi d'uso nel settore sanitario e delle scienze della vita.

Allineamento al Well-Architected AWS Framework

La soluzione si allinea a tutti e sei i pilastri del AWS Well-Architected Framework come segue:

- Eccellenza operativa: l'architettura è disaccoppiata per un monitoraggio e un aggiornamento efficienti. Agenti Amazon Bedrock e ti AWS Lambda aiutano a distribuire e ripristinare rapidamente gli strumenti.
- Sicurezza: questa soluzione è progettata per rispettare le normative sanitarie, come l'HIPAA. Puoi anche implementare la crittografia, il controllo granulare degli accessi e i guardrail di Amazon Bedrock per proteggere i dati dei pazienti.
- Affidabilità: i servizi AWS gestiti, come Amazon OpenSearch Service e Amazon Bedrock, forniscono l'infrastruttura per l'interazione continua dei modelli.
- Efficienza delle prestazioni: la soluzione RAG recupera rapidamente i dati pertinenti utilizzando la ricerca semantica ottimizzata e le query Cypher, mentre un router agente identifica i percorsi ottimali per le query degli utenti.
- Ottimizzazione dei costi: il pay-per-token modello dell'architettura Amazon Bedrock e RAG riduce i costi di inferenza e pre-formazione.
- Sostenibilità: l'utilizzo dell'infrastruttura e dell' pay-per-tokenelaborazione serverless riduce al minimo l'utilizzo delle risorse e migliora la sostenibilità.

Caso d'uso: previsione degli esiti dei pazienti e dei tassi di riammissione

L'analisi predittiva basata sull'intelligenza artificiale offre ulteriori vantaggi prevedendo gli esiti dei pazienti e abilitando piani di trattamento personalizzati. Ciò può migliorare la soddisfazione dei pazienti e gli esiti sanitari. Integrando queste funzionalità di intelligenza artificiale con Amazon Bedrock e altre tecnologie, gli operatori sanitari possono ottenere significativi aumenti di produttività, ridurre i costi ed elevare la qualità complessiva dell'assistenza ai pazienti.

È possibile archiviare dati medici, come anamnesi dei pazienti, note cliniche, farmaci e trattamenti, in un Knowledge Graph. Combinando la profonda comprensione contestuale LLMs con i dati strutturati e temporali contenuti in un grafico delle conoscenze mediche, gli operatori sanitari possono ottenere ulteriori informazioni sui modelli individuali dei pazienti. Utilizzando l'analisi predittiva, è possibile identificare tempestivamente potenziali casi di mancata aderenza o complicanze terapeutiche e generare punteggi personalizzati di propensione alla riammissione.

Questa soluzione consente di prevedere la probabilità di una riammissione. Queste previsioni possono migliorare gli esiti dei pazienti e ridurre i costi sanitari. Questa soluzione può anche aiutare i medici e gli amministratori ospedalieri a concentrare la propria attenzione sui pazienti con un rischio più elevato di riammissione. Inoltre, li aiuta ad avviare interventi proattivi con tali pazienti tramite avvisi, azioni self-service e basate sui dati.

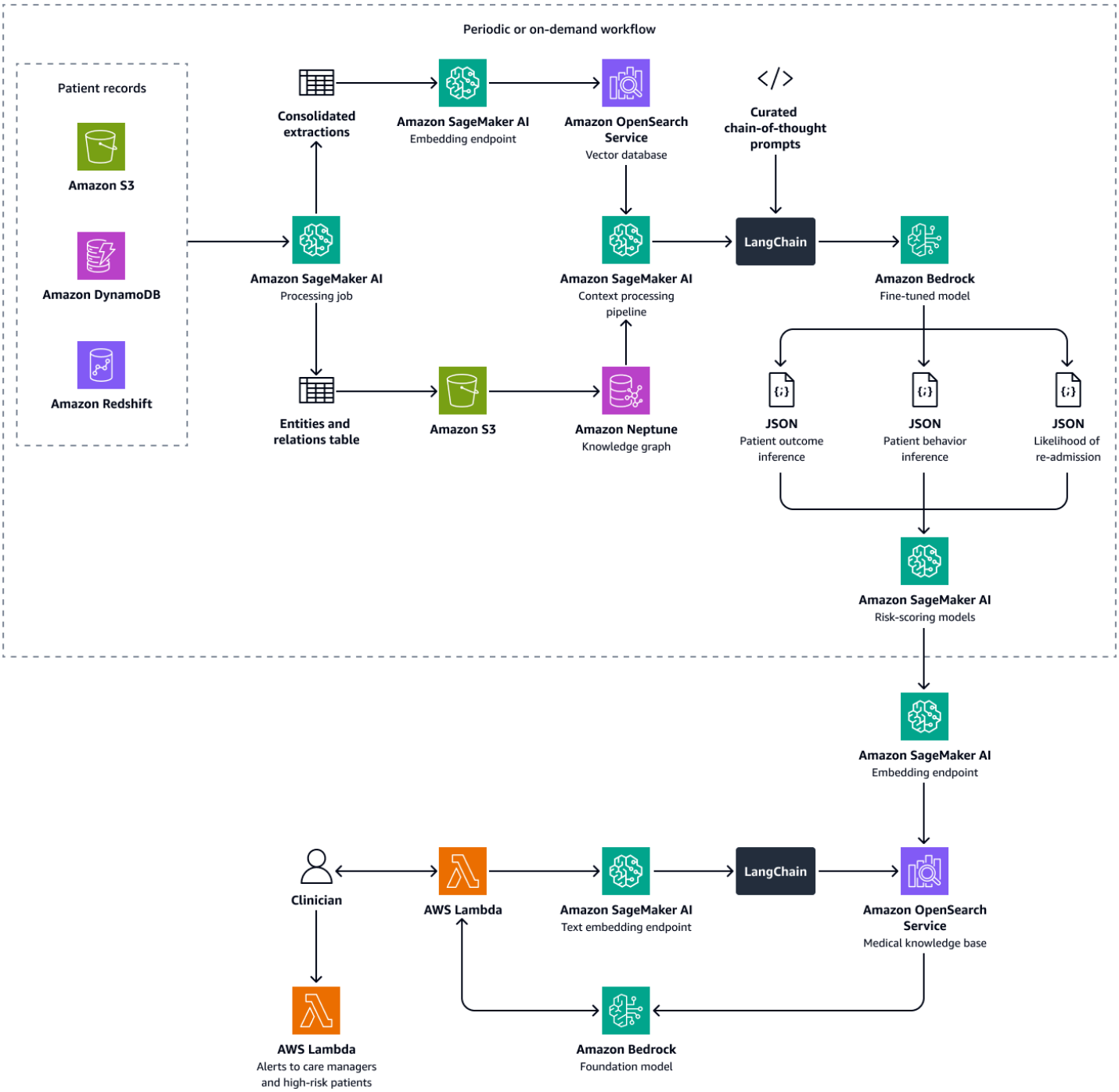
Panoramica della soluzione

Questa soluzione utilizza un framework multi-retrieval Augmented Generation (RAG) per analizzare i dati dei pazienti. Prevede la probabilità di riammissione ospedaliera per i singoli pazienti e aiuta a calcolare un punteggio di propensione alla riammissione a livello ospedaliero. Questa soluzione integra le seguenti funzionalità:

- Knowledge graph: archivia i dati strutturati e cronologici dei pazienti, come gli incontri ospedalieri, i precedenti ricoveri, i sintomi, i risultati di laboratorio, i trattamenti prescritti e l'anamnesi di aderenza ai farmaci
- Database vettoriale: archivia dati clinici non strutturati, come riepiloghi delle dimissioni, note mediche e registrazioni degli appuntamenti mancati o degli effetti collaterali segnalati dei farmaci

- LLM ottimizzato: utilizza sia i dati strutturati del Knowledge Graph che i dati non strutturati del database vettoriale per generare inferenze sul comportamento del paziente, sull'aderenza al trattamento e sulla probabilità di riammissione

I modelli di valutazione del rischio quantificano le inferenze del LLM in punteggi numerici. È possibile aggregare i punteggi in un punteggio di propensione alla riammissione a livello ospedaliero. Questo punteggio definisce l'esposizione al rischio di ogni paziente ed è possibile calcolarlo periodicamente o in base alle necessità. Tutte le inferenze e i punteggi di rischio sono indicizzati e archiviati in Amazon OpenSearch Service in modo che i responsabili sanitari e i medici possano recuperarli. Integrando un agente di intelligenza artificiale conversazionale con questo database vettoriale, i medici e i responsabili dell'assistenza possono estrarre informazioni senza problemi a livello di singolo paziente, a livello di struttura o per specialità medica. Puoi anche impostare avvisi automatici basati sui punteggi di rischio, per incoraggiare interventi proattivi.



La creazione di questa soluzione prevede i seguenti passaggi:

- Fase 1: Previsione degli esiti dei pazienti utilizzando un grafico delle conoscenze mediche
- Fase 2: Previsione del comportamento del paziente rispetto ai farmaci o ai trattamenti prescritti
- Fase 3: Previsione della probabilità di riammissione del paziente

- Fase 4: Calcolo del punteggio di propensione alla riammissione in ospedale

Fase 1: Previsione degli esiti dei pazienti utilizzando un grafico delle conoscenze mediche

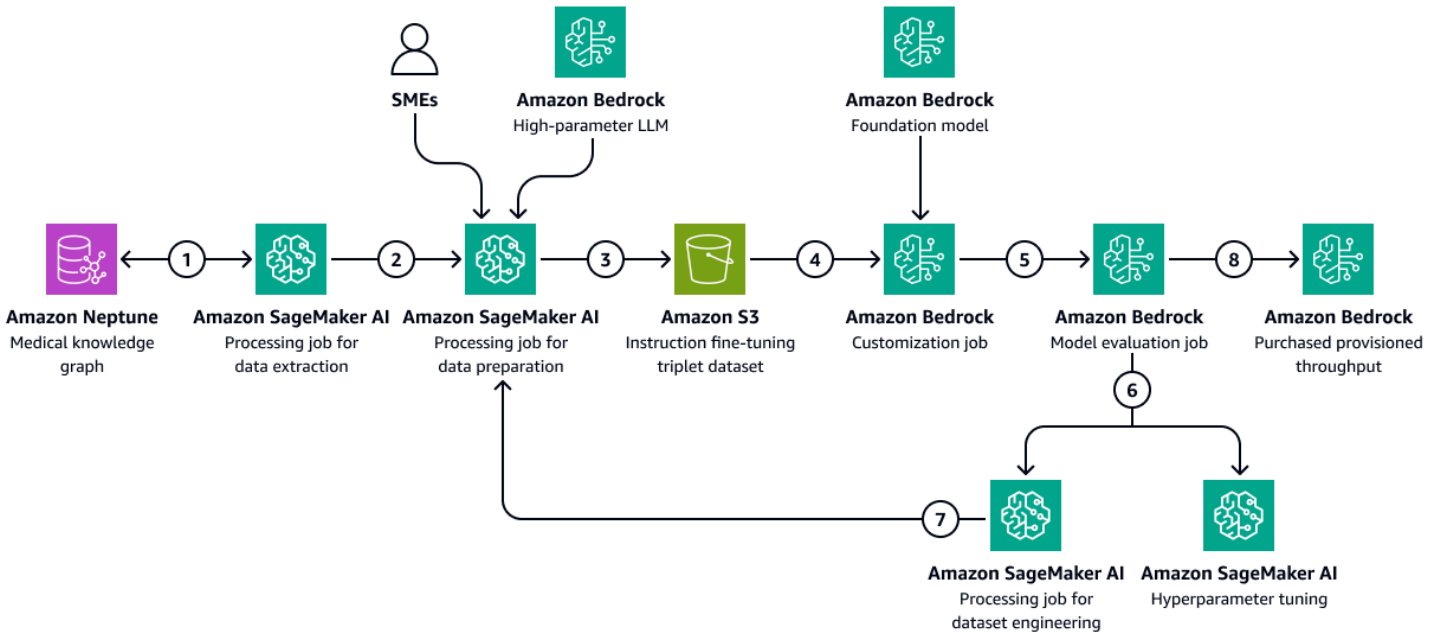
In Amazon Neptune, puoi utilizzare un Knowledge Graph per archiviare informazioni temporali sulle visite e sugli esiti dei pazienti nel tempo. Il modo più efficace per creare e archiviare un knowledge graph consiste nell'utilizzare un modello grafico e un database grafico. I database a grafi sono progettati appositamente per archiviare e gestire le relazioni. I database a grafi semplificano la modellazione e la gestione di dati altamente connessi e dispongono di schemi flessibili.

Il Knowledge Graph consente di eseguire analisi di serie temporali. Di seguito sono riportati gli elementi chiave del database grafico utilizzato per la previsione temporale degli esiti dei pazienti:

- Dati storici: diagnosi precedenti, terapia continuativa, farmaci usati in precedenza e risultati di laboratorio relativi al paziente
- Visite dei pazienti (cronologiche): date delle visite, sintomi, allergie osservate, note cliniche, diagnosi, procedure, trattamenti, farmaci prescritti e risultati di laboratorio
- Sintomi e parametri clinici: informazioni cliniche e basate sui sintomi, tra cui la gravità, i modelli di progressione e la risposta del paziente al farmaco

Puoi utilizzare gli approfondimenti del Medical Knowledge Graph per perfezionare un LLM in Amazon Bedrock, come Llama 3. Puoi perfezionare il LLM con dati sequenziali del paziente sulla risposta del paziente a una serie di farmaci o trattamenti nel tempo. Utilizzate un set di dati etichettato che classifica una serie di farmaci o trattamenti e i dati di interazione paziente-clinica in categorie predefinite che indicano lo stato di salute di un paziente. Esempi di queste categorie sono il deterioramento della salute, il miglioramento o il progresso stabile. Quando il medico inserisce un nuovo contesto sul paziente e sui suoi sintomi, il LLM ottimizzato può utilizzare i modelli del set di dati di formazione per prevedere il potenziale esito del paziente.

L'immagine seguente mostra i passaggi sequenziali necessari per perfezionare un LLM in Amazon Bedrock utilizzando un set di dati di formazione specifico per il settore sanitario. Questi dati potrebbero includere le condizioni mediche dei pazienti e le risposte ai trattamenti nel tempo. Questo set di dati di formazione aiuterebbe il modello a fare previsioni generalizzate sugli esiti dei pazienti.



Il diagramma mostra il flusso di lavoro seguente:

1. Il processo di estrazione dei dati di Amazon SageMaker AI interroga il Knowledge Graph per recuperare dati cronologici sulle risposte dei diversi pazienti a una serie di farmaci o trattamenti nel tempo.
2. Il processo di preparazione dei dati SageMaker AI integra un Amazon Bedrock LLM e input di esperti in materia (). SMEs Il lavoro classifica i dati recuperati dal Knowledge Graph in categorie predefinite (come deterioramento della salute, miglioramento o progresso stabile) che indicano lo stato di salute di ciascun paziente.
3. Il lavoro crea un set di dati di ottimizzazione che include le informazioni estratte dal Knowledge Graph, le istruzioni e la categoria di esiti del paziente. chain-of-thought Carica questo set di dati di addestramento in un bucket Amazon S3.
4. Un processo di personalizzazione di Amazon Bedrock utilizza questo set di dati di formazione per mettere a punto un LLM.
5. Il processo di personalizzazione di Amazon Bedrock integra il modello base di Amazon Bedrock preferito nell'ambiente di formazione. Avvia il processo di ottimizzazione e utilizza il set di dati di addestramento e gli iperparametri di allenamento che configuri.
6. Un processo di valutazione di Amazon Bedrock valuta il modello ottimizzato utilizzando un framework di valutazione del modello predefinito.
7. Se il modello necessita di miglioramenti, il processo di formazione viene riavviato con più dati dopo un'attenta valutazione del set di dati di addestramento. Se il modello non dimostra

un miglioramento incrementale delle prestazioni, valuta anche la possibilità di modificare gli iperparametri di allenamento.

8. Dopo che la valutazione del modello soddisfa gli standard definiti dagli stakeholder aziendali, esegui l'hosting del modello ottimizzato in base al throughput assegnato da Amazon Bedrock.

Fase 2: Previsione del comportamento del paziente rispetto ai farmaci o ai trattamenti prescritti

Fine-tuned è in LLMs grado di elaborare note cliniche, riepiloghi delle dimissioni e altri documenti specifici del paziente tratti dal grafico temporale delle conoscenze mediche. Possono valutare se è probabile che il paziente segua i farmaci o i trattamenti prescritti.

Questo passaggio utilizza il Knowledge Graph creato in[Fase 1: Previsione degli esiti dei pazienti utilizzando un grafico delle conoscenze mediche](). Il Knowledge Graph contiene i dati del profilo del paziente, inclusa l'aderenza storica del paziente come nodo. Tra le caratteristiche di tali nodi sono inclusi anche casi di mancata aderenza a farmaci o trattamenti, effetti collaterali dei farmaci, mancanza di accesso o barriere di costo ai farmaci o regimi di dosaggio complessi.

Fine-tuned LLMs può utilizzare i dati di adempimento delle prescrizioni precedenti tratti dal Medical Knowledge Graph e i riepiloghi descrittivi delle note cliniche da un database vettoriale di Amazon Service. OpenSearch Queste note cliniche potrebbero menzionare appuntamenti saltati frequentemente o la mancata osservanza dei trattamenti. L'LLM può utilizzare queste note per prevedere la probabilità di future non aderenze.

1. Preparare i dati di input come segue:
   - Dati strutturati: estrai i dati recenti dei pazienti, come le ultime tre visite e i risultati di laboratorio, dal grafico delle conoscenze mediche.
   - Dati non strutturati: recupera le note cliniche recenti dal database vettoriale di Amazon OpenSearch Service.
2. Crea un prompt di input che includa l'anamnesi del paziente e il contesto attuale. Di seguito è riportato un prompt di esempio:

```
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. Passa il prompt al LLM ottimizzato. L'LLM elabora il prompt e prevede il risultato. Di seguito è riportato un esempio di risposta del 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. Analizza la risposta del modello per estrarre la categoria di risultati prevista. Ad esempio, la categoria della risposta di esempio nel passaggio precedente potrebbe essere un'elevata probabilità di non aderenza.
5. (Facoltativo) Utilizzate i log dei modelli o metodi aggiuntivi per assegnare punteggi di confidenza. I logit sono le probabilità non normalizzate dell'elemento appartenente a una determinata classe o categoria.

Fase 3: Previsione della probabilità di riammissione del paziente

Le riammissioni ospedaliere sono una delle principali preoccupazioni a causa degli elevati costi di amministrazione sanitaria e del loro impatto sul benessere del paziente. Il calcolo dei tassi di riammissione ospedaliera è un modo per misurare la qualità dell'assistenza ai pazienti e le prestazioni di un operatore sanitario.

Per calcolare il tasso di riammissione, hai definito un indicatore, ad esempio un tasso di riammissione di 7 giorni. Questo indicatore è la percentuale di pazienti ricoverati che tornano in ospedale per una visita non programmata entro sette giorni dalla dimissione. Per prevedere la possibilità di riammissione di un paziente, un LLM ottimizzato può utilizzare i dati temporali del grafico delle conoscenze mediche in cui è stato creato. Fase 1: Previsione degli esiti dei pazienti utilizzando un grafico delle conoscenze mediche Questo Knowledge Graph conserva registrazioni cronologiche degli incontri, delle procedure, dei farmaci e dei sintomi dei pazienti. Questi record di dati contengono quanto segue:

- Periodo di tempo trascorso dall'ultima dimissione del paziente
- La risposta del paziente ai trattamenti e ai farmaci precedenti
- La progressione dei sintomi o delle condizioni nel tempo

È possibile elaborare queste serie temporali per prevedere la probabilità di riammissione di un paziente tramite un prompt di sistema curato. Il prompt impartisce la logica di previsione al LLM ottimizzato.

1. Preparate i dati di input come segue:
   - Cronologia di aderenza: estrai le date di ritiro dei farmaci, le frequenze di ricarica dei farmaci, la diagnosi e i dettagli del farmaco, la storia medica cronologica e altre informazioni dal grafico delle conoscenze mediche.
   - Indicatori comportamentali: recupera e includi note cliniche sugli appuntamenti mancati e sugli effetti collaterali segnalati dai pazienti.
2. Crea un prompt di input che includa la cronologia di aderenza e gli indicatori comportamentali. Di seguito è riportato un prompt di esempio:

```
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. Passa il prompt al LLM ottimizzato. L'LLM elabora la richiesta e prevede la probabilità e le ragioni della riammissione. Di seguito è riportato un esempio di risposta del 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. Categorizza la previsione in una scala standardizzata, ad esempio bassa, media o alta.
5. Esamina il ragionamento fornito dal LLM e identifica i fattori chiave che contribuiscono alla previsione.
6. Mappa i risultati qualitativi in punteggi quantitativi. Ad esempio, molto alto potrebbe corrispondere a una probabilità di 0,9.
7. Utilizza set di dati di convalida per calibrare gli output del modello rispetto ai tassi di riammissione effettivi.
```

Fase 4: Calcolo del punteggio di propensione alla riammissione in ospedale

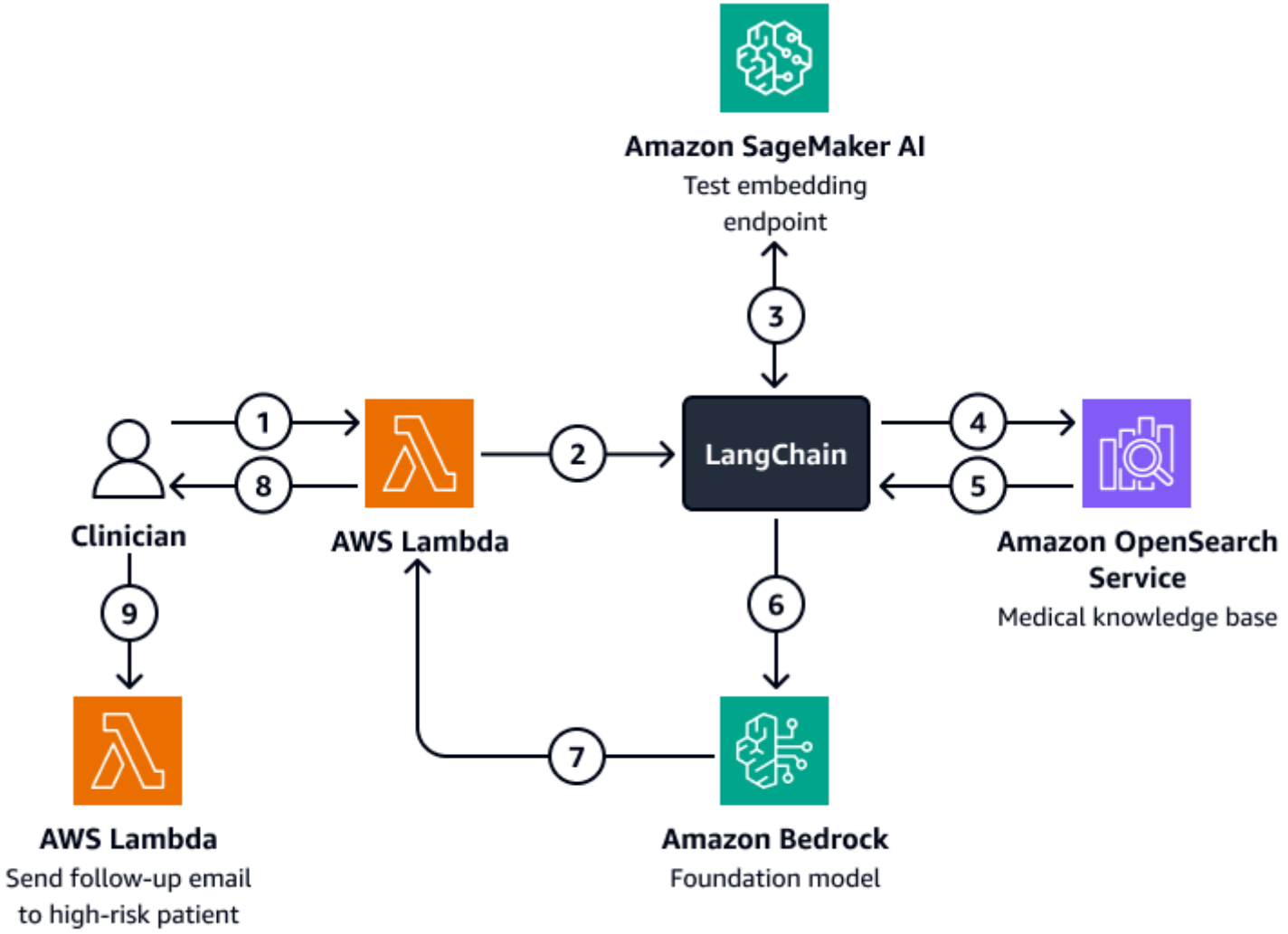
Successivamente, si calcola un punteggio di propensione alla riammissione ospedaliera per paziente. Questo punteggio riflette l'impatto netto delle tre analisi eseguite nelle fasi precedenti: esiti potenziali per i pazienti, comportamento del paziente nei confronti di farmaci e trattamenti e probabilità di riammissione del paziente. Aggregando il punteggio di propensione alla riammissione a livello di paziente a livello di specialità e quindi a livello ospedaliero, è possibile ottenere informazioni utili a medici, responsabili dell'assistenza e amministratori. Il punteggio di propensione alla riammissione ospedaliera aiuta a valutare le prestazioni complessive per struttura, specialità o condizione. Quindi, è possibile utilizzare questo punteggio per implementare interventi proattivi.

1. Assegna dei pesi a ciascuno dei diversi fattori (previsione dei risultati, probabilità di aderenza, riammissione). Di seguito sono riportati alcuni esempi di pesi:
   - Peso della previsione del risultato: 0,4
   - Peso di previsione dell'aderenza: 0,3
   - Peso della probabilità di riammissione: 0,3
2. Usa il seguente calcolo per calcolare il punteggio composito:

   `ReadadmissionPropensityScore` = (`OutcomeScore` × `OutcomeWeight`) + (`AdherenceScore` × `AdherenceWeight`) + (`ReadmissionLikelihoodScore` × `ReadmissionLikelihoodWeight`)
3. Assicurati che tutti i punteggi individuali siano sulla stessa scala, ad esempio da 0 a 1.
4. Definite le soglie di azione. Ad esempio, i punteggi superiori a 0,7 avviano gli avvisi.

Sulla base delle analisi di cui sopra e del punteggio di propensione alla riammissione di un paziente, i medici o i responsabili sanitari possono impostare avvisi per monitorare i singoli pazienti in base al punteggio calcolato. Se supera una soglia predefinita, vengono avvisati quando viene raggiunta tale soglia. Questo aiuta i responsabili sanitari a essere proattivi anziché reattivi nella creazione di piani di assistenza alla dimissione per i loro pazienti. Salva i punteggi degli esiti, del comportamento e della propensione alla riammissione dei pazienti in forma indicizzata in un database vettoriale di Amazon OpenSearch Service in modo che i responsabili dell'assistenza possano recuperarli senza problemi utilizzando un agente di intelligenza artificiale conversazionale.

Il diagramma seguente mostra il flusso di lavoro di un agente di intelligenza artificiale conversazionale che un medico o un responsabile sanitario può utilizzare per recuperare informazioni sugli esiti dei pazienti, sul comportamento previsto e sulla propensione alla riammissione. Gli utenti possono recuperare informazioni a livello di paziente, di reparto o di ospedale. L'agente AI recupera queste informazioni, che vengono archiviate in forma indicizzata in un database vettoriale Amazon OpenSearch Service. L'agente utilizza la query per recuperare i dati pertinenti e fornisce risposte personalizzate, comprese le azioni suggerite per i pazienti ad alto rischio di riammissione. In base al livello di rischio, l'operatore può anche impostare promemoria per pazienti e operatori sanitari.



Il diagramma mostra il flusso di lavoro seguente:

1. Il medico pone una domanda a un agente di intelligenza artificiale conversazionale, che ospita una funzione. AWS Lambda
2. La funzione Lambda avvia un LangChain agente.

3. Il LangChain l'agente invia la domanda dell'utente a un endpoint di incorporamento di testo Amazon SageMaker AI. L'endpoint incorpora la domanda.
4. Il LangChain l'agente passa la domanda incorporata a una knowledge base medica in Amazon OpenSearch Service.
5. Amazon OpenSearch Service restituisce le informazioni specifiche più pertinenti alla richiesta dell'utente al LangChain agente.
6. Il LangChain agent invia la query e il contesto recuperato dalla knowledge base a un modello di base Amazon Bedrock.
7. Il modello di base Amazon Bedrock genera una risposta e la invia alla funzione Lambda.
8. La funzione Lambda restituisce la risposta al medico.
9. Il medico avvia una funzione Lambda che invia un'e-mail di follow-email a un paziente ad alto rischio di riammissione.

Allineamento al Well-Architected AWS Framework

L'architettura per tracciare il comportamento dei pazienti e prevedere i tassi di riammissione ospedaliera integra Servizi AWS, rappresenta grafici delle conoscenze mediche e migliora i risultati sanitari, allineandosi LLMs ai sei pilastri del Well-Architected Framework:AWS

- Eccellenza operativa: la soluzione è un sistema automatizzato e disaccoppiato che utilizza Amazon Bedrock e AWS Lambda per avvisi in tempo reale.
- Sicurezza: questa soluzione è progettata per rispettare le normative sanitarie, come l'HIPAA. Puoi anche implementare la crittografia, il controllo granulare degli accessi e i guardrail Amazon Bedrock per proteggere i dati dei pazienti.
- Affidabilità: l'architettura utilizza sistemi serverless e con tolleranza ai guasti. Servizi AWS
- Efficienza delle prestazioni: Amazon OpenSearch Service e il fine-tuned LLMs possono fornire previsioni rapide e accurate.
- Ottimizzazione dei costi: le tecnologie e i modelli serverless aiutano a ridurre al minimo i pay-per-inference costi. Sebbene l'utilizzo di un LLM ottimizzato possa comportare costi aggiuntivi, il modello utilizza un approccio RAG che riduce i dati e il tempo di calcolo necessari per il processo di messa a punto.
- Sostenibilità: l'architettura riduce al minimo il consumo di risorse attraverso l'uso di un'infrastruttura serverless. Supporta inoltre operazioni sanitarie efficienti e scalabili.

Caso d'uso: gestione e miglioramento delle competenze del personale sanitario

L'implementazione di strategie di trasformazione e miglioramento delle competenze dei talenti aiuta la forza lavoro a rimanere abile nell'uso di nuove tecnologie e pratiche nei servizi medici e sanitari. Le iniziative proattive di miglioramento delle competenze assicurano che gli operatori sanitari possano fornire un'assistenza ai pazienti di alta qualità, ottimizzare l'efficienza operativa e rimanere conformi agli standard normativi. Inoltre, la trasformazione dei talenti promuove una cultura dell'apprendimento continuo. Questo è fondamentale per adattarsi al mutevole panorama sanitario e affrontare le sfide emergenti in materia di salute pubblica. Gli approcci formativi tradizionali, come la formazione in aula e i moduli di apprendimento statico, offrono contenuti uniformi a un vasto pubblico. Spesso mancano percorsi di apprendimento personalizzati, fondamentali per soddisfare le esigenze specifiche e i livelli di competenza dei singoli professionisti. Questa one-size-fits-all strategia può portare al disimpegno e alla conservazione non ottimale delle conoscenze.

Di conseguenza, le organizzazioni sanitarie devono adottare soluzioni innovative, scalabili e basate sulla tecnologia in grado di determinare il divario per ciascuno dei dipendenti nello stato attuale e in quello futuro potenziale. Queste soluzioni dovrebbero consigliare percorsi di apprendimento iperpersonalizzati e il giusto set di contenuti formativi. Ciò prepara efficacemente la forza lavoro per le future cure sanitarie.

Nel settore sanitario, puoi applicare l'intelligenza artificiale generativa per aiutarti a comprendere e migliorare le competenze della tua forza lavoro. Attraverso la connessione di modelli linguistici di grandi dimensioni (LLMs) e strumenti di recupero avanzati, le organizzazioni possono comprendere quali competenze possiedono attualmente e identificare le competenze chiave che potrebbero essere necessarie in futuro. Queste informazioni aiutano a colmare il divario assumendo nuovi lavoratori e migliorando le competenze della forza lavoro attuale. Utilizzando Amazon Bedrock e Knowledge Graphs, le organizzazioni sanitarie possono sviluppare applicazioni specifiche del dominio che facilitano l'apprendimento continuo e lo sviluppo delle competenze.

Le conoscenze fornite da questa soluzione ti aiutano a gestire efficacemente i talenti, ottimizzare le prestazioni della forza lavoro, promuovere il successo organizzativo, identificare le competenze esistenti e elaborare una strategia per i talenti. Questa soluzione può aiutarti a svolgere queste attività in settimane anziché mesi.

Panoramica della soluzione

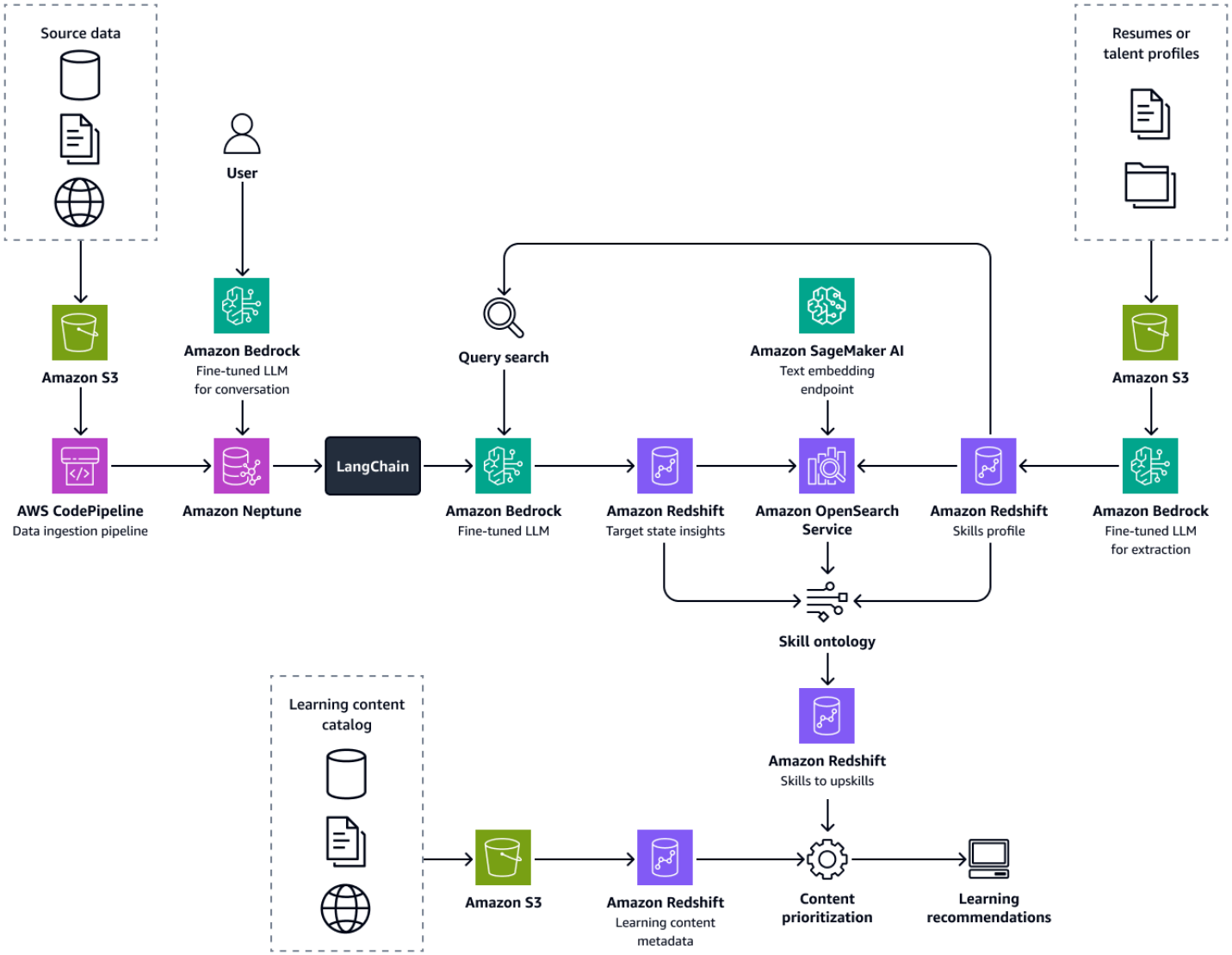
Questa soluzione è un framework per la trasformazione dei talenti nel settore sanitario che comprende i seguenti componenti:

- Analisi intelligente del curriculum: questo componente è in grado di leggere il curriculum di un candidato ed estrarre con precisione le informazioni sul candidato, comprese le competenze. Soluzione intelligente di estrazione delle informazioni creata utilizzando il modello Llama 2 ottimizzato in Amazon Bedrock su un set di dati di formazione proprietario che copre curriculum e profili di talenti di oltre 19 settori. Questo processo basato sul LLM consente di risparmiare centinaia di ore automatizzando il processo di revisione manuale dei curriculum e associando i migliori candidati ai ruoli vacanti.
- Knowledge graph: un knowledge graph basato su Amazon Neptune, un archivio unificato di informazioni sui talenti che include la tassonomia dei ruoli e delle competenze dell'organizzazione e del settore, che cattura la semantica dei talenti sanitari utilizzando definizioni di competenze, ruoli e relative proprietà, relazioni e vincoli logici.
- Ontologia delle competenze: la scoperta delle prossimità tra le competenze dei candidati e le competenze ideali dello stato attuale o futuro (recuperate utilizzando un grafico della conoscenza) viene ottenuta attraverso algoritmi ontologici che misurano la somiglianza semantica tra le abilità dei candidati e le abilità dello stato obiettivo.
- Percorso e contenuti di apprendimento: questo componente è un motore di raccomandazione formativa in grado di consigliare i contenuti didattici giusti tra un catalogo di materiali didattici di qualsiasi fornitore in base alle lacune di competenze identificate. Identificazione dei percorsi di miglioramento delle competenze più ottimali per ciascun candidato analizzando le lacune nelle competenze e consigliando contenuti di apprendimento prioritari, per consentire uno sviluppo professionale continuo e senza interruzioni per ogni candidato durante la transizione verso un nuovo ruolo.

Questa soluzione automatizzata basata sul cloud è alimentata da servizi di machine learning, knowledge graphs e Retrieval Augmented Generation (LLMsRAG). Può scalare per elaborare decine o migliaia di curriculum in un lasso di tempo minimo, creare profili istantanei dei candidati, identificare le lacune nel loro stato attuale o potenziale futuro e quindi consigliare in modo efficiente i contenuti formativi giusti per colmare queste lacune.

L'immagine seguente mostra il end-to-end flusso del framework. La soluzione è costruita e ottimizzata in LLMs Amazon Bedrock. Questi LLMs recuperano dati dalla knowledge base di

talenti sanitari di Amazon Neptune. Gli algoritmi basati sui dati forniscono consigli per percorsi di apprendimento ottimali per ogni candidato.



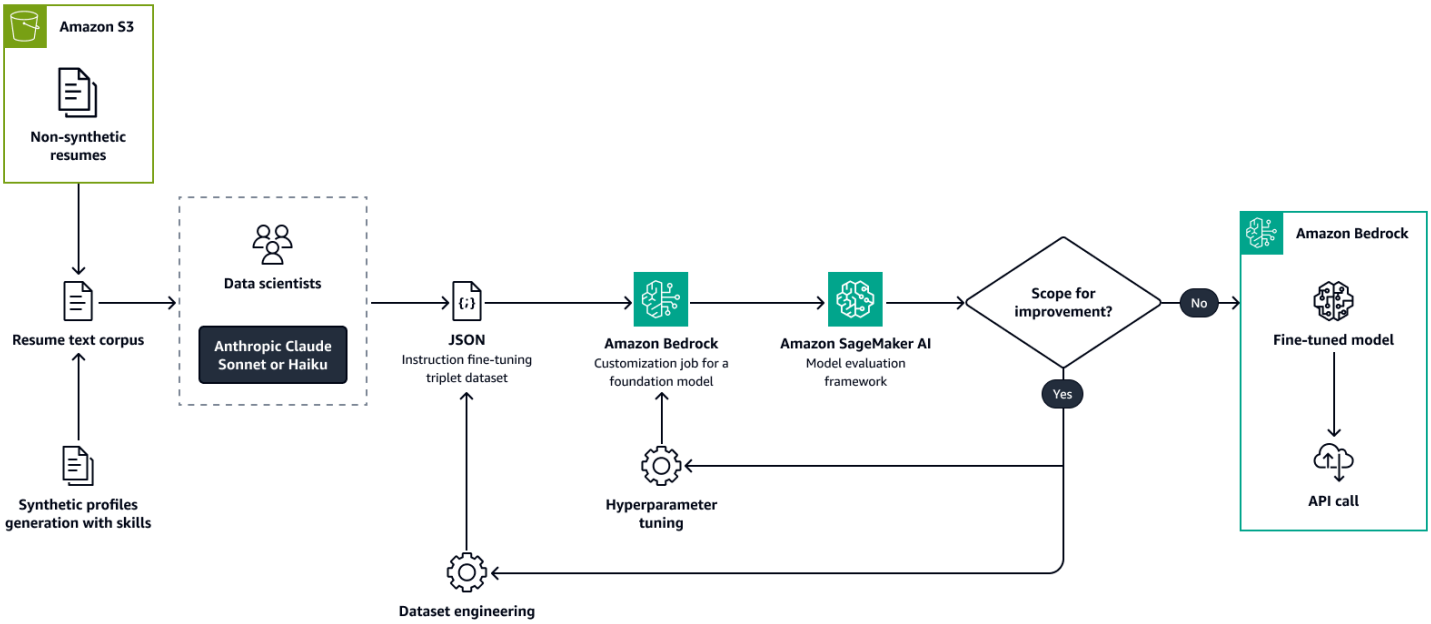
La creazione di questa soluzione prevede i seguenti passaggi:

- Fase 1: Estrazione di informazioni sui talenti e creazione di un profilo di competenze
- Fase 2: Scoprire role-to-skill la rilevanza grazie a un grafico della conoscenza
- Fase 3: Identificare le lacune nelle competenze e consigliare la formazione

Fase 1: Estrazione di informazioni sui talenti e creazione di un profilo di competenze

Innanzitutto, ottimizzi un modello linguistico di grandi dimensioni, come Llama 2, in Amazon Bedrock con un set di dati personalizzato. Questo adatta l'LLM al caso d'uso. Durante la formazione, estrai in modo accurato e coerente gli attributi chiave dei talenti dai curriculum dei candidati o da profili di talento simili. Questi attributi di talento includono competenze, titolo di ruolo attuale, titoli di esperienza con intervalli di tempo, istruzione e certificazioni. Per ulteriori informazioni, consulta Personalizza il tuo modello per migliorarne le prestazioni per il tuo caso d'uso nella documentazione di Amazon Bedrock.

L'immagine seguente mostra il processo di ottimizzazione di un modello di analisi dei curriculum utilizzando Amazon Bedrock. Sia i curriculum reali che quelli creati sinteticamente vengono passati a un LLM per estrarre le informazioni chiave. Un gruppo di data scientist convalida le informazioni estratte confrontandole con il testo originale non elaborato. Le informazioni estratte vengono quindi concatenate utilizzando il chain-of-thoughtprompting e il testo originale per ricavare un set di dati di addestramento da perfezionare. Questo set di dati viene quindi passato a un processo di personalizzazione di Amazon Bedrock, che perfeziona il modello. Un processo batch di Amazon SageMaker AI esegue un framework di valutazione del modello che valuta il modello perfezionato. Se il modello necessita di miglioramenti, il processo viene eseguito nuovamente con più dati o iperparametri diversi. Dopo che la valutazione soddisfa gli standard, esegui l'hosting del modello personalizzato tramite il throughput fornito da Amazon Bedrock.



Fase 2: Scoprire role-to-skill la rilevanza grazie a un grafico della conoscenza

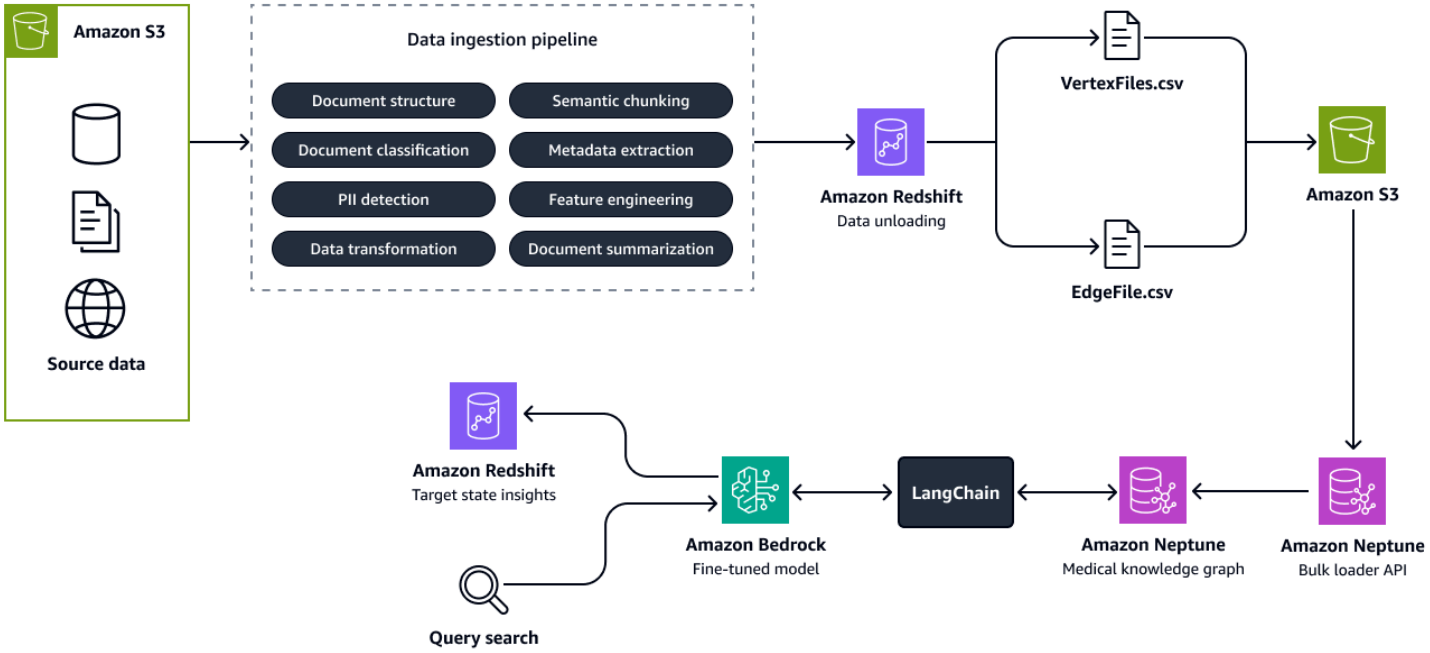
Successivamente, creerai un grafico della conoscenza che racchiuda le competenze e la tassonomia dei ruoli della tua organizzazione e di altre organizzazioni del settore sanitario. Questa base di conoscenze arricchita proviene dai dati aggregati di talenti e organizzazioni in Amazon Redshift. Puoi raccogliere dati sui talenti da una serie di fornitori di dati sul mercato del lavoro e da fonti di dati strutturate e non strutturate specifiche dell'organizzazione, come sistemi di pianificazione delle risorse aziendali (ERP), un sistema informativo per le risorse umane (HRIS), curriculum dei dipendenti, descrizioni delle mansioni e documenti sull'architettura dei talenti.

Crea il knowledge graph su Amazon Neptune. I nodi rappresentano competenze e ruoli, mentre i bordi rappresentano le relazioni tra di essi. Arricchisci questo grafico con metadati per includere dettagli come il nome dell'organizzazione, il settore, la categoria professionale, il tipo di competenza, il tipo di ruolo e i tag del settore.

Successivamente, si sviluppa un'applicazione Graph Retrieval Augmented Generation (Graph RAG). Graph RAG è un approccio RAG che recupera i dati da un database grafico. I seguenti sono i componenti dell'applicazione Graph RAG:

- Integrazione con un LLM in Amazon Bedrock: l'applicazione utilizza un LLM in Amazon Bedrock per la comprensione del linguaggio naturale e la generazione di query. Gli utenti possono interagire con il sistema utilizzando il linguaggio naturale. Ciò lo rende accessibile alle parti interessate non tecniche.
- Orchestrazione e recupero delle informazioni: utilizzo o LlamaIndexLangChainorchestratori per facilitare l'integrazione tra LLM e il knowledge graph di Neptune. Gestiscono il processo di conversione delle query in linguaggio naturale in query OpenCypher. Quindi, eseguono le interrogazioni sul Knowledge Graph. Usa prompt engineering per istruire l'LLM sulle migliori pratiche per la creazione di query OpenCypher. Questo aiuta a ottimizzare le interrogazioni per recuperare il sottografo pertinente, che contiene tutte le entità e le relazioni pertinenti relative ai ruoli e alle competenze richiesti.
- Generazione di informazioni: l'LLM di Amazon Bedrock elabora i dati del grafico recuperati. Genera informazioni dettagliate sullo stato attuale e proietta gli stati futuri per il ruolo richiesto e le competenze associate.

L'immagine seguente mostra i passaggi per creare un Knowledge Graph a partire dai dati di origine. I dati di origine strutturati e non strutturati vengono passati alla pipeline di inserimento dei dati. La pipeline estrae e trasforma le informazioni in una formazione di carichi di massa CSV compatibile con Amazon Neptune. L'API bulk loader carica i file CSV archiviati in un bucket Amazon S3 nel knowledge graph di Neptune. Per le domande degli utenti relative al talento, allo stato futuro, ai ruoli o alle competenze pertinenti, il LLM ottimizzato di Amazon Bedrock interagisce con il knowledge graph attraverso un LangChain orchestratore. L'orchestratore recupera il contesto pertinente dal knowledge graph e invia le risposte alla tabella degli approfondimenti in Amazon Redshift. Il LangChain orchestrator, come Graph QAChain, converte la query in linguaggio naturale dell'utente in una query OpenCypher per interrogare il knowledge graph. Il modello ottimizzato di Amazon Bedrock genera una risposta basata sul contesto recuperato.



Fase 3: Identificare le lacune nelle competenze e consigliare la formazione

In questa fase, si calcola con precisione la prossimità tra lo stato attuale di un operatore sanitario e i potenziali ruoli statali futuri. A tal fine, si esegue un'analisi dell'affinità delle competenze confrontando le competenze individuali con il ruolo lavorativo. In un database vettoriale di Amazon OpenSearch Service, memorizzi le informazioni sulla tassonomia delle competenze e i metadati delle competenze, come la descrizione delle abilità, il tipo di abilità e i cluster di competenze. Utilizza un modello di incorporamento Amazon Bedrock, come i modelli Amazon Titan Text Embeddings, per incorporare

l'abilità chiave identificata nei vettori. Tramite una ricerca vettoriale, recuperi le descrizioni delle competenze statali attuali e delle competenze dello stato target ed esegui un'analisi ontologica. L'analisi fornisce punteggi di prossimità tra le coppie di abilità dello stato attuale e dello stato bersaglio. Per ogni coppia, si utilizzano i punteggi ontologici calcolati per identificare le lacune nelle affinità di abilità. Quindi, consiglierai il percorso ottimale per il miglioramento delle competenze, che il candidato può prendere in considerazione durante le transizioni di ruolo.

Per ogni ruolo, consigliare i contenuti didattici corretti per il miglioramento delle competenze o la riqualificazione implica un approccio sistematico che inizia con la creazione di un catalogo completo di contenuti formativi. Questo catalogo, archiviato in un database Amazon Redshift, aggrega i contenuti di vari provider e include metadati, come la durata del contenuto, il livello di difficoltà e la modalità di apprendimento. Il passaggio successivo consiste nell'estrarre le competenze chiave offerte da ciascun contenuto e quindi mapparle alle competenze individuali richieste per il ruolo target. È possibile ottenere questa mappatura analizzando la copertura fornita dai contenuti attraverso un'analisi della prossimità delle competenze. Questa analisi valuta quanto strettamente le competenze insegnate dal contenuto siano in linea con le competenze desiderate per il ruolo. I metadati svolgono un ruolo fondamentale nella selezione dei contenuti più appropriati per ciascuna competenza, assicurando che gli studenti ricevano consigli personalizzati adatti alle loro esigenze di apprendimento. Utilizzalo LLMs in Amazon Bedrock per estrarre competenze dai metadati dei contenuti, eseguire l'ingegneria delle funzionalità e convalidare i consigli sui contenuti. Ciò migliora l'accuratezza e la pertinenza nel processo di miglioramento delle competenze o riqualificazione.

Allineamento al Well-Architected AWS Framework

La soluzione è in linea con tutti e sei i pilastri del Well-Architected AWS Framework:

- Eccellenza operativa: una pipeline modulare e automatizzata migliora l'eccellenza operativa. I componenti chiave della pipeline sono disaccoppiati e automatizzati, il che consente aggiornamenti più rapidi dei modelli e un monitoraggio più semplice. Inoltre, le pipeline di formazione automatizzate supportano rilasci più rapidi di modelli ottimizzati.
- Sicurezza: questa soluzione elabora informazioni sensibili e di identificazione personale (PII), come i dati contenuti nei curriculum e nei profili dei talenti. In AWS Identity and Access Management (IAM), implementa politiche di controllo degli accessi dettagliate e assicurati che solo il personale autorizzato abbia accesso a questi dati.
- Affidabilità: la soluzione utilizza Servizi AWS, ad esempio, Neptune, Amazon Bedrock e Service OpenSearch , che forniscono tolleranza agli errori, alta disponibilità e accesso ininterrotto alle informazioni anche in caso di forte domanda.

- Efficienza delle prestazioni: ottimizzati in LLMs Amazon Bedrock and OpenSearch Service, i database vettoriali sono progettati per elaborare in modo rapido e preciso set di dati di grandi dimensioni al fine di fornire consigli di apprendimento tempestivi e personalizzati.
- Ottimizzazione dei costi: questa soluzione utilizza un approccio RAG, che riduce la necessità di un addestramento preliminare continuo dei modelli. Invece di perfezionare ripetutamente l'intero modello, il sistema perfeziona solo processi specifici, come l'estrazione di informazioni dai curriculum e la strutturazione degli output. Ciò si traduce in significativi risparmi sui costi. Riducendo al minimo la frequenza e la portata dei modelli di formazione ad alta intensità di risorse e utilizzando i servizi pay-per-use cloud, le organizzazioni sanitarie possono ottimizzare i costi operativi mantenendo prestazioni elevate.
- Sostenibilità: questa soluzione utilizza servizi scalabili e nativi del cloud che allocano le risorse di elaborazione in modo dinamico. Ciò riduce il consumo di energia e l'impatto ambientale, pur continuando a supportare iniziative di trasformazione dei talenti su larga scala e ad alta intensità di dati.

Sviluppo e orchestrazione di soluzioni di intelligenza artificiale generativa per l'assistenza sanitaria

Per creare le soluzioni illustrate in questa guida, è necessario creare un'architettura RAG che utilizzi la tecnologia fine-tuned LLMs per fornire agli operatori sanitari dati aumentati sui pazienti, approfondimenti clinici e diagnostici e previsioni degli esiti dei pazienti. Ciò richiede l'integrazione di più strumenti per creare un flusso di lavoro Servizi AWS coeso ed efficiente. In questa sezione vengono descritti i seguenti argomenti:

- Amazon Q Developer— Usa Amazon Q Developer per rispondere a domande tecniche ed errori di codice durante il processo di sviluppo.
- Design RAG multi-retriever— Progetta e implementa soluzioni RAG che utilizzano più retriever per recuperare il contesto medico corretto per la domanda dell'utente.
- ReAct agenti— Implementa agenti che combinano il ragionamento con l'azione dinamica.

Amazon Q Developer

Quando si crea una soluzione di intelligenza artificiale generativa, può essere difficile creare agenti di intelligenza artificiale e connettere i servizi chiave. Tuttavia, Amazon Q Developer aiuta i data scientist e gli ingegneri di intelligenza artificiale fornendo l'accesso a un assistente AI generativo avanzato. Amazon Q può rispondere in modo rapido e preciso alle domande degli utenti e agli errori di codice, il che può aiutarti a ottimizzare il processo di sviluppo LLM. Amazon Q offre vantaggi significativi agli sviluppatori che creano applicazioni che utilizzano i modelli di base di Amazon Bedrock. Può semplificare i flussi di lavoro e migliorare la qualità del codice. Automatizza la generazione di script Python e configurazioni Infrastructure as Code (IaC), riducendo significativamente i tempi e gli sforzi di sviluppo. Grazie a funzionalità di refactoring avanzate, Amazon Q può migliorare le prestazioni del codice, identificare le vulnerabilità di sicurezza e garantire che gli sviluppatori aderiscano alle migliori pratiche. Inoltre, facilita l'apprendimento e l'adozione per i principianti fornendo suggerimenti e spiegazioni sensibili al contesto, rendendo le attività di codifica complesse più accessibili ed efficienti.

Design RAG multi-retriever

In un'applicazione di intelligenza artificiale generativa, una pipeline RAG multi-retriever può recuperare in modo efficiente informazioni da più fonti di dati per aiutare gli operatori sanitari e i medici a rispondere a domande mediche. Questa pipeline utilizza diversi tipi di retriever per estrarre dati pertinenti da diverse basi di conoscenza. Ogni retriever è specializzato nel recupero di un particolare tipo di informazioni, come anamnesi dei pazienti, approfondimenti diagnostici, note cliniche o contenuti tratti da ricerche mediche e testi accademici.

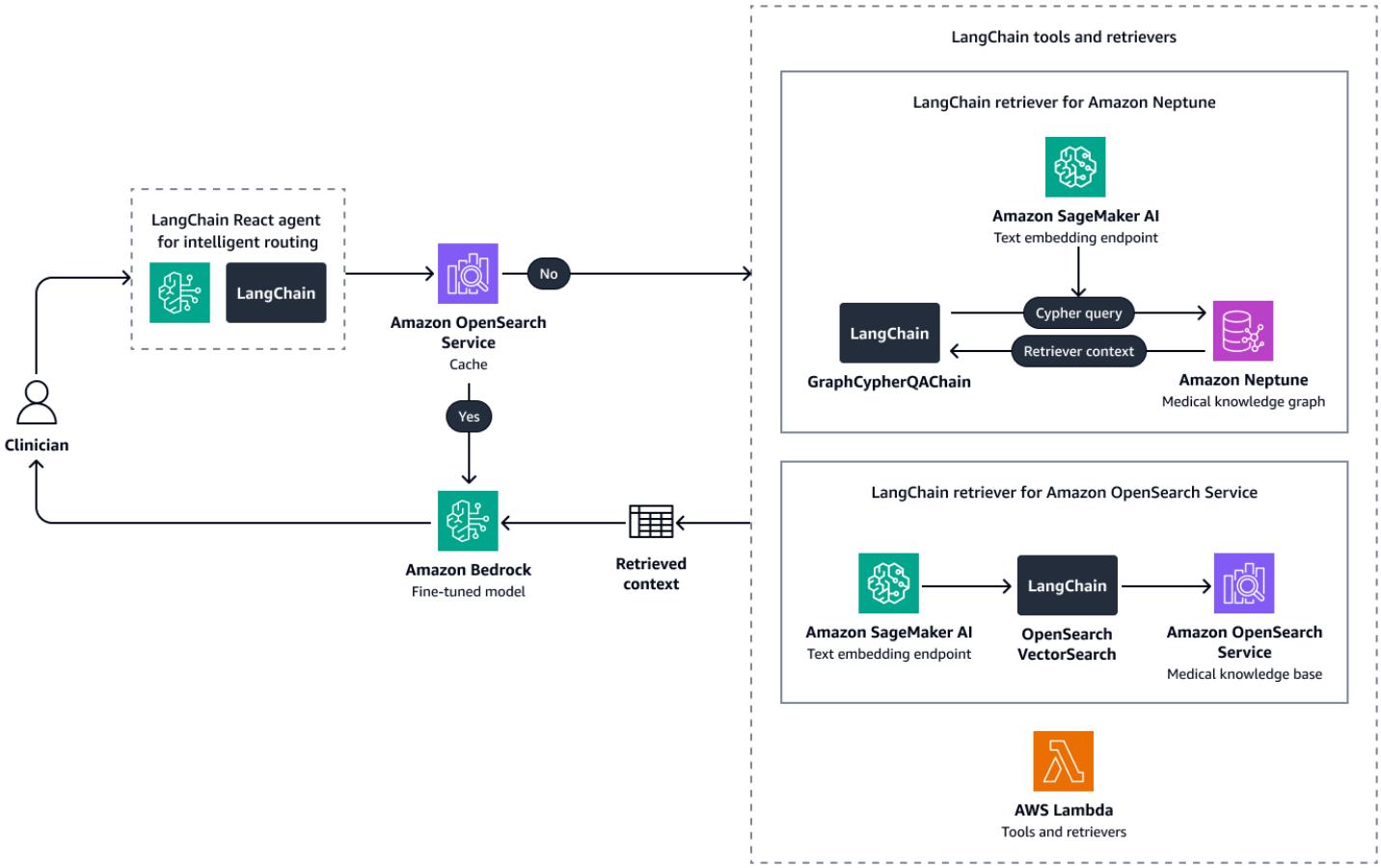
Utilizza la natura dei dati e i requisiti specifici dell'applicazione per determinare quale sia la knowledge base di backend corretta per il tuo caso d'uso. Un database vettoriale di Amazon OpenSearch Service è ideale per grandi volumi di dati sanitari non strutturati o semistrutturati, tra cui riepiloghi di valutazione delle diagnosi di immagini, riepiloghi delle dimissioni, report clinici, ricerche mediche e contenuti di testo accademico. D'altra parte, un servizio di database grafico, come Amazon Neptune, può essere ideale per i casi d'uso nel settore sanitario che richiedono un'esplorazione approfondita delle relazioni temporali tra entità, come paziente, anamnesi del paziente, operatore sanitario, farmaci, sintomi e trattamenti.

Un componente fondamentale di questa pipeline è la previsione dell'intento delle interrogazioni degli utenti. Ciò garantisce che il sistema indirizzi la query alla catena di retriever corretta. Ad esempio, se un medico chiede informazioni sulla storia del trattamento di un paziente, sui sintomi, sull'interazione con l'ospedale, sulla probabilità di riammissione in ospedale o sui potenziali esiti del paziente, il modulo di previsione dell'intento di interrogazione identifica tale intento. Indirizza la richiesta alla catena di recupero che può recuperare le cartelle cliniche dei pazienti o i dati cronologici del trattamento dal grafico delle conoscenze mediche. In alternativa, se la domanda riguarda l'individuazione di malattie, valutazioni diagnostiche specifiche o dettagli di procedure cliniche specifiche dai libri di testo accademici, la query viene indirizzata alla catena di retriever che può recuperare queste informazioni dal database Service Vector. OpenSearch È possibile utilizzare la funzionalità di chiamata degli strumenti di LangChain per associare uno strumento personalizzato ad Amazon Bedrock LLM in grado di classificare una domanda dell'utente in intenti predefiniti.

Questo sistema RAG multi-retriever include LangChain agenti progettati per gestire l'accesso alla knowledge base specifica. È possibile utilizzare… LangChain per orchestrare l'interazione tra Amazon Bedrock LLM, i diversi retriever e gli strumenti. LangChain include una classe tool-calling che consente di creare strumenti personalizzati, come un classificatore di intenti, un retriever per Neptune, un retriever per OpenSearch Service o qualsiasi altro strumento che può essere sviluppato per classificare l'intento dell'utente e accedere ai dati da una specifica base di conoscenza in un

formato strutturato. Quindi fornite questi strumenti alla classe per creare un agente Reasoning and Acting (). ReAct L' ReAct agente elabora la domanda dell'utente, pianifica i passaggi sequenziali per rispondere alla domanda, quindi esegue iterativamente gli strumenti disponibili ed elabora le risposte dello strumento per rispondere finalmente alla richiesta dell'utente.

L'immagine seguente mostra come funziona un sistema RAG multi-retriever progettato per un recupero efficiente delle conoscenze e una risoluzione intelligente delle query. A LangChain ReAct l'agente analizza l'intento dell'utente, formula un piano strutturato di esecuzione e seleziona gli strumenti di recupero più pertinenti. Il sistema interroga una cache di domande precedenti e verifica la presenza di domande simili in base ad attributi chiave, come l'ID del paziente, la condizione medica e la data della visita. Se viene trovata una domanda molto simile, la risposta corrispondente viene recuperata direttamente. In caso contrario, l'agente esegue il retriever appropriato. Per recuperare informazioni incentrate sul paziente, come la storia del trattamento, i sintomi, le interazioni ospedaliere o la probabilità di riammissione, il sistema utilizza un graph retriever. Per le valutazioni diagnostiche, le procedure cliniche e i risultati medici strutturati, l'agente utilizza un recuperatore di database vettoriali. In scenari che richiedono una combinazione di conoscenze contestuali provenienti da entrambi gli archivi di dati, per generare una risposta completa, il sistema utilizza una strategia di recupero ibrida che integra i risultati sia del knowledge graph che del database vettoriale.



ReAct agenti

Gli agenti Reasoning e Acting (ReAct) sono progettati per applicazioni RAG multiformi. Questi agenti forniscono una potente combinazione di ragionamento e azione dinamica, in particolare per applicazioni complesse che coinvolgono step-by-step flussi di lavoro logici di recupero delle informazioni. Per ulteriori informazioni, vedere ReAct: Synergizing Reasoning and Acting in Language Models.

In contesti medici e sanitari, le domande di un medico o di un medico sono spesso multiformi. Ad esempio, un medico potrebbe chiedere «Quali trattamenti sono stati somministrati a pazienti simili affetti sia da ipertensione che da diabete di tipo 2?» Dopo aver identificato l'intento dell'utente, ovvero quello di recuperare i trattamenti per l'ipertensione e il diabete di tipo 2, l'agente di intelligenza artificiale deve suddividere questa query in attività secondarie e quindi scegliere la strategia di recupero più efficiente. In questo caso, l'agente di intelligenza artificiale deve identificare i nodi più rilevanti (come l'età del paziente, il sesso, le condizioni, i trattamenti e i farmaci) e quindi interrogare il grafico per individuare tali entità e i relativi attributi e relazioni. ReAct gli agenti sono molto utili perché

combinano la capacità di ragionamento (inferenza logica) di un LLM con un'azione (interrogare o interagire con risorse o basi di conoscenza esterne).

Per rispondere alla domanda dell'utente «Quali trattamenti sono stati somministrati a pazienti simili affetti sia da ipertensione che da diabete di tipo 2?» , l'esempio seguente illustra come funziona un ReAct agente:

1. Ragionamento dell'agente: l' ReAct agente deduce che la domanda implica il recupero di informazioni sulle condizioni (diabete e ipertensione). Considera l'età del paziente, i trattamenti, i farmaci e il periodo da analizzare.
2. Azione dell'agente: l'agente utilizza OpenCypher per interrogare il Knowledge Graph sui trattamenti specifici per il diabete di tipo 2 e l'ipertensione. Inoltre, recupera i farmaci somministrati, le date delle visite ospedaliere, gli effetti collaterali dei farmaci, gli esiti noti dei pazienti e i dati incrociati per pazienti simili (ad esempio pazienti dello stesso sesso ed età).
3. Osservazione degli agenti: dal Knowledge Graph, l'agente recupera i dati tabulari degli ultimi sei mesi sui trattamenti somministrati a pazienti affetti sia da ipertensione che da diabete di tipo 2.
4. Ragionamento dell'agente: per classificare i risultati dei record recuperati, l'agente identifica attributi importanti, come la frequenza, gli effetti collaterali dei farmaci o gli esiti noti dei pazienti.
5. Azione dell'agente: l'agente riordina i record in base agli attributi identificati e alla logica predefinita impartita tramite il prompt di sistema.
6. Generazione di risposte: il LLM in Amazon Bedrock genera una risposta basata sul contesto preparato dall' ReAct agente.

Valutazione di soluzioni di intelligenza artificiale generativa per l'assistenza sanitaria

Valutare le soluzioni di intelligenza artificiale per il settore sanitario che crei è fondamentale per garantire che siano efficaci, affidabili e scalabili negli ambienti medici del mondo reale. Utilizza un approccio sistematico per valutare le prestazioni di ogni componente della soluzione. Di seguito è riportato un riepilogo delle metodologie e delle metriche che è possibile utilizzare per valutare la soluzione.

Argomenti

Valutazione dell'estrazione delle informazioni

Valuta le prestazioni delle soluzioni di estrazione delle informazioni, come l'intelligente parser di curriculum e l'estrattore di entità personalizzato. È possibile misurare l'allineamento delle risposte di queste soluzioni utilizzando un set di dati di test. Se non disponi di un set di dati che copra profili versatili di talenti sanitari e cartelle cliniche dei pazienti, puoi creare un set di dati di test personalizzato utilizzando la capacità di ragionamento di un LLM. Ad esempio, è possibile utilizzare un modello di parametri di grandi dimensioni, come Anthropic Claude modelli, per generare un set di dati di test.

Di seguito sono riportate tre metriche chiave che è possibile utilizzare per valutare i modelli di estrazione delle informazioni:

- Precisione e completezza: queste metriche valutano la misura in cui l'output ha acquisito le informazioni corrette e complete presenti nei dati di base. Ciò implica il controllo sia della correttezza delle informazioni estratte sia della presenza di tutti i dettagli pertinenti nelle informazioni estratte.
- Somiglianza e pertinenza: queste metriche valutano le somiglianze semantiche, strutturali e contestuali tra l'output e i dati di base (la somiglianza) e il grado in cui l'output si allinea e affronta il contenuto, il contesto e l'intento dei dati di base relativi alla verità (la rilevanza).

- Frequenza di richiamo o acquisizione adeguata: queste percentuali determinano empiricamente quanti dei valori attuali nei dati di base della verità sono stati identificati correttamente dal modello. La tariffa dovrebbe includere una penalizzazione per tutti i valori falsi estratti dal modello.
- Punteggio di precisione: il punteggio di precisione consente di determinare quanti falsi positivi sono presenti nelle previsioni, rispetto ai veri positivi. Ad esempio, puoi utilizzare metriche di precisione per misurare la correttezza della competenza acquisita.

Valutazione delle soluzioni RAG con più retriever

Per valutare l'efficacia con cui il sistema recupera le informazioni pertinenti e l'efficacia con cui utilizza tali informazioni per generare risposte accurate e contestualmente appropriate, puoi utilizzare le seguenti metriche:

- Pertinenza della risposta: misura la pertinenza della risposta generata, che utilizza il contesto recuperato, rispetto alla query originale.
- Precisione del contesto: rispetto al totale dei risultati recuperati, valuta la percentuale di documenti o frammenti recuperati pertinenti alla query. Una maggiore precisione del contesto indica che il meccanismo di recupero è efficace nella selezione delle informazioni pertinenti.
- Fedeltà: valuta la precisione con cui la risposta generata riflette le informazioni nel contesto recuperato. In altre parole, misura se la risposta rimane fedele alle informazioni di origine.

Valutazione di una soluzione utilizzando un LLM

Puoi utilizzare una tecnica chiamata LLM- as-a-judge per valutare le risposte testuali della tua soluzione di intelligenza artificiale generativa. Implica l'utilizzo LLMs per valutare e valutare le prestazioni degli output del modello. Questa tecnica utilizza le funzionalità di Amazon Bedrock per fornire giudizi su vari attributi, come la qualità della risposta, la coerenza, l'aderenza, l'accuratezza e la completezza rispetto alle preferenze umane o ai dati di base. Utilizzi tecniche chain-of-thought (CoT) e few-shot prompting per una valutazione completa. Il prompt indica all'LLM di valutare la risposta generata con la rubrica del punteggio e i pochi esempi contenuti nel prompt dimostrano l'effettivo processo di valutazione. Il prompt include anche le linee guida da seguire per il valutatore LLM. Ad esempio, potresti prendere in considerazione l'utilizzo di una o più delle seguenti tecniche di valutazione che utilizzano un LLM per giudicare le risposte generate:

- Confronto a coppie: poni al valutatore LLM una domanda medica e risposte multiple generate da diverse versioni iterative dei sistemi RAG che hai creato. Chiedi al valutatore LLM di determinare la risposta migliore in base alla qualità della risposta, alla coerenza e all'aderenza alla domanda originale.
- Classificazione a risposta singola: questa tecnica è ideale per i casi d'uso in cui è necessario valutare l'accuratezza della categorizzazione, come la classificazione degli esiti dei pazienti, la categorizzazione del comportamento del paziente, la probabilità di riammissione del paziente e la categorizzazione del rischio. Utilizzate lo strumento di valutazione LLM per analizzare isolatamente la categorizzazione o la classificazione individuale e valutare le argomentazioni fornite rispetto a dati fondati sulla realtà.
- Valutazione guidata da riferimenti: fornisci al valutatore LLM una serie di domande mediche che richiedono risposte descrittive. Crea risposte di esempio a queste domande, come risposte di riferimento o risposte ideali. Chiedi al valutatore LLM di confrontare la risposta generata dal LLM con le risposte di riferimento o le risposte ideali e chiedi al valutatore LLM di valutare la risposta generata per accuratezza, completezza, somiglianza, pertinenza o altri attributi. Questa tecnica consente di valutare se le risposte generate sono in linea con una risposta standard o esemplare ben definita.

Risorse

AWS documentazione

- Documentazione Amazon Bedrock
- Documentazione di Amazon Neptune
- Documentazione OpenSearch del servizio Amazon
- Applicazione del AWS Well-Architected Framework per Amazon Neptune (guida prescrittiva)AWS
- Best practice operative per Amazon OpenSearch Service (documentazione OpenSearch del servizio)
- Utilizzo di Amazon Comprehend Medical LLMs e per il settore sanitario e delle scienze biologiche AWS (Prescriptive Guidance)

AWS post di blog

- Crea applicazioni di intelligenza artificiale generativa basate su RAG e agenti con il nuovo modello Amazon Titan Text Premier, disponibile in Amazon Bedrock
- Completa l'intelligence commerciale creando un Knowledge Graph da un data warehouse con Amazon Neptune
- Utilizzo dei knowledge graphs per creare applicazioni GraphRag con Amazon Bedrock e Amazon Neptune

Altre risorse

- Integrazione della generazione basata sul recupero con modelli linguistici di grandi dimensioni in Nefrologia: applicazioni pratiche avanzate (Central, National Library of Medicine) PubMed
- Introduzione a LangChain (LangChain documentazione)

Collaboratori

Creazione di testi

- Nitu Nivedita, amministratore delegato, responsabile dell'intelligenza artificiale, dati e intelligenza artificiale, Accenture
- Manoj Appully, fondatore e CTO di Cadiem
- Conor Folan, consulente per dati e intelligenza artificiale, Accenture
- Deepak Krishna AR, consulente — Dati e intelligenza artificiale, Accenture
- Almore Cato, responsabile per dati e intelligenza artificiale, Accenture
- Soonam Kurian, architetto principale delle soluzioni, AWS

Revisione

- Sally Lin, Senior Manager di Data Science, Data & AI, Accenture
- Terry Huang, responsabile della scienza dei dati — Dati e intelligenza artificiale, Accenture
- William Lorenz, Architetto delle soluzioni per i partner, AWS

Scrittura tecnica

- Lilly AbouHarb, scrittrice tecnica senior, AWS

Cronologia dei documenti

La tabella seguente descrive le modifiche significative apportate a questa guida. Per ricevere notifiche sugli aggiornamenti futuri, puoi abbonarti a un feed RSS.

Modifica	Descrizione	Data
Pubblicazione iniziale	—	14 marzo 2025

AWS Glossario delle linee guida prescrittive

I seguenti sono termini di uso comune nelle strategie, nelle guide e nei modelli forniti da AWS Prescriptive Guidance. Per suggerire voci, utilizza il link Fornisci feedback alla fine del glossario.

Numeri

7 R

Sette strategie di migrazione comuni per trasferire le applicazioni sul cloud. Queste strategie si basano sulle 5 R identificate da Gartner nel 2011 e sono le seguenti:

- Rifattorizzare/riprogettare: trasferisci un'applicazione e modifica la sua architettura sfruttando appieno le funzionalità native del cloud per migliorare l'agilità, le prestazioni e la scalabilità. Ciò comporta in genere la portabilità del sistema operativo e del database. Esempio: migra il tuo database Oracle locale all'edizione compatibile con Amazon Aurora PostgreSQL.
- Ridefinire la piattaforma (lift and reshape): trasferisci un'applicazione nel cloud e introduci un certo livello di ottimizzazione per sfruttare le funzionalità del cloud. Esempio: migra il tuo database Oracle locale ad Amazon Relational Database Service (Amazon RDS) per Oracle in. Cloud AWS
- Riacquistare (drop and shop): passa a un prodotto diverso, in genere effettuando la transizione da una licenza tradizionale a un modello SaaS. Esempio: migra il tuo sistema di gestione delle relazioni con i clienti (CRM) su Salesforce.com.
- Eseguire il rehosting (lift and shift): trasferisci un'applicazione sul cloud senza apportare modifiche per sfruttare le funzionalità del cloud. Esempio: migra il database Oracle locale su Oracle su un'istanza in. EC2 Cloud AWS
- Trasferire (eseguire il rehosting a livello hypervisor): trasferisci l'infrastruttura sul cloud senza acquistare nuovo hardware, riscrivere le applicazioni o modificare le operazioni esistenti. Si esegue la migrazione dei server da una piattaforma locale a un servizio cloud per la stessa piattaforma. Esempio: migra un'applicazione suMicrosoft Hyper-V. AWS
- Riesaminare (mantenere): mantieni le applicazioni nell'ambiente di origine. Queste potrebbero includere applicazioni che richiedono una rifattorizzazione significativa che desideri rimandare a un momento successivo e applicazioni legacy che desideri mantenere, perché non vi è alcuna giustificazione aziendale per effettuarne la migrazione.
- Ritirare: disattiva o rimuovi le applicazioni che non sono più necessarie nell'ambiente di origine.

A

ABAC

Vedi controllo degli accessi basato sugli attributi.

servizi astratti

Vedi servizi gestiti.

ACIDO

Vedi atomicità, consistenza, isolamento, durata.

migrazione attiva-attiva

Un metodo di migrazione del database in cui i database di origine e di destinazione vengono mantenuti sincronizzati (utilizzando uno strumento di replica bidirezionale o operazioni di doppia scrittura) ed entrambi i database gestiscono le transazioni provenienti dalle applicazioni di connessione durante la migrazione. Questo metodo supporta la migrazione in piccoli batch controllati anziché richiedere una conversione una tantum. È più flessibile ma richiede più lavoro rispetto alla migrazione attiva-passiva.

migrazione attiva-passiva

Un metodo di migrazione di database in cui i database di origine e di destinazione vengono mantenuti sincronizzati, ma solo il database di origine gestisce le transazioni provenienti dalle applicazioni di connessione mentre i dati vengono replicati nel database di destinazione. Il database di destinazione non accetta alcuna transazione durante la migrazione.

funzione aggregata

Una funzione SQL che opera su un gruppo di righe e calcola un singolo valore restituito per il gruppo. Esempi di funzioni aggregate includono `SUM` e. `MAX`

Intelligenza artificiale

Vedi intelligenza artificiale.

AIOps

Guarda le operazioni di intelligenza artificiale.

anonimizzazione

Il processo di eliminazione permanente delle informazioni personali in un set di dati. L'anonimizzazione può aiutare a proteggere la privacy personale. I dati anonimi non sono più considerati dati personali.

anti-modello

Una soluzione utilizzata di frequente per un problema ricorrente in cui la soluzione è controproducente, inefficace o meno efficace di un'alternativa.

controllo delle applicazioni

Un approccio alla sicurezza che consente l'uso solo di applicazioni approvate per proteggere un sistema dal malware.

portfolio di applicazioni

Una raccolta di informazioni dettagliate su ogni applicazione utilizzata da un'organizzazione, compresi i costi di creazione e manutenzione dell'applicazione e il relativo valore aziendale. Queste informazioni sono fondamentali per il processo di scoperta e analisi del portfolio e aiutano a identificare e ad assegnare la priorità alle applicazioni da migrare, modernizzare e ottimizzare.

intelligenza artificiale (IA)

Il campo dell'informatica dedicato all'uso delle tecnologie informatiche per svolgere funzioni cognitive tipicamente associate agli esseri umani, come l'apprendimento, la risoluzione di problemi e il riconoscimento di schemi. Per ulteriori informazioni, consulta la sezione Che cos'è l'intelligenza artificiale?

operazioni di intelligenza artificiale (AIOps)

Il processo di utilizzo delle tecniche di machine learning per risolvere problemi operativi, ridurre gli incidenti operativi e l'intervento umano e aumentare la qualità del servizio. Per ulteriori informazioni su come AIOps viene utilizzato nella strategia di AWS migrazione, consulta la guida all'integrazione delle operazioni.

crittografia asimmetrica

Un algoritmo di crittografia che utilizza una coppia di chiavi, una chiave pubblica per la crittografia e una chiave privata per la decrittografia. Puoi condividere la chiave pubblica perché non viene utilizzata per la decrittografia, ma l'accesso alla chiave privata deve essere altamente limitato.

atomicità, consistenza, isolamento, durabilità (ACID)

Un insieme di proprietà del software che garantiscono la validità dei dati e l'affidabilità operativa di un database, anche in caso di errori, interruzioni di corrente o altri problemi.

Controllo degli accessi basato su attributi (ABAC)

La pratica di creare autorizzazioni dettagliate basate su attributi utente, come reparto, ruolo professionale e nome del team. Per ulteriori informazioni, consulta ABAC AWS nella documentazione AWS Identity and Access Management (IAM).

fonte di dati autorevole

Una posizione in cui è archiviata la versione principale dei dati, considerata la fonte di informazioni più affidabile. È possibile copiare i dati dalla fonte di dati autorevole in altre posizioni allo scopo di elaborarli o modificarli, ad esempio anonimizzandoli, oscurandoli o pseudonimizzandoli.

Zona di disponibilità

Una posizione distinta all'interno di un edificio Regione AWS che è isolata dai guasti in altre zone di disponibilità e offre una connettività di rete economica e a bassa latenza verso altre zone di disponibilità nella stessa regione.

AWS Cloud Adoption Framework (CAF)AWS

Un framework di linee guida e best practice AWS per aiutare le organizzazioni a sviluppare un piano efficiente ed efficace per passare con successo al cloud. AWS CAF organizza le linee guida in sei aree di interesse chiamate prospettive: business, persone, governance, piattaforma, sicurezza e operazioni. Le prospettive relative ad azienda, persone e governance si concentrano sulle competenze e sui processi aziendali; le prospettive relative alla piattaforma, alla sicurezza e alle operazioni si concentrano sulle competenze e sui processi tecnici. Ad esempio, la prospettiva relativa alle persone si rivolge alle parti interessate che gestiscono le risorse umane (HR), le funzioni del personale e la gestione del personale. In questa prospettiva, AWS CAF fornisce linee guida per lo sviluppo delle persone, la formazione e le comunicazioni per aiutare a preparare l'organizzazione all'adozione del cloud di successo. Per ulteriori informazioni, consulta il sito web di AWS CAF e il white paper AWS CAF.

AWS Workload Qualification Framework (WQF)AWS

Uno strumento che valuta i carichi di lavoro di migrazione dei database, consiglia strategie di migrazione e fornisce stime del lavoro. AWS WQF è incluso in (). AWS Schema Conversion Tool AWS SCT Analizza gli schemi di database e gli oggetti di codice, il codice dell'applicazione, le dipendenze e le caratteristiche delle prestazioni e fornisce report di valutazione.

B

bot difettoso

Un bot che ha lo scopo di interrompere o causare danni a individui o organizzazioni.

BCP

Vedi la pianificazione della continuità operativa.

grafico comportamentale

Una vista unificata, interattiva dei comportamenti delle risorse e delle interazioni nel tempo. Puoi utilizzare un grafico comportamentale con Amazon Detective per esaminare tentativi di accesso non riusciti, chiamate API sospette e azioni simili. Per ulteriori informazioni, consulta Dati in un grafico comportamentale nella documentazione di Detective.

sistema big-endian

Un sistema che memorizza per primo il byte più importante. Vedi anche endianness.

Classificazione binaria

Un processo che prevede un risultato binario (una delle due classi possibili). Ad esempio, il modello di machine learning potrebbe dover prevedere problemi come "Questa e-mail è spam o non è spam?" o "Questo prodotto è un libro o un'auto?"

filtro Bloom

Una struttura di dati probabilistica ed efficiente in termini di memoria che viene utilizzata per verificare se un elemento fa parte di un set.

distribuzioni blu/verdi

Una strategia di implementazione in cui si creano due ambienti separati ma identici. La versione corrente dell'applicazione viene eseguita in un ambiente (blu) e la nuova versione dell'applicazione nell'altro ambiente (verde). Questa strategia consente di ripristinare rapidamente il sistema con un impatto minimo.

bot

Un'applicazione software che esegue attività automatizzate su Internet e simula l'attività o l'interazione umana. Alcuni bot sono utili o utili, come i web crawler che indicizzano le informazioni su Internet. Alcuni altri bot, noti come bot dannosi, hanno lo scopo di disturbare o causare danni a individui o organizzazioni.

botnet

Reti di bot infettate da malware e controllate da un'unica parte, nota come bot herder o bot operator. Le botnet sono il meccanismo più noto per scalare i bot e il loro impatto.

ramo

Un'area contenuta di un repository di codice. Il primo ramo creato in un repository è il ramo principale. È possibile creare un nuovo ramo a partire da un ramo esistente e quindi sviluppare funzionalità o correggere bug al suo interno. Un ramo creato per sviluppare una funzionalità viene comunemente detto ramo di funzionalità. Quando la funzionalità è pronta per il rilascio, il ramo di funzionalità viene ricongiunto al ramo principale. Per ulteriori informazioni, consulta Informazioni sulle filiali (documentazione). GitHub

accesso break-glass

In circostanze eccezionali e tramite una procedura approvata, un mezzo rapido per consentire a un utente di accedere a un sito a Account AWS cui in genere non dispone delle autorizzazioni necessarie. Per ulteriori informazioni, vedere l'indicatore Implementate break-glass procedures nella guida Well-Architected AWS .

strategia brownfield

L'infrastruttura esistente nell'ambiente. Quando si adotta una strategia brownfield per un'architettura di sistema, si progetta l'architettura in base ai vincoli dei sistemi e dell'infrastruttura attuali. Per l'espansione dell'infrastruttura esistente, è possibile combinare strategie brownfield e greenfield.

cache del buffer

L'area di memoria in cui sono archiviati i dati a cui si accede con maggiore frequenza.

capacità di business

Azioni intraprese da un'azienda per generare valore (ad esempio vendite, assistenza clienti o marketing). Le architetture dei microservizi e le decisioni di sviluppo possono essere guidate dalle capacità aziendali. Per ulteriori informazioni, consulta la sezione Organizzazione in base alle funzionalità aziendali del whitepaper Esecuzione di microservizi containerizzati su AWS.

pianificazione della continuità operativa (BCP)

Un piano che affronta il potenziale impatto di un evento che comporta l'interruzione dell'attività, come una migrazione su larga scala, sulle operazioni e consente a un'azienda di riprendere rapidamente le operazioni.

C

CAF

Vedi AWS Cloud Adoption Framework.

implementazione canaria

Il rilascio lento e incrementale di una versione agli utenti finali. Quando sei sicuro, distribuisci la nuova versione e sostituisci la versione corrente nella sua interezza.

CCoE

Vedi Cloud Center of Excellence.

CDC

Vedi Change Data Capture.

Change Data Capture (CDC)

Il processo di tracciamento delle modifiche a un'origine dati, ad esempio una tabella di database, e di registrazione dei metadati relativi alla modifica. È possibile utilizzare CDC per vari scopi, ad esempio il controllo o la replica delle modifiche in un sistema di destinazione per mantenere la sincronizzazione.

ingegneria del caos

Introduzione intenzionale di guasti o eventi dirompenti per testare la resilienza di un sistema. Puoi usare AWS Fault Injection Service (AWS FIS) per eseguire esperimenti che stressano i tuoi AWS carichi di lavoro e valutarne la risposta.

CI/CD

Vedi integrazione continua e distribuzione continua.

classificazione

Un processo di categorizzazione che aiuta a generare previsioni. I modelli di ML per problemi di classificazione prevedono un valore discreto. I valori discreti sono sempre distinti l'uno dall'altro. Ad esempio, un modello potrebbe dover valutare se in un'immagine è presente o meno un'auto.

crittografia lato client

Crittografia dei dati a livello locale, prima che il destinatario li Servizio AWS riceva.

Centro di eccellenza cloud (CCoE)

Un team multidisciplinare che guida le iniziative di adozione del cloud in tutta l'organizzazione, tra cui lo sviluppo di best practice per il cloud, la mobilitazione delle risorse, la definizione delle tempistiche di migrazione e la guida dell'organizzazione attraverso trasformazioni su larga scala. Per ulteriori informazioni, consulta gli CCoE post sull' Cloud AWS Enterprise Strategy Blog.

cloud computing

La tecnologia cloud generalmente utilizzata per l'archiviazione remota di dati e la gestione dei dispositivi IoT. Il cloud computing è generalmente collegato alla tecnologia di edge computing.

modello operativo cloud

In un'organizzazione IT, il modello operativo utilizzato per creare, maturare e ottimizzare uno o più ambienti cloud. Per ulteriori informazioni, consulta Building your Cloud Operating Model.

fasi di adozione del cloud

Le quattro fasi che le organizzazioni in genere attraversano quando migrano verso Cloud AWS:

- Progetto: esecuzione di alcuni progetti relativi al cloud per scopi di dimostrazione e apprendimento
- Fondamento: effettuare investimenti fondamentali per scalare l'adozione del cloud (ad esempio, creazione di una landing zone, definizione di una CCo E, definizione di un modello operativo)
- Migrazione: migrazione di singole applicazioni
- Reinvenzione: ottimizzazione di prodotti e servizi e innovazione nel cloud

Queste fasi sono state definite da Stephen Orban nel post sul blog The Journey Toward Cloud-First & the Stages of Adoption on the Enterprise Strategy. Cloud AWS Per informazioni su come si relazionano alla strategia di AWS migrazione, consulta la guida alla preparazione alla migrazione.

CMDB

Vedi database di gestione della configurazione.

repository di codice

Una posizione in cui il codice di origine e altri asset, come documentazione, esempi e script, vengono archiviati e aggiornati attraverso processi di controllo delle versioni. Gli archivi cloud più comuni includono GitHub oBitbucket Cloud. Ogni versione del codice è denominata ramo. In una

struttura a microservizi, ogni repository è dedicato a una singola funzionalità. Una singola pipeline CI/CD può utilizzare più repository.

cache fredda

Una cache del buffer vuota, non ben popolata o contenente dati obsoleti o irrilevanti. Ciò influisce sulle prestazioni perché l'istanza di database deve leggere dalla memoria o dal disco principale, il che richiede più tempo rispetto alla lettura dalla cache del buffer.

dati freddi

Dati a cui si accede raramente e che in genere sono storici. Quando si eseguono interrogazioni di questo tipo di dati, le interrogazioni lente sono in genere accettabili. Lo spostamento di questi dati su livelli o classi di storage meno costosi e con prestazioni inferiori può ridurre i costi.

visione artificiale (CV)

Un campo dell'intelligenza artificiale che utilizza l'apprendimento automatico per analizzare ed estrarre informazioni da formati visivi come immagini e video digitali. Ad esempio, Amazon SageMaker AI fornisce algoritmi di elaborazione delle immagini per CV.

deriva della configurazione

Per un carico di lavoro, una modifica della configurazione rispetto allo stato previsto. Potrebbe causare la non conformità del carico di lavoro e in genere è graduale e involontaria.

database di gestione della configurazione (CMDB)

Un repository che archivia e gestisce le informazioni su un database e il relativo ambiente IT, inclusi i componenti hardware e software e le relative configurazioni. In genere si utilizzano i dati di un CMDB nella fase di individuazione e analisi del portafoglio della migrazione.

Pacchetto di conformità

Una raccolta di AWS Config regole e azioni correttive che puoi assemblare per personalizzare i controlli di conformità e sicurezza. È possibile distribuire un pacchetto di conformità come singola entità in una regione Account AWS and o all'interno di un'organizzazione utilizzando un modello YAML. Per ulteriori informazioni, consulta i Conformance Pack nella documentazione. AWS Config

integrazione e distribuzione continua (continuous integration and continuous delivery, CI/CD)

Il processo di automazione delle fasi di origine, compilazione, test, gestione temporanea e produzione del processo di rilascio del software. CI/CD is commonly described as a pipeline. CI/

CDpuò aiutarvi ad automatizzare i processi, migliorare la produttività, migliorare la qualità del codice e velocizzare le consegne. Per ulteriori informazioni, consulta Vantaggi della distribuzione continua. CD può anche significare continuous deployment (implementazione continua). Per ulteriori informazioni, consulta Distribuzione continua e implementazione continua a confronto.

CV

Vedi visione artificiale.

D

dati a riposo

Dati stazionari nella rete, ad esempio i dati archiviati.

classificazione dei dati

Un processo per identificare e classificare i dati nella rete in base alla loro criticità e sensibilità. È un componente fondamentale di qualsiasi strategia di gestione dei rischi di sicurezza informatica perché consente di determinare i controlli di protezione e conservazione appropriati per i dati. La classificazione dei dati è un componente del pilastro della sicurezza nel AWS Well-Architected Framework. Per ulteriori informazioni, consulta Classificazione dei dati.

deriva dei dati

Una variazione significativa tra i dati di produzione e i dati utilizzati per addestrare un modello di machine learning o una modifica significativa dei dati di input nel tempo. La deriva dei dati può ridurre la qualità, l'accuratezza e l'equità complessive nelle previsioni dei modelli ML.

dati in transito

Dati che si spostano attivamente attraverso la rete, ad esempio tra le risorse di rete.

rete di dati

Un framework architettonico che fornisce la proprietà distribuita e decentralizzata dei dati con gestione e governance centralizzate.

riduzione al minimo dei dati

Il principio della raccolta e del trattamento dei soli dati strettamente necessari. Praticare la riduzione al minimo dei dati in the Cloud AWS può ridurre i rischi per la privacy, i costi e l'impronta di carbonio delle analisi.

perimetro dei dati

Una serie di barriere preventive nell' AWS ambiente che aiutano a garantire che solo le identità attendibili accedano alle risorse attendibili delle reti previste. Per ulteriori informazioni, consulta Building a data perimeter on. AWS

pre-elaborazione dei dati

Trasformare i dati grezzi in un formato che possa essere facilmente analizzato dal modello di ML. La pre-elaborazione dei dati può comportare la rimozione di determinate colonne o righe e l'eliminazione di valori mancanti, incoerenti o duplicati.

provenienza dei dati

Il processo di tracciamento dell'origine e della cronologia dei dati durante il loro ciclo di vita, ad esempio il modo in cui i dati sono stati generati, trasmessi e archiviati.

soggetto dei dati

Un individuo i cui dati vengono raccolti ed elaborati.

data warehouse

Un sistema di gestione dei dati che supporta la business intelligence, come l'analisi. I data warehouse contengono in genere grandi quantità di dati storici e vengono generalmente utilizzati per interrogazioni e analisi.

linguaggio di definizione del database (DDL)

Istruzioni o comandi per creare o modificare la struttura di tabelle e oggetti in un database.

linguaggio di manipolazione del database (DML)

Istruzioni o comandi per modificare (inserire, aggiornare ed eliminare) informazioni in un database.

DDL

Vedi linguaggio di definizione del database.

deep ensemble

Combinare più modelli di deep learning per la previsione. È possibile utilizzare i deep ensemble per ottenere una previsione più accurata o per stimare l'incertezza nelle previsioni.

deep learning

Un sottocampo del ML che utilizza più livelli di reti neurali artificiali per identificare la mappatura tra i dati di input e le variabili target di interesse.

defense-in-depth

Un approccio alla sicurezza delle informazioni in cui una serie di meccanismi e controlli di sicurezza sono accuratamente stratificati su una rete di computer per proteggere la riservatezza, l'integrità e la disponibilità della rete e dei dati al suo interno. Quando si adotta questa strategia AWS, si aggiungono più controlli a diversi livelli della AWS Organizations struttura per proteggere le risorse. Ad esempio, un defense-in-depth approccio potrebbe combinare l'autenticazione a più fattori, la segmentazione della rete e la crittografia.

amministratore delegato

In AWS Organizations, un servizio compatibile può registrare un account AWS membro per amministrare gli account dell'organizzazione e gestire le autorizzazioni per quel servizio. Questo account è denominato amministratore delegato per quel servizio specifico. Per ulteriori informazioni e un elenco di servizi compatibili, consulta Servizi che funzionano con AWS Organizations nella documentazione di AWS Organizations .

implementazione

Il processo di creazione di un'applicazione, di nuove funzionalità o di correzioni di codice disponibili nell'ambiente di destinazione. L'implementazione prevede l'applicazione di modifiche in una base di codice, seguita dalla creazione e dall'esecuzione di tale base di codice negli ambienti applicativi.

Ambiente di sviluppo

Vedi ambiente.

controllo di rilevamento

Un controllo di sicurezza progettato per rilevare, registrare e avvisare dopo che si è verificato un evento. Questi controlli rappresentano una seconda linea di difesa e avvisano l'utente in caso di eventi di sicurezza che aggirano i controlli preventivi in vigore. Per ulteriori informazioni, consulta Controlli di rilevamento in Implementazione dei controlli di sicurezza in AWS.

mappatura del flusso di valore dello sviluppo (DVSM)

Un processo utilizzato per identificare e dare priorità ai vincoli che influiscono negativamente sulla velocità e sulla qualità nel ciclo di vita dello sviluppo del software. DVSM estende il processo di

mappatura del flusso di valore originariamente progettato per pratiche di produzione snella. Si concentra sulle fasi e sui team necessari per creare e trasferire valore attraverso il processo di sviluppo del software.

gemello digitale

Una rappresentazione virtuale di un sistema reale, ad esempio un edificio, una fabbrica, un'attrezzatura industriale o una linea di produzione. I gemelli digitali supportano la manutenzione predittiva, il monitoraggio remoto e l'ottimizzazione della produzione.

tabella delle dimensioni

In uno schema a stella, una tabella più piccola che contiene gli attributi dei dati quantitativi in una tabella dei fatti. Gli attributi della tabella delle dimensioni sono in genere campi di testo o numeri discreti che si comportano come testo. Questi attributi vengono comunemente utilizzati per il vincolo delle query, il filtraggio e l'etichettatura dei set di risultati.

disastro

Un evento che impedisce a un carico di lavoro o a un sistema di raggiungere gli obiettivi aziendali nella sua sede principale di implementazione. Questi eventi possono essere disastri naturali, guasti tecnici o il risultato di azioni umane, come errori di configurazione involontari o attacchi di malware.

disaster recovery (DR)

La strategia e il processo utilizzati per ridurre al minimo i tempi di inattività e la perdita di dati causati da un disastro. Per ulteriori informazioni, consulta Disaster Recovery of Workloads su AWS: Recovery in the Cloud in the AWS Well-Architected Framework.

DML

Vedi linguaggio di manipolazione del database.

progettazione basata sul dominio

Un approccio allo sviluppo di un sistema software complesso collegandone i componenti a domini in evoluzione, o obiettivi aziendali principali, perseguiti da ciascun componente. Questo concetto è stato introdotto da Eric Evans nel suo libro,Domain-Driven Design: Tackling Complexity in the Heart of Software (Boston: Addison-Wesley Professional, 2003). Per informazioni su come utilizzare la progettazione basata sul dominio con il modello del fico strangolatore (Strangler Fig), consulta la sezione Modernizzazione incrementale dei servizi Web Microsoft ASP.NET (ASMX) legacy utilizzando container e il Gateway Amazon API.

DOTT.

Vedi disaster recovery.

rilevamento della deriva

Tracciamento delle deviazioni da una configurazione di base. Ad esempio, puoi utilizzarlo AWS CloudFormation per rilevare la deriva nelle risorse di sistema oppure puoi usarlo AWS Control Tower per rilevare cambiamenti nella tua landing zone che potrebbero influire sulla conformità ai requisiti di governance.

DVSM

Vedi la mappatura del flusso di valore dello sviluppo.

E

EDA

Vedi analisi esplorativa dei dati.

MODIFICA

Vedi scambio elettronico di dati.

edge computing

La tecnologia che aumenta la potenza di calcolo per i dispositivi intelligenti all'edge di una rete IoT. Rispetto al cloud computing, l'edge computing può ridurre la latenza di comunicazione e migliorare i tempi di risposta.

scambio elettronico di dati (EDI)

Lo scambio automatizzato di documenti aziendali tra organizzazioni. Per ulteriori informazioni, vedere Cos'è lo scambio elettronico di dati.

crittografia

Un processo di elaborazione che trasforma i dati in chiaro, leggibili dall'uomo, in testo cifrato.

chiave crittografica

Una stringa crittografica di bit randomizzati generata da un algoritmo di crittografia. Le chiavi possono variare di lunghezza e ogni chiave è progettata per essere imprevedibile e univoca.

endianità

L'ordine in cui i byte vengono archiviati nella memoria del computer. I sistemi big-endian memorizzano per primo il byte più importante. I sistemi little-endian memorizzano per primo il byte meno importante.

endpoint

Vedi service endpoint.

servizio endpoint

Un servizio che puoi ospitare in un cloud privato virtuale (VPC) da condividere con altri utenti. Puoi creare un servizio endpoint con AWS PrivateLink e concedere autorizzazioni ad altri Account AWS o a AWS Identity and Access Management (IAM) principali. Questi account o principali possono connettersi al servizio endpoint in privato creando endpoint VPC di interfaccia. Per ulteriori informazioni, consulta Creazione di un servizio endpoint nella documentazione di Amazon Virtual Private Cloud (Amazon VPC).

pianificazione delle risorse aziendali (ERP)

Un sistema che automatizza e gestisce i processi aziendali chiave (come contabilità, MES e gestione dei progetti) per un'azienda.

crittografia envelope

Il processo di crittografia di una chiave di crittografia con un'altra chiave di crittografia. Per ulteriori informazioni, vedete Envelope encryption nella documentazione AWS Key Management Service (AWS KMS).

ambiente

Un'istanza di un'applicazione in esecuzione. Di seguito sono riportati i tipi di ambiente più comuni nel cloud computing:

- ambiente di sviluppo: un'istanza di un'applicazione in esecuzione disponibile solo per il team principale responsabile della manutenzione dell'applicazione. Gli ambienti di sviluppo vengono utilizzati per testare le modifiche prima di promuoverle negli ambienti superiori. Questo tipo di ambiente viene talvolta definito ambiente di test.
- ambienti inferiori: tutti gli ambienti di sviluppo di un'applicazione, ad esempio quelli utilizzati per le build e i test iniziali.

- ambiente di produzione: un'istanza di un'applicazione in esecuzione a cui gli utenti finali possono accedere. In una pipeline CI/CD, l'ambiente di produzione è l'ultimo ambiente di implementazione.
- ambienti superiori: tutti gli ambienti a cui possono accedere utenti diversi dal team di sviluppo principale. Si può trattare di un ambiente di produzione, ambienti di preproduzione e ambienti per i test di accettazione da parte degli utenti.

epica

Nelle metodologie agili, categorie funzionali che aiutano a organizzare e dare priorità al lavoro. Le epiche forniscono una descrizione di alto livello dei requisiti e delle attività di implementazione. Ad esempio, le epopee della sicurezza AWS CAF includono la gestione delle identità e degli accessi, i controlli investigativi, la sicurezza dell'infrastruttura, la protezione dei dati e la risposta agli incidenti. Per ulteriori informazioni sulle epiche, consulta la strategia di migrazione AWS , consulta la guida all'implementazione del programma.

ERP

Vedi pianificazione delle risorse aziendali.

analisi esplorativa dei dati (EDA)

Il processo di analisi di un set di dati per comprenderne le caratteristiche principali. Si raccolgono o si aggregano dati e quindi si eseguono indagini iniziali per trovare modelli, rilevare anomalie e verificare ipotesi. L'EDA viene eseguita calcolando statistiche di riepilogo e creando visualizzazioni di dati.

F

tabella dei fatti

Il tavolo centrale con schema a stella. Memorizza dati quantitativi sulle operazioni aziendali. In genere, una tabella dei fatti contiene due tipi di colonne: quelle che contengono misure e quelle che contengono una chiave esterna per una tabella di dimensioni.

fallire velocemente

Una filosofia che utilizza test frequenti e incrementali per ridurre il ciclo di vita dello sviluppo. È una parte fondamentale di un approccio agile.

limite di isolamento dei guasti

Nel Cloud AWS, un limite come una zona di disponibilità Regione AWS, un piano di controllo o un piano dati che limita l'effetto di un errore e aiuta a migliorare la resilienza dei carichi di lavoro. Per ulteriori informazioni, consulta AWS Fault Isolation Boundaries.

ramo di funzionalità

Vedi filiale.

caratteristiche

I dati di input che usi per fare una previsione. Ad esempio, in un contesto di produzione, le caratteristiche potrebbero essere immagini acquisite periodicamente dalla linea di produzione.

importanza delle caratteristiche

Quanto è importante una caratteristica per le previsioni di un modello. Di solito viene espresso come punteggio numerico che può essere calcolato con varie tecniche, come Shapley Additive Explanations (SHAP) e gradienti integrati. Per ulteriori informazioni, consulta Interpretabilità del modello di machine learning con AWS.

trasformazione delle funzionalità

Per ottimizzare i dati per il processo di machine learning, incluso l'arricchimento dei dati con fonti aggiuntive, il dimensionamento dei valori o l'estrazione di più set di informazioni da un singolo campo di dati. Ciò consente al modello di ML di trarre vantaggio dai dati. Ad esempio, se suddividi la data "2021-05-27 00:15:37" in "2021", "maggio", "giovedì" e "15", puoi aiutare l'algoritmo di apprendimento ad apprendere modelli sfumati associati a diversi componenti dei dati.

prompt con pochi scatti

Fornire a un LLM un numero limitato di esempi che dimostrino l'attività e il risultato desiderato prima di chiedergli di eseguire un'attività simile. Questa tecnica è un'applicazione dell'apprendimento contestuale, in cui i modelli imparano da esempi (immagini) incorporati nei prompt. I prompt con pochi passaggi possono essere efficaci per attività che richiedono una formattazione, un ragionamento o una conoscenza del dominio specifici. Vedi anche zero-shot prompting.

FGAC

Vedi il controllo granulare degli accessi.

controllo granulare degli accessi (FGAC)

L'uso di più condizioni per consentire o rifiutare una richiesta di accesso.

migrazione flash-cut

Un metodo di migrazione del database che utilizza la replica continua dei dati tramite l'acquisizione dei dati delle modifiche per migrare i dati nel più breve tempo possibile, anziché utilizzare un approccio graduale. L'obiettivo è ridurre al minimo i tempi di inattività.

FM

Vedi il modello di base.

modello di fondazione (FM)

Una grande rete neurale di deep learning che si è addestrata su enormi set di dati generalizzati e non etichettati. FMs sono in grado di svolgere un'ampia varietà di attività generali, come comprendere il linguaggio, generare testo e immagini e conversare in linguaggio naturale. Per ulteriori informazioni, consulta Cosa sono i modelli Foundation.

G

AI generativa

Un sottoinsieme di modelli di intelligenza artificiale che sono stati addestrati su grandi quantità di dati e che possono utilizzare un semplice prompt di testo per creare nuovi contenuti e artefatti, come immagini, video, testo e audio. Per ulteriori informazioni, consulta Cos'è l'IA generativa.

blocco geografico

Vedi restrizioni geografiche.

limitazioni geografiche (blocco geografico)

In Amazon CloudFront, un'opzione per impedire agli utenti di determinati paesi di accedere alle distribuzioni di contenuti. Puoi utilizzare un elenco consentito o un elenco di blocco per specificare i paesi approvati e vietati. Per ulteriori informazioni, consulta Limitare la distribuzione geografica dei contenuti nella CloudFront documentazione.

Flusso di lavoro di GitFlow

Un approccio in cui gli ambienti inferiori e superiori utilizzano rami diversi in un repository di codice di origine. Il flusso di lavoro Gitflow è considerato obsoleto e il flusso di lavoro basato su trunk è l'approccio moderno e preferito.

immagine dorata

Un'istantanea di un sistema o di un software che viene utilizzata come modello per distribuire nuove istanze di quel sistema o software. Ad esempio, nella produzione, un'immagine dorata può essere utilizzata per fornire software su più dispositivi e contribuire a migliorare la velocità, la scalabilità e la produttività nelle operazioni di produzione dei dispositivi.

strategia greenfield

L'assenza di infrastrutture esistenti in un nuovo ambiente. Quando si adotta una strategia greenfield per un'architettura di sistema, è possibile selezionare tutte le nuove tecnologie senza il vincolo della compatibilità con l'infrastruttura esistente, nota anche come brownfield. Per l'espansione dell'infrastruttura esistente, è possibile combinare strategie brownfield e greenfield.

guardrail

Una regola di alto livello che aiuta a governare le risorse, le politiche e la conformità tra le unità organizzative (). OUs I guardrail preventivi applicano le policy per garantire l'allineamento agli standard di conformità. Vengono implementati utilizzando le policy di controllo dei servizi e i limiti delle autorizzazioni IAM. I guardrail di rilevamento rilevano le violazioni delle policy e i problemi di conformità e generano avvisi per porvi rimedio. Sono implementati utilizzando Amazon AWS Config AWS Security Hub GuardDuty AWS Trusted Advisor, Amazon Inspector e controlli personalizzati AWS Lambda .

H

AH

Vedi disponibilità elevata.

migrazione di database eterogenea

Migrazione del database di origine in un database di destinazione che utilizza un motore di database diverso (ad esempio, da Oracle ad Amazon Aurora). La migrazione eterogenea fa in

genere parte di uno sforzo di riprogettazione e la conversione dello schema può essere un'attività complessa. AWS offre AWS SCT che aiuta con le conversioni dello schema.

alta disponibilità (HA)

La capacità di un carico di lavoro di funzionare in modo continuo, senza intervento, in caso di sfide o disastri. I sistemi HA sono progettati per il failover automatico, fornire costantemente prestazioni di alta qualità e gestire carichi e guasti diversi con un impatto minimo sulle prestazioni.

modernizzazione storica

Un approccio utilizzato per modernizzare e aggiornare i sistemi di tecnologia operativa (OT) per soddisfare meglio le esigenze dell'industria manifatturiera. Uno storico è un tipo di database utilizzato per raccogliere e archiviare dati da varie fonti in una fabbrica.

dati di esclusione

Una parte di dati storici etichettati che viene trattenuta da un set di dati utilizzata per addestrare un modello di apprendimento automatico. È possibile utilizzare i dati di holdout per valutare le prestazioni del modello confrontando le previsioni del modello con i dati di holdout.

migrazione di database omogenea

Migrazione del database di origine in un database di destinazione che condivide lo stesso motore di database (ad esempio, da Microsoft SQL Server ad Amazon RDS per SQL Server). La migrazione omogenea fa in genere parte di un'operazione di rehosting o ridefinizione della piattaforma. Per migrare lo schema è possibile utilizzare le utilità native del database.

dati caldi

Dati a cui si accede frequentemente, come dati in tempo reale o dati di traduzione recenti. Questi dati richiedono in genere un livello o una classe di storage ad alte prestazioni per fornire risposte rapide alle query.

hotfix

Una soluzione urgente per un problema critico in un ambiente di produzione. A causa della sua urgenza, un hotfix viene in genere creato al di fuori del tipico DevOps flusso di lavoro di rilascio.

periodo di hypercare

Subito dopo la conversione, il periodo di tempo in cui un team di migrazione gestisce e monitora le applicazioni migrate nel cloud per risolvere eventuali problemi. In genere, questo periodo dura

da 1 a 4 giorni. Al termine del periodo di hypercare, il team addetto alla migrazione in genere trasferisce la responsabilità delle applicazioni al team addetto alle operazioni cloud.

I

IaC

Considera l'infrastruttura come codice.

Policy basata su identità

Una policy associata a uno o più principi IAM che definisce le relative autorizzazioni all'interno dell' Cloud AWS ambiente.

applicazione inattiva

Un'applicazione che prevede un uso di CPU e memoria medio compreso tra il 5% e il 20% in un periodo di 90 giorni. In un progetto di migrazione, è normale ritirare queste applicazioni o mantenerle on-premise.

IIoT

Vedi Industrial Internet of Things.

infrastruttura immutabile

Un modello che implementa una nuova infrastruttura per i carichi di lavoro di produzione anziché aggiornare, applicare patch o modificare l'infrastruttura esistente. Le infrastrutture immutabili sono intrinsecamente più coerenti, affidabili e prevedibili delle infrastrutture mutabili. Per ulteriori informazioni, consulta la best practice Deploy using immutable infrastructure in Well-Architected AWS Framework.

VPC in ingresso (ingress)

In un'architettura AWS multi-account, un VPC che accetta, ispeziona e indirizza le connessioni di rete dall'esterno di un'applicazione. La AWS Security Reference Architecture consiglia di configurare l'account di rete con funzionalità in entrata, in uscita e di ispezione VPCs per proteggere l'interfaccia bidirezionale tra l'applicazione e la rete Internet in generale.

migrazione incrementale

Una strategia di conversione in cui si esegue la migrazione dell'applicazione in piccole parti anziché eseguire una conversione singola e completa. Ad esempio, inizialmente potresti spostare

solo alcuni microservizi o utenti nel nuovo sistema. Dopo aver verificato che tutto funzioni correttamente, puoi spostare in modo incrementale microservizi o utenti aggiuntivi fino alla disattivazione del sistema legacy. Questa strategia riduce i rischi associati alle migrazioni di grandi dimensioni.

Industria 4.0

Un termine introdotto da Klaus Schwab nel 2016 per riferirsi alla modernizzazione dei processi di produzione attraverso progressi in termini di connettività, dati in tempo reale, automazione, analisi e AI/ML.

infrastruttura

Tutte le risorse e gli asset contenuti nell'ambiente di un'applicazione.

infrastruttura come codice (IaC)

Il processo di provisioning e gestione dell'infrastruttura di un'applicazione tramite un insieme di file di configurazione. Il processo IaC è progettato per aiutarti a centralizzare la gestione dell'infrastruttura, a standardizzare le risorse e a dimensionare rapidamente, in modo che i nuovi ambienti siano ripetibili, affidabili e coerenti.

IIoInternet delle cose industriale (T)

L'uso di sensori e dispositivi connessi a Internet nei settori industriali, come quello manifatturiero, energetico, automobilistico, sanitario, delle scienze della vita e dell'agricoltura. Per ulteriori informazioni, vedere Creazione di una strategia di trasformazione digitale per l'Internet of Things (IIoT) industriale.

VPC di ispezione

In un'architettura AWS multi-account, un VPC centralizzato che gestisce le ispezioni del traffico di rete tra VPCs (nello stesso o in modo diverso Regioni AWS), Internet e le reti locali. La AWS Security Reference Architecture consiglia di configurare l'account di rete con informazioni in entrata, in uscita e di ispezione VPCs per proteggere l'interfaccia bidirezionale tra l'applicazione e Internet in generale.

Internet of Things (IoT)

La rete di oggetti fisici connessi con sensori o processori incorporati che comunicano con altri dispositivi e sistemi tramite Internet o una rete di comunicazione locale. Per ulteriori informazioni, consulta Cos'è l'IoT?

interpretabilità

Una caratteristica di un modello di machine learning che descrive il grado in cui un essere umano è in grado di comprendere in che modo le previsioni del modello dipendono dai suoi input. Per ulteriori informazioni, vedere Interpretabilità del modello di machine learning con. AWS

IoT

Vedi Internet of Things.

libreria di informazioni IT (ITIL)

Una serie di best practice per offrire servizi IT e allinearli ai requisiti aziendali. ITIL fornisce le basi per ITSM.

gestione dei servizi IT (ITSM)

Attività associate alla progettazione, implementazione, gestione e supporto dei servizi IT per un'organizzazione. Per informazioni sull'integrazione delle operazioni cloud con gli strumenti ITSM, consulta la guida all'integrazione delle operazioni.

ITIL

Vedi la libreria di informazioni IT.

ITSM

Vedi Gestione dei servizi IT.

L

controllo degli accessi basato su etichette (LBAC)

Un'implementazione del controllo di accesso obbligatorio (MAC) in cui agli utenti e ai dati stessi viene assegnato esplicitamente un valore di etichetta di sicurezza. L'intersezione tra l'etichetta di sicurezza utente e l'etichetta di sicurezza dei dati determina quali righe e colonne possono essere visualizzate dall'utente.

zona di destinazione

Una landing zone è un AWS ambiente multi-account ben progettato, scalabile e sicuro. Questo è un punto di partenza dal quale le organizzazioni possono avviare e distribuire rapidamente carichi di lavoro e applicazioni con fiducia nel loro ambiente di sicurezza e infrastruttura. Per ulteriori

informazioni sulle zone di destinazione, consulta la sezione Configurazione di un ambiente AWS multi-account sicuro e scalabile.

modello linguistico di grandi dimensioni (LLM)

Un modello di intelligenza artificiale di deep learning preaddestrato su una grande quantità di dati. Un LLM può svolgere più attività, come rispondere a domande, riepilogare documenti, tradurre testo in altre lingue e completare frasi. Per ulteriori informazioni, consulta Cosa sono. LLMs

migrazione su larga scala

Una migrazione di 300 o più server.

BIANCO

Vedi controllo degli accessi basato su etichette.

Privilegio minimo

La best practice di sicurezza per la concessione delle autorizzazioni minime richieste per eseguire un'attività. Per ulteriori informazioni, consulta Applicazione delle autorizzazioni del privilegio minimo nella documentazione di IAM.

eseguire il rehosting (lift and shift)

Vedi 7 R.

sistema little-endian

Un sistema che memorizza per primo il byte meno importante. Vedi anche endianità.

LLM

Vedi modello linguistico di grandi dimensioni.

ambienti inferiori

Vedi ambiente.

M

machine learning (ML)

Un tipo di intelligenza artificiale che utilizza algoritmi e tecniche per il riconoscimento e l'apprendimento di schemi. Il machine learning analizza e apprende dai dati registrati, come i dati

dell'Internet delle cose (IoT), per generare un modello statistico basato su modelli. Per ulteriori informazioni, consulta la sezione Machine learning.

ramo principale

Vedi filiale.

malware

Software progettato per compromettere la sicurezza o la privacy del computer. Il malware potrebbe interrompere i sistemi informatici, divulgare informazioni sensibili o ottenere accessi non autorizzati. Esempi di malware includono virus, worm, ransomware, trojan horse, spyware e keylogger.

servizi gestiti

Servizi AWS per cui AWS gestisce il livello di infrastruttura, il sistema operativo e le piattaforme e si accede agli endpoint per archiviare e recuperare i dati. Amazon Simple Storage Service (Amazon S3) Simple Storage Service (Amazon S3) e Amazon DynamoDB sono esempi di servizi gestiti. Questi sono noti anche come servizi astratti.

sistema di esecuzione della produzione (MES)

Un sistema software per tracciare, monitorare, documentare e controllare i processi di produzione che convertono le materie prime in prodotti finiti in officina.

MAP

Vedi Migration Acceleration Program.

meccanismo

Un processo completo in cui si crea uno strumento, si promuove l'adozione dello strumento e quindi si esaminano i risultati per apportare le modifiche. Un meccanismo è un ciclo che si rafforza e si migliora man mano che funziona. Per ulteriori informazioni, consulta Creazione di meccanismi nel AWS Well-Architected Framework.

account membro

Tutti gli account Account AWS diversi dall'account di gestione che fanno parte di un'organizzazione in. AWS Organizations Un account può essere membro di una sola organizzazione alla volta.

MEH.

Vedi sistema di esecuzione della produzione.

Message Queuing Telemetry Transport (MQTT)

Un protocollo di comunicazione machine-to-machine (M2M) leggero, basato sul modello di pubblicazione/sottoscrizione, per dispositivi IoT con risorse limitate.

microservizio

Un servizio piccolo e indipendente che comunica tramite canali ben definiti ed è in genere di proprietà di piccoli team autonomi. APIs Ad esempio, un sistema assicurativo potrebbe includere microservizi che si riferiscono a funzionalità aziendali, come vendite o marketing, o sottodomini, come acquisti, reclami o analisi. I vantaggi dei microservizi includono agilità, dimensionamento flessibile, facilità di implementazione, codice riutilizzabile e resilienza. Per ulteriori informazioni, consulta Integrazione dei microservizi utilizzando servizi serverless. AWS

architettura di microservizi

Un approccio alla creazione di un'applicazione con componenti indipendenti che eseguono ogni processo applicativo come microservizio. Questi microservizi comunicano attraverso un'interfaccia ben definita utilizzando sistemi leggeri. APIs Ogni microservizio in questa architettura può essere aggiornato, distribuito e dimensionato per soddisfare la richiesta di funzioni specifiche di un'applicazione. Per ulteriori informazioni, vedere Implementazione dei microservizi su. AWS

Programma di accelerazione della migrazione (MAP)

Un AWS programma che fornisce consulenza, supporto, formazione e servizi per aiutare le organizzazioni a costruire una solida base operativa per il passaggio al cloud e per contribuire a compensare il costo iniziale delle migrazioni. MAP include una metodologia di migrazione per eseguire le migrazioni precedenti in modo metodico e un set di strumenti per automatizzare e accelerare gli scenari di migrazione comuni.

migrazione su larga scala

Il processo di trasferimento della maggior parte del portfolio di applicazioni sul cloud avviene a ondate, con più applicazioni trasferite a una velocità maggiore in ogni ondata. Questa fase utilizza le migliori pratiche e le lezioni apprese nelle fasi precedenti per implementare una fabbrica di migrazione di team, strumenti e processi per semplificare la migrazione dei carichi di lavoro attraverso l'automazione e la distribuzione agile. Questa è la terza fase della strategia di migrazione AWS.

fabbrica di migrazione

Team interfunzionali che semplificano la migrazione dei carichi di lavoro attraverso approcci automatizzati e agili. I team di Migration Factory in genere includono addetti alle operazioni,

analisti e proprietari aziendali, ingegneri addetti alla migrazione, sviluppatori e DevOps professionisti che lavorano nell'ambito degli sprint. Tra il 20% e il 50% di un portfolio di applicazioni aziendali è costituito da schemi ripetuti che possono essere ottimizzati con un approccio di fabbrica. Per ulteriori informazioni, consulta la discussione sulle fabbriche di migrazione e la Guida alla fabbrica di migrazione al cloud in questo set di contenuti.

metadati di migrazione

Le informazioni sull'applicazione e sul server necessarie per completare la migrazione. Ogni modello di migrazione richiede un set diverso di metadati di migrazione. Esempi di metadati di migrazione includono la sottorete, il gruppo di sicurezza e l'account di destinazione. AWS

modello di migrazione

Un'attività di migrazione ripetibile che descrive in dettaglio la strategia di migrazione, la destinazione della migrazione e l'applicazione o il servizio di migrazione utilizzati. Esempio: riorganizza la migrazione su Amazon EC2 con AWS Application Migration Service.

Valutazione del portfolio di migrazione (MPA)

Uno strumento online che fornisce informazioni per la convalida del business case per la migrazione a. Cloud AWS MPA offre una valutazione dettagliata del portfolio (dimensionamento corretto dei server, prezzi, confronto del TCO, analisi dei costi di migrazione) e pianificazione della migrazione (analisi e raccolta dei dati delle applicazioni, raggruppamento delle applicazioni, prioritizzazione delle migrazioni e pianificazione delle ondate). Lo strumento MPA (richiede l'accesso) è disponibile gratuitamente per tutti i AWS consulenti e i consulenti dei partner APN.

valutazione della preparazione alla migrazione (MRA)

Il processo di acquisizione di informazioni sullo stato di preparazione al cloud di un'organizzazione, l'identificazione dei punti di forza e di debolezza e la creazione di un piano d'azione per colmare le lacune identificate, utilizzando il CAF. AWS Per ulteriori informazioni, consulta la guida di preparazione alla migrazione. MRA è la prima fase della strategia di migrazione AWS.

strategia di migrazione

L'approccio utilizzato per migrare un carico di lavoro verso. Cloud AWS Per ulteriori informazioni, consulta la voce 7 R in questo glossario e consulta Mobilita la tua organizzazione per accelerare le migrazioni su larga scala.

ML

Vedi machine learning.

modernizzazione

Trasformazione di un'applicazione obsoleta (legacy o monolitica) e della relativa infrastruttura in un sistema agile, elastico e altamente disponibile nel cloud per ridurre i costi, aumentare l'efficienza e sfruttare le innovazioni. Per ulteriori informazioni, vedere Strategia per la modernizzazione delle applicazioni in. Cloud AWS

valutazione della preparazione alla modernizzazione

Una valutazione che aiuta a determinare la preparazione alla modernizzazione delle applicazioni di un'organizzazione, identifica vantaggi, rischi e dipendenze e determina in che misura l'organizzazione può supportare lo stato futuro di tali applicazioni. Il risultato della valutazione è uno schema dell'architettura di destinazione, una tabella di marcia che descrive in dettaglio le fasi di sviluppo e le tappe fondamentali del processo di modernizzazione e un piano d'azione per colmare le lacune identificate. Per ulteriori informazioni, vedere Valutazione della preparazione alla modernizzazione per le applicazioni in. Cloud AWS

applicazioni monolitiche (monoliti)

Applicazioni eseguite come un unico servizio con processi strettamente collegati. Le applicazioni monolitiche presentano diversi inconvenienti. Se una funzionalità dell'applicazione registra un picco di domanda, l'intera architettura deve essere dimensionata. L'aggiunta o il miglioramento delle funzionalità di un'applicazione monolitica diventa inoltre più complessa man mano che la base di codice cresce. Per risolvere questi problemi, puoi utilizzare un'architettura di microservizi. Per ulteriori informazioni, consulta la sezione Scomposizione dei monoliti in microservizi.

MAPPA

Vedi Migration Portfolio Assessment.

MQTT

Vedi Message Queuing Telemetry Transport.

classificazione multiclasse

Un processo che aiuta a generare previsioni per più classi (prevedendo uno o più di due risultati). Ad esempio, un modello di machine learning potrebbe chiedere "Questo prodotto è un libro, un'auto o un telefono?" oppure "Quale categoria di prodotti è più interessante per questo cliente?"

infrastruttura mutabile

Un modello che aggiorna e modifica l'infrastruttura esistente per i carichi di lavoro di produzione. Per migliorare la coerenza, l'affidabilità e la prevedibilità, il AWS Well-Architected Framework consiglia l'uso di un'infrastruttura immutabile come best practice.

O

OAC

Vedi Origin Access Control.

QUERCIA

Vedi Origin Access Identity.

OCM

Vedi gestione delle modifiche organizzative.

migrazione offline

Un metodo di migrazione in cui il carico di lavoro di origine viene eliminato durante il processo di migrazione. Questo metodo prevede tempi di inattività prolungati e viene in genere utilizzato per carichi di lavoro piccoli e non critici.

OI

Vedi l'integrazione delle operazioni.

OLA

Vedi accordo a livello operativo.

migrazione online

Un metodo di migrazione in cui il carico di lavoro di origine viene copiato sul sistema di destinazione senza essere messo offline. Le applicazioni connesse al carico di lavoro possono continuare a funzionare durante la migrazione. Questo metodo comporta tempi di inattività pari a zero o comunque minimi e viene in genere utilizzato per carichi di lavoro di produzione critici.

OPC-UA

Vedi Open Process Communications - Unified Architecture.

Comunicazioni a processo aperto - Architettura unificata (OPC-UA)

Un protocollo di comunicazione machine-to-machine (M2M) per l'automazione industriale. OPC-UA fornisce uno standard di interoperabilità con schemi di crittografia, autenticazione e autorizzazione dei dati.

accordo a livello operativo (OLA)

Un accordo che chiarisce quali sono gli impegni reciproci tra i gruppi IT funzionali, a supporto di un accordo sul livello di servizio (SLA).

revisione della prontezza operativa (ORR)

Un elenco di domande e best practice associate che aiutano a comprendere, valutare, prevenire o ridurre la portata degli incidenti e dei possibili guasti. Per ulteriori informazioni, vedere Operational Readiness Reviews (ORR) nel Well-Architected AWS Framework.

tecnologia operativa (OT)

Sistemi hardware e software che interagiscono con l'ambiente fisico per controllare le operazioni, le apparecchiature e le infrastrutture industriali. Nella produzione, l'integrazione di sistemi OT e di tecnologia dell'informazione (IT) è un obiettivo chiave per le trasformazioni dell'Industria 4.0.

integrazione delle operazioni (OI)

Il processo di modernizzazione delle operazioni nel cloud, che prevede la pianificazione, l'automazione e l'integrazione della disponibilità. Per ulteriori informazioni, consulta la guida all'integrazione delle operazioni.

trail organizzativo

Un percorso creato da noi AWS CloudTrail che registra tutti gli eventi di un'organizzazione per tutti Account AWS . AWS Organizations Questo percorso viene creato in ogni Account AWS che fa parte dell'organizzazione e tiene traccia dell'attività in ogni account. Per ulteriori informazioni, consulta Creazione di un percorso per un'organizzazione nella CloudTrail documentazione.

gestione del cambiamento organizzativo (OCM)

Un framework per la gestione di trasformazioni aziendali importanti e che comportano l'interruzione delle attività dal punto di vista delle persone, della cultura e della leadership. OCM aiuta le organizzazioni a prepararsi e passare a nuovi sistemi e strategie accelerando l'adozione del cambiamento, affrontando i problemi di transizione e promuovendo cambiamenti culturali e organizzativi. Nella strategia di AWS migrazione, questo framework si chiama accelerazione delle

persone, a causa della velocità di cambiamento richiesta nei progetti di adozione del cloud. Per ulteriori informazioni, consultare la Guida OCM.

controllo dell'accesso all'origine (OAC)

In CloudFront, un'opzione avanzata per limitare l'accesso per proteggere i contenuti di Amazon Simple Storage Service (Amazon S3). OAC supporta tutti i bucket S3 in generale Regioni AWS, la crittografia lato server con AWS KMS (SSE-KMS) e le richieste dinamiche e dirette al bucket S3. `PUT` `DELETE`

identità di accesso origine (OAI)

Nel CloudFront, un'opzione per limitare l'accesso per proteggere i tuoi contenuti Amazon S3. Quando usi OAI, CloudFront crea un principale con cui Amazon S3 può autenticarsi. I principali autenticati possono accedere ai contenuti in un bucket S3 solo tramite una distribuzione specifica. CloudFront Vedi anche OAC, che fornisce un controllo degli accessi più granulare e avanzato.

ORR

Vedi la revisione della prontezza operativa.

- NON

Vedi la tecnologia operativa.

VPC in uscita (egress)

In un'architettura AWS multi-account, un VPC che gestisce le connessioni di rete avviate dall'interno di un'applicazione. La AWS Security Reference Architecture consiglia di configurare l'account di rete con funzionalità in entrata, in uscita e di ispezione VPCs per proteggere l'interfaccia bidirezionale tra l'applicazione e Internet in generale.

P

limite delle autorizzazioni

Una policy di gestione IAM collegata ai principali IAM per impostare le autorizzazioni massime che l'utente o il ruolo possono avere. Per ulteriori informazioni, consulta Limiti delle autorizzazioni nella documentazione di IAM.

informazioni di identificazione personale (PII)

Informazioni che, se visualizzate direttamente o abbinate ad altri dati correlati, possono essere utilizzate per dedurre ragionevolmente l'identità di un individuo. Esempi di informazioni personali includono nomi, indirizzi e informazioni di contatto.

Informazioni che consentono l'identificazione personale degli utenti

Visualizza le informazioni di identificazione personale.

playbook

Una serie di passaggi predefiniti che raccolgono il lavoro associato alle migrazioni, come l'erogazione delle funzioni operative principali nel cloud. Un playbook può assumere la forma di script, runbook automatici o un riepilogo dei processi o dei passaggi necessari per gestire un ambiente modernizzato.

PLC

Vedi controllore logico programmabile.

PLM

Vedi la gestione del ciclo di vita del prodotto.

policy

Un oggetto in grado di definire le autorizzazioni (vedi politica basata sull'identità), specificare le condizioni di accesso (vedi politicabasata sulle risorse) o definire le autorizzazioni massime per tutti gli account di un'organizzazione in (vedi politica di controllo dei servizi). AWS Organizations

persistenza poliglotta

Scelta indipendente della tecnologia di archiviazione di dati di un microservizio in base ai modelli di accesso ai dati e ad altri requisiti. Se i microservizi utilizzano la stessa tecnologia di archiviazione di dati, possono incontrare problemi di implementazione o registrare prestazioni scadenti. I microservizi vengono implementati più facilmente e ottengono prestazioni e scalabilità migliori se utilizzano l'archivio dati più adatto alle loro esigenze. Per ulteriori informazioni, consulta la sezione Abilitazione della persistenza dei dati nei microservizi.

valutazione del portfolio

Un processo di scoperta, analisi e definizione delle priorità del portfolio di applicazioni per pianificare la migrazione. Per ulteriori informazioni, consulta la pagina Valutazione della preparazione alla migrazione.

predicate

Una condizione di interrogazione che restituisce o, in genere, si trova in una clausola`WHERE`. `true` `false`

predicato pushdown

Una tecnica di ottimizzazione delle query del database che filtra i dati della query prima del trasferimento. Ciò riduce la quantità di dati che devono essere recuperati ed elaborati dal database relazionale e migliora le prestazioni delle query.

controllo preventivo

Un controllo di sicurezza progettato per impedire il verificarsi di un evento. Questi controlli sono la prima linea di difesa per impedire accessi non autorizzati o modifiche indesiderate alla rete. Per ulteriori informazioni, consulta Controlli preventivi in Implementazione dei controlli di sicurezza in AWS.

principale

Un'entità in AWS grado di eseguire azioni e accedere alle risorse. Questa entità è in genere un utente root per un Account AWS ruolo IAM o un utente. Per ulteriori informazioni, consulta Principali in Termini e concetti dei ruoli nella documentazione di IAM.

privacy fin dalla progettazione

Un approccio di ingegneria dei sistemi che tiene conto della privacy durante l'intero processo di sviluppo.

zone ospitate private

Un contenitore che contiene informazioni su come desideri che Amazon Route 53 risponda alle query DNS per un dominio e i relativi sottodomini all'interno di uno o più. VPCs Per ulteriori informazioni, consulta Utilizzo delle zone ospitate private nella documentazione di Route 53.

controllo proattivo

Un controllo di sicurezza progettato per impedire l'implementazione di risorse non conformi. Questi controlli analizzano le risorse prima del loro provisioning. Se la risorsa non è conforme al controllo, non viene fornita. Per ulteriori informazioni, consulta la guida di riferimento sui controlli nella AWS Control Tower documentazione e consulta Controlli proattivi in Implementazione dei controlli di sicurezza su. AWS

gestione del ciclo di vita del prodotto (PLM)

La gestione dei dati e dei processi di un prodotto durante l'intero ciclo di vita, dalla progettazione, sviluppo e lancio, attraverso la crescita e la maturità, fino al declino e alla rimozione.

Ambiente di produzione

Vedi ambiente.

controllore logico programmabile (PLC)

Nella produzione, un computer altamente affidabile e adattabile che monitora le macchine e automatizza i processi di produzione.

concatenamento rapido

Utilizzo dell'output di un prompt LLM come input per il prompt successivo per generare risposte migliori. Questa tecnica viene utilizzata per suddividere un'attività complessa in sottoattività o per perfezionare o espandere iterativamente una risposta preliminare. Aiuta a migliorare l'accuratezza e la pertinenza delle risposte di un modello e consente risultati più granulari e personalizzati.

pseudonimizzazione

Il processo di sostituzione degli identificatori personali in un set di dati con valori segnaposto. La pseudonimizzazione può aiutare a proteggere la privacy personale. I dati pseudonimizzati sono ancora considerati dati personali.

publish/subscribe (pub/sub)

Un modello che consente comunicazioni asincrone tra microservizi per migliorare la scalabilità e la reattività. Ad esempio, in un MES basato su microservizi, un microservizio può pubblicare messaggi di eventi su un canale a cui altri microservizi possono abbonarsi. Il sistema può aggiungere nuovi microservizi senza modificare il servizio di pubblicazione.

Q

Piano di query

Una serie di passaggi, come le istruzioni, utilizzati per accedere ai dati in un sistema di database relazionale SQL.

regressione del piano di query

Quando un ottimizzatore del servizio di database sceglie un piano non ottimale rispetto a prima di una determinata modifica all'ambiente di database. Questo può essere causato da modifiche a statistiche, vincoli, impostazioni dell'ambiente, associazioni dei parametri di query e aggiornamenti al motore di database.

R

Matrice RACI

Vedi responsabile, responsabile, consultato, informato (RACI).

STRACCIO

Vedi Retrieval Augmented Generation.

ransomware

Un software dannoso progettato per bloccare l'accesso a un sistema informatico o ai dati fino a quando non viene effettuato un pagamento.

Matrice RASCI

Vedi responsabile, responsabile, consultato, informato (RACI).

RCAC

Vedi controllo dell'accesso a righe e colonne.

replica di lettura

Una copia di un database utilizzata per scopi di sola lettura. È possibile indirizzare le query alla replica di lettura per ridurre il carico sul database principale.

riprogettare

Vedi 7 Rs.

obiettivo del punto di ripristino (RPO)

Il periodo di tempo massimo accettabile dall'ultimo punto di ripristino dei dati. Questo determina ciò che si considera una perdita di dati accettabile tra l'ultimo punto di ripristino e l'interruzione del servizio.

obiettivo del tempo di ripristino (RTO)

Il ritardo massimo accettabile tra l'interruzione del servizio e il ripristino del servizio.

rifattorizzare

Vedi 7 R.

Regione

Una raccolta di AWS risorse in un'area geografica. Ciascuna Regione AWS è isolata e indipendente dalle altre per fornire tolleranza agli errori, stabilità e resilienza. Per ulteriori informazioni, consulta Specificare cosa può usare Regioni AWS il tuo account.

regressione

Una tecnica di ML che prevede un valore numerico. Ad esempio, per risolvere il problema "A che prezzo verrà venduta questa casa?" un modello di ML potrebbe utilizzare un modello di regressione lineare per prevedere il prezzo di vendita di una casa sulla base di dati noti sulla casa (ad esempio, la metratura).

riospitare

Vedi 7 R.

rilascio

In un processo di implementazione, l'atto di promuovere modifiche a un ambiente di produzione.

trasferisco

Vedi 7 Rs.

ripiattaforma

Vedi 7 Rs.

riacquisto

Vedi 7 Rs.

resilienza

La capacità di un'applicazione di resistere o ripristinare le interruzioni. L'elevata disponibilità e il disaster recovery sono considerazioni comuni quando si pianifica la resilienza in. Cloud AWSPer ulteriori informazioni, vedere Cloud AWS Resilience.

policy basata su risorse

Una policy associata a una risorsa, ad esempio un bucket Amazon S3, un endpoint o una chiave di crittografia. Questo tipo di policy specifica a quali principali è consentito l'accesso, le azioni supportate e qualsiasi altra condizione che deve essere soddisfatta.

matrice di assegnazione di responsabilità (RACI)

Una matrice che definisce i ruoli e le responsabilità di tutte le parti coinvolte nelle attività di migrazione e nelle operazioni cloud. Il nome della matrice deriva dai tipi di responsabilità definiti nella matrice: responsabile (R), responsabile (A), consultato (C) e informato (I). Il tipo di supporto (S) è facoltativo. Se includi il supporto, la matrice viene chiamata matrice RASCI e, se la escludi, viene chiamata matrice RACI.

controllo reattivo

Un controllo di sicurezza progettato per favorire la correzione di eventi avversi o deviazioni dalla baseline di sicurezza. Per ulteriori informazioni, consulta Controlli reattivi in Implementazione dei controlli di sicurezza in AWS.

retain

Vedi 7 R.

andare in pensione

Vedi 7 Rs.

Retrieval Augmented Generation (RAG)

Una tecnologia di intelligenza artificiale generativa in cui un LLM fa riferimento a una fonte di dati autorevole esterna alle sue fonti di dati di formazione prima di generare una risposta. Ad esempio, un modello RAG potrebbe eseguire una ricerca semantica nella knowledge base o nei dati personalizzati di un'organizzazione. Per ulteriori informazioni, consulta Cos'è il RAG.

rotazione

Processo di aggiornamento periodico di un segreto per rendere più difficile l'accesso alle credenziali da parte di un utente malintenzionato.

controllo dell'accesso a righe e colonne (RCAC)

L'uso di espressioni SQL di base e flessibili con regole di accesso definite. RCAC è costituito da autorizzazioni di riga e maschere di colonna.

RPO

Vedi l'obiettivo del punto di ripristino.

RTO

Vedi l'obiettivo del tempo di ripristino.

runbook

Un insieme di procedure manuali o automatizzate necessarie per eseguire un'attività specifica. In genere sono progettati per semplificare operazioni o procedure ripetitive con tassi di errore elevati.

S

SAML 2.0

Uno standard aperto utilizzato da molti provider di identità (IdPs). Questa funzionalità abilita il single sign-on (SSO) federato, in modo che gli utenti possano accedere AWS Management Console o chiamare le operazioni AWS API senza che tu debba creare un utente in IAM per tutti i membri dell'organizzazione. Per ulteriori informazioni sulla federazione basata su SAML 2.0, consulta Informazioni sulla federazione basata su SAML 2.0 nella documentazione di IAM.

SCADA

Vedi controllo di supervisione e acquisizione dati.

SCP

Vedi la politica di controllo del servizio.

Secret

In AWS Secrets Manager, informazioni riservate o riservate, come una password o le credenziali utente, archiviate in forma crittografata. È costituito dal valore segreto e dai relativi metadati. Il valore segreto può essere binario, una stringa singola o più stringhe. Per ulteriori informazioni, consulta Cosa c'è in un segreto di Secrets Manager? nella documentazione di Secrets Manager.

sicurezza fin dalla progettazione

Un approccio di ingegneria dei sistemi che tiene conto della sicurezza durante l'intero processo di sviluppo.

controllo di sicurezza

Un guardrail tecnico o amministrativo che impedisce, rileva o riduce la capacità di un autore di minacce di sfruttare una vulnerabilità di sicurezza. Esistono quattro tipi principali di controlli di sicurezza: preventivi, investigativi, reattivi e proattivi.

rafforzamento della sicurezza

Il processo di riduzione della superficie di attacco per renderla più resistente agli attacchi. Può includere azioni come la rimozione di risorse che non sono più necessarie, l'implementazione di best practice di sicurezza che prevedono la concessione del privilegio minimo o la disattivazione di funzionalità non necessarie nei file di configurazione.

sistema di gestione delle informazioni e degli eventi di sicurezza (SIEM)

Strumenti e servizi che combinano sistemi di gestione delle informazioni di sicurezza (SIM) e sistemi di gestione degli eventi di sicurezza (SEM). Un sistema SIEM raccoglie, monitora e analizza i dati da server, reti, dispositivi e altre fonti per rilevare minacce e violazioni della sicurezza e generare avvisi.

automazione della risposta alla sicurezza

Un'azione predefinita e programmata progettata per rispondere o porre rimedio automaticamente a un evento di sicurezza. Queste automazioni fungono da controlli di sicurezza investigativi o reattivi che aiutano a implementare le migliori pratiche di sicurezza. AWS Esempi di azioni di risposta automatizzate includono la modifica di un gruppo di sicurezza VPC, l'applicazione di patch a un'istanza EC2 Amazon o la rotazione delle credenziali.

Crittografia lato server

Crittografia dei dati a destinazione, da parte di chi li riceve. Servizio AWS

Policy di controllo dei servizi (SCP)

Una politica che fornisce il controllo centralizzato sulle autorizzazioni per tutti gli account di un'organizzazione in. AWS Organizations SCPs definire barriere o fissare limiti alle azioni che un amministratore può delegare a utenti o ruoli. È possibile utilizzarli SCPs come elenchi consentiti o elenchi di rifiuto, per specificare quali servizi o azioni sono consentiti o proibiti. Per ulteriori informazioni, consulta le politiche di controllo del servizio nella AWS Organizations documentazione.

endpoint del servizio

L'URL del punto di ingresso per un Servizio AWS. Puoi utilizzare l'endpoint per connetterti a livello di programmazione al servizio di destinazione. Per ulteriori informazioni, consulta Endpoint del Servizio AWS nei Riferimenti generali di AWS.

accordo sul livello di servizio (SLA)

Un accordo che chiarisce ciò che un team IT promette di offrire ai propri clienti, ad esempio l'operatività e le prestazioni del servizio.

indicatore del livello di servizio (SLI)

Misurazione di un aspetto prestazionale di un servizio, ad esempio il tasso di errore, la disponibilità o la velocità effettiva.

obiettivo a livello di servizio (SLO)

Una metrica target che rappresenta lo stato di un servizio, misurato da un indicatore del livello di servizio.

Modello di responsabilità condivisa

Un modello che descrive la responsabilità condivisa AWS per la sicurezza e la conformità del cloud. AWS è responsabile della sicurezza del cloud, mentre tu sei responsabile della sicurezza nel cloud. Per ulteriori informazioni, consulta Modello di responsabilità condivisa.

SIEM

Vedi il sistema di gestione delle informazioni e degli eventi sulla sicurezza.

punto di errore singolo (SPOF)

Un guasto in un singolo componente critico di un'applicazione che può disturbare il sistema.

SLAM

Vedi il contratto sul livello di servizio.

SLI

Vedi l'indicatore del livello di servizio.

LENTA

Vedi obiettivo del livello di servizio.

split-and-seed modello

Un modello per dimensionare e accelerare i progetti di modernizzazione. Man mano che vengono definite nuove funzionalità e versioni dei prodotti, il team principale si divide per creare nuovi team di prodotto. Questo aiuta a dimensionare le capacità e i servizi dell'organizzazione, migliora la produttività degli sviluppatori e supporta una rapida innovazione. Per ulteriori informazioni, vedere Approccio graduale alla modernizzazione delle applicazioni in. Cloud AWS

SPOF

Vedi punto di errore singolo.

schema a stella

Una struttura organizzativa di database che utilizza un'unica tabella dei fatti di grandi dimensioni per archiviare i dati transazionali o misurati e utilizza una o più tabelle dimensionali più piccole per memorizzare gli attributi dei dati. Questa struttura è progettata per l'uso in un data warehouse o per scopi di business intelligence.

modello del fico strangolatore

Un approccio alla modernizzazione dei sistemi monolitici mediante la riscrittura e la sostituzione incrementali delle funzionalità del sistema fino alla disattivazione del sistema legacy. Questo modello utilizza l'analogia di una pianta di fico che cresce fino a diventare un albero robusto e alla fine annienta e sostituisce il suo ospite. Il modello è stato introdotto da Martin Fowler come metodo per gestire il rischio durante la riscrittura di sistemi monolitici. Per un esempio di come applicare questo modello, consulta Modernizzazione incrementale dei servizi Web legacy di Microsoft ASP.NET (ASMX) mediante container e Gateway Amazon API.

sottorete

Un intervallo di indirizzi IP nel VPC. Una sottorete deve risiedere in una singola zona di disponibilità.

controllo di supervisione e acquisizione dati (SCADA)

Nella produzione, un sistema che utilizza hardware e software per monitorare gli asset fisici e le operazioni di produzione.

crittografia simmetrica

Un algoritmo di crittografia che utilizza la stessa chiave per crittografare e decrittografare i dati.

test sintetici

Test di un sistema in modo da simulare le interazioni degli utenti per rilevare potenziali problemi o monitorare le prestazioni. Puoi usare Amazon CloudWatch Synthetics per creare questi test.

prompt di sistema

Una tecnica per fornire contesto, istruzioni o linee guida a un LLM per indirizzarne il comportamento. I prompt di sistema aiutano a impostare il contesto e stabilire regole per le interazioni con gli utenti.

T

tags

Coppie chiave-valore che fungono da metadati per l'organizzazione delle risorse. AWS Con i tag è possibile a gestire, identificare, organizzare, cercare e filtrare le risorse. Per ulteriori informazioni, consulta Tagging delle risorse AWS.

variabile di destinazione

Il valore che stai cercando di prevedere nel machine learning supervisionato. Questo è indicato anche come variabile di risultato. Ad esempio, in un ambiente di produzione la variabile di destinazione potrebbe essere un difetto del prodotto.

elenco di attività

Uno strumento che viene utilizzato per tenere traccia dei progressi tramite un runbook. Un elenco di attività contiene una panoramica del runbook e un elenco di attività generali da completare. Per ogni attività generale, include la quantità stimata di tempo richiesta, il proprietario e lo stato di avanzamento.

Ambiente di test

Vedi ambiente.

training

Fornire dati da cui trarre ispirazione dal modello di machine learning. I dati di training devono contenere la risposta corretta. L'algoritmo di apprendimento trova nei dati di addestramento i pattern che mappano gli attributi dei dati di input al target (la risposta che si desidera prevedere). Produce un modello di ML che acquisisce questi modelli. Puoi quindi utilizzare il modello di ML per creare previsioni su nuovi dati di cui non si conosce il target.

Transit Gateway

Un hub di transito di rete che puoi utilizzare per interconnettere le tue reti VPCs e quelle locali. Per ulteriori informazioni, consulta Cos'è un gateway di transito nella AWS Transit Gateway documentazione.

flusso di lavoro basato su trunk

Un approccio in cui gli sviluppatori creano e testano le funzionalità localmente in un ramo di funzionalità e quindi uniscono tali modifiche al ramo principale. Il ramo principale viene quindi integrato negli ambienti di sviluppo, preproduzione e produzione, in sequenza.

Accesso attendibile

Concessione delle autorizzazioni a un servizio specificato dall'utente per eseguire attività all'interno dell'organizzazione AWS Organizations e nei suoi account per conto dell'utente. Il servizio attendibile crea un ruolo collegato al servizio in ogni account, quando tale ruolo è necessario, per eseguire attività di gestione per conto dell'utente. Per ulteriori informazioni, consulta Utilizzo AWS Organizations con altri AWS servizi nella AWS Organizations documentazione.

regolazione

Modificare alcuni aspetti del processo di training per migliorare la precisione del modello di ML. Ad esempio, puoi addestrare il modello di ML generando un set di etichette, aggiungendo etichette e quindi ripetendo questi passaggi più volte con impostazioni diverse per ottimizzare il modello.

team da due pizze

Una piccola DevOps squadra che puoi sfamare con due pizze. Un team composto da due persone garantisce la migliore opportunità possibile di collaborazione nello sviluppo del software.

U

incertezza

Un concetto che si riferisce a informazioni imprecise, incomplete o sconosciute che possono minare l'affidabilità dei modelli di machine learning predittivi. Esistono due tipi di incertezza: l'incertezza epistemica, che è causata da dati limitati e incompleti, mentre l'incertezza aleatoria è causata dal rumore e dalla casualità insiti nei dati. Per ulteriori informazioni, consulta la guida Quantificazione dell'incertezza nei sistemi di deep learning.

compiti indifferenziati

Conosciuto anche come sollevamento di carichi pesanti, è un lavoro necessario per creare e far funzionare un'applicazione, ma che non apporta valore diretto all'utente finale né offre vantaggi competitivi. Esempi di attività indifferenziate includono l'approvvigionamento, la manutenzione e la pianificazione della capacità.

ambienti superiori

Vedi ambiente.

V

vacuum

Un'operazione di manutenzione del database che prevede la pulizia dopo aggiornamenti incrementali per recuperare lo spazio di archiviazione e migliorare le prestazioni.

controllo delle versioni

Processi e strumenti che tengono traccia delle modifiche, ad esempio le modifiche al codice di origine in un repository.

Peering VPC

Una connessione tra due VPCs che consente di indirizzare il traffico utilizzando indirizzi IP privati. Per ulteriori informazioni, consulta Che cos'è il peering VPC? nella documentazione di Amazon VPC.

vulnerabilità

Un difetto software o hardware che compromette la sicurezza del sistema.

W

cache calda

Una cache del buffer che contiene dati correnti e pertinenti a cui si accede frequentemente. L'istanza di database può leggere dalla cache del buffer, il che richiede meno tempo rispetto alla lettura dalla memoria dal disco principale.

dati caldi

Dati a cui si accede raramente. Quando si eseguono interrogazioni di questo tipo di dati, in genere sono accettabili query moderatamente lente.

funzione finestra

Una funzione SQL che esegue un calcolo su un gruppo di righe che si riferiscono in qualche modo al record corrente. Le funzioni della finestra sono utili per l'elaborazione di attività, come il calcolo di una media mobile o l'accesso al valore delle righe in base alla posizione relativa della riga corrente.

Carico di lavoro

Una raccolta di risorse e codice che fornisce valore aziendale, ad esempio un'applicazione rivolta ai clienti o un processo back-end.

flusso di lavoro

Gruppi funzionali in un progetto di migrazione responsabili di una serie specifica di attività. Ogni flusso di lavoro è indipendente ma supporta gli altri flussi di lavoro del progetto. Ad esempio, il flusso di lavoro del portfolio è responsabile della definizione delle priorità delle applicazioni, della pianificazione delle ondate e della raccolta dei metadati di migrazione. Il flusso di lavoro del portfolio fornisce queste risorse al flusso di lavoro di migrazione, che quindi migra i server e le applicazioni.

VERME

Vedi scrivere una volta, leggere molti.

WQF

Vedi AWS Workload Qualification Framework.

scrivi una volta, leggi molte (WORM)

Un modello di storage che scrive i dati una sola volta e ne impedisce l'eliminazione o la modifica. Gli utenti autorizzati possono leggere i dati tutte le volte che è necessario, ma non possono modificarli. Questa infrastruttura di archiviazione dei dati è considerata immutabile.

Z

exploit zero-day

Un attacco, in genere malware, che sfrutta una vulnerabilità zero-day.

vulnerabilità zero-day

Un difetto o una vulnerabilità assoluta in un sistema di produzione. Gli autori delle minacce possono utilizzare questo tipo di vulnerabilità per attaccare il sistema. Gli sviluppatori vengono spesso a conoscenza della vulnerabilità causata dall'attacco.

prompt zero-shot

Fornire a un LLM le istruzioni per eseguire un'attività ma non esempi (immagini) che possano aiutarla. Il LLM deve utilizzare le sue conoscenze pre-addestrate per gestire l'attività. L'efficacia del prompt zero-shot dipende dalla complessità dell'attività e dalla qualità del prompt. Vedi anche few-shot prompting.

applicazione zombie

Un'applicazione che prevede un utilizzo CPU e memoria inferiore al 5%. In un progetto di migrazione, è normale ritirare queste applicazioni.

Le traduzioni sono generate tramite traduzione automatica. In caso di conflitto tra il contenuto di una traduzione e la versione originale in Inglese, quest'ultima prevarrà.