Criação de soluções de geração aumentada de recuperação para assistência médica AWS

AWS Orientação prescritiva

AWS Orientação prescritiva: Criação de soluções de geração aumentada de recuperação para assistência médica AWS

Table of Contents

Criação de soluções de geração aumentada de recuperação para assistência médica AWS

Amazon Web Services, Accenture e Cadiem (contribuidores)

Março de 2025 (histórico do documento)

Antes dos grandes modelos de linguagem (LLMs) e da IA generativa, a tarefa de desenvolver aplicativos automatizados e de alta precisão no setor de saúde era desafiadora. Os métodos tradicionais dependiam muito da entrada e análise manuais de dados. A complexidade de analisar imagens médicas e registros de pacientes exigiu ampla intervenção humana, o que muitas vezes resultou em fluxos de trabalho fragmentados e ineficientes. O avanço das tecnologias de IA ajuda você a criar aplicativos hiperpersonalizados em grande escala. Os aplicativos de saúde agora podem se integrar às bases de conhecimento médico, interpretar imagens de diagnóstico com maior precisão e prever os resultados dos pacientes usando modelos preditivos.

Este guia explora como LLMs estão revolucionando a assistência médica por meio de aplicativos de geração aumentada de recuperação com os quais você pode criar. Serviços da AWS A Retrieval Augmented Generation (RAG) é uma tecnologia generativa de IA na qual um LLM faz referência a uma fonte de dados autorizada que está fora de suas fontes de dados de treinamento antes de gerar uma resposta. Os aplicativos RAG baseiam a produção do modelo no conhecimento do mundo real, o que reduz as alucinações e aumenta a relevância da resposta. No setor de saúde, o RAG pode ser usado para fornecer informações up-to-date médicas precisas, garantindo que os profissionais de saúde tenham acesso às pesquisas e diretrizes clínicas mais recentes. Ao transformar dados em insights acionáveis e automatizar processos complexos, essas tecnologias ajudam a aprimorar o atendimento ao paciente, agilizar as operações e melhorar a produtividade dos profissionais de saúde.

No Amazon Bedrock, você pode ajustá-los LLMs e integrá-los com agentes inteligentes para criar soluções avançadas de saúde. Destacando a sinergia entre o Amazon OpenSearch Service e o Amazon Neptune, o guia demonstra como esses serviços elevam as soluções RAG por meio de maior relevância de pesquisa e recuperação avançada de dados de várias fontes. Você pode orquestrar soluções abrangentes do Amazon Bedrock que usam agentes e LangChainpara coordenar perfeitamente as interações em diversos repositórios de dados. Essa integração demonstra o poder de combinar serviços especializados para criar sistemas orientados por IA mais eficazes e eficientes.

Atendimento ao paciente e produtividade

Este guia apresenta dois casos de uso reais para atendimento e produtividade do paciente: aumento dos dados do paciente e previsão de riscos de readmissão. Ele fornece planos estratégicos para implementar essas soluções em grande escala, oferecendo às organizações de saúde um caminho claro para a industrialização de processos orientados por IA. Por meio desses insights, as instituições de saúde podem usar tecnologias avançadas de IA para criar fluxos de trabalho mais eficientes e inteligentes.

Gestão de talentos

Este guia também descreve estratégias para requalificar e capacitar profissionais de saúde para integrar perfeitamente a IA generativa em suas rotinas diárias. Isso pode aumentar a produtividade e a qualidade do atendimento ao paciente. Ao equipar sua força de trabalho com as habilidades para usar com eficácia ferramentas avançadas de IA, as organizações de saúde podem maximizar o retorno sobre o investimento e impulsionar a inovação no atendimento ao paciente.

Essa solução de gerenciamento de talentos baseada em IA inclui os seguintes recursos principais:

- Analisador inteligente de currículos de talentos — Ao usar o avançado LLMs disponível no Amazon Bedrock, essa ferramenta extrai e analisa com eficiência as habilidades e os atributos essenciais dos currículos. Essa ferramenta pode agilizar o processo de recrutamento.
- Base de conhecimento de talentos — Desenvolvido pelo Amazon Neptune, esse banco de dados dinâmico fornece informações em tempo real sobre os níveis de pessoal, distribuição de habilidades e tendências do setor. Isso ajuda você a tomar decisões baseadas em dados sobre o gerenciamento da força de trabalho.
- Mecanismo de recomendação de aprendizado — Essa ferramenta baseada em IA identifica lacunas de habilidades dentro da organização e recomenda programas de treinamento personalizados para a equipe médica. Essa ferramenta promove o desenvolvimento profissional contínuo e ajuda sua força de trabalho a se adaptar às tecnologias de saúde em evolução.

Juntos, esses recursos baseados em IA ajudam a otimizar o desempenho da força de trabalho, revolucionando o gerenciamento de talentos com maior inteligência e eficiência.

Oportunidades e desafios

O Amazon Bedrock pode fornecer maior produtividade, escalabilidade, economia e insights baseados em dados. O Amazon Bedrock capacita as organizações de saúde a usarem de LLMs forma eficaz em vários casos de uso, desde a criação de conteúdo e análise de dados até a tomada de decisão automatizada. Este guia fornece abordagens para superar desafios comuns de IA generativa, como problemas de qualidade de dados, escalabilidade da infraestrutura, manutenção do desempenho do modelo e requisitos de melhoria contínua durante a transição da prova de conceito para a produção.

Oportunidades para aplicações generativas de IA na área da saúde

O setor de saúde está pronto para uma mudança transformadora, impulsionada pelas oportunidades apresentadas pelos aplicativos generativos de IA. A IA generativa tem o potencial de aprimorar o atendimento ao paciente, agilizar as operações e acelerar a pesquisa médica. Ao usar modelos avançados de IA, os profissionais de saúde podem automatizar o aumento dos registros médicos. Histórias abrangentes e de up-to-date pacientes facilitam diagnósticos e planos de tratamento mais precisos. A análise de imagens baseada em IA, como a interpretação de ultrassonografias e outras imagens médicas, pode fornecer informações rápidas e precisas, reduzindo a carga de trabalho dos profissionais médicos e minimizando o risco de erro humano.

Além do diagnóstico e do tratamento, a IA generativa pode desempenhar um papel fundamental na análise preditiva. A análise preditiva ajuda as organizações de saúde a antecipar os resultados dos pacientes e personalizar os planos de tratamento adequadamente. Essa tecnologia também pode otimizar os processos administrativos, desde o gerenciamento de dados do paciente até a simplificação da comunicação entre provedores e pacientes. Ao integrar soluções generativas de IA aos sistemas de saúde existentes, as instituições médicas podem obter maior eficiência, reduzir custos e, por fim, oferecer cuidados de maior qualidade. A integração da IA com a assistência médica não é apenas um aprimoramento, mas uma mudança fundamental em direção a um atendimento mais inteligente, responsivo e centrado no paciente.

Análise avançada de imagens

A combinação do Amazon Bedrock com armazenamentos de dados, como o Amazon Neptune OpenSearch e o Amazon Service, pode ajudar você a lidar com as complexidades da análise avançada de imagens na área da saúde. As soluções de recuperação de informações podem

aumentar o processo de descoberta de doenças e melhorar a precisão da interpretação, avaliando imagens de diagnóstico e interpretando ultrassonografias. A solução pode integrar os dados de avaliação visual e textual com a revisão manual da avaliação do paciente pelos médicos.

Desafios da industrialização das soluções

Os principais obstáculos a serem enfrentados ao industrializar soluções de IA na área da saúde são a qualidade e a disponibilidade dos dados. Os dados de saúde geralmente existem em formatos fragmentados e inconsistentes. Garantir que os modelos de IA tenham acesso a dados limpos, estruturados e representativos é crucial para manter o desempenho em cenários do mundo real. A escalabilidade da infraestrutura pode se tornar um desafio devido aos ambientes de produção. Esses ambientes precisam lidar com grandes volumes de dados de pacientes em tempo real, fornecendo tempos de resposta rápidos e mantendo a conformidade com os regulamentos de privacidade de dados, como o Health Insurance Portability and Accountability Act (HIPAA). Além disso, com informações médicas emergentes e dados de pacientes que evoluem com o tempo, os modelos de IA precisam ser retreinados e atualizados para permanecerem relevantes e fornecerem recomendações precisas. Por fim, integrar essas soluções de IA aos sistemas de saúde existentes pode ser complexo devido a problemas de interoperabilidade e à necessidade de alinhamento com os fluxos de trabalho clínicos atuais. Essa integração requer mudanças técnicas e operacionais.

Caso de uso: criação de um aplicativo de inteligência médica com dados de pacientes aumentados

A IA generativa pode ajudar a aumentar o atendimento ao paciente e a produtividade da equipe, aprimorando as funções clínicas e administrativas. A análise de imagens baseada em IA, como a interpretação de ultrassonografias, acelera os processos de diagnóstico e melhora a precisão. Ele pode fornecer informações críticas que apoiam intervenções médicas oportunas.

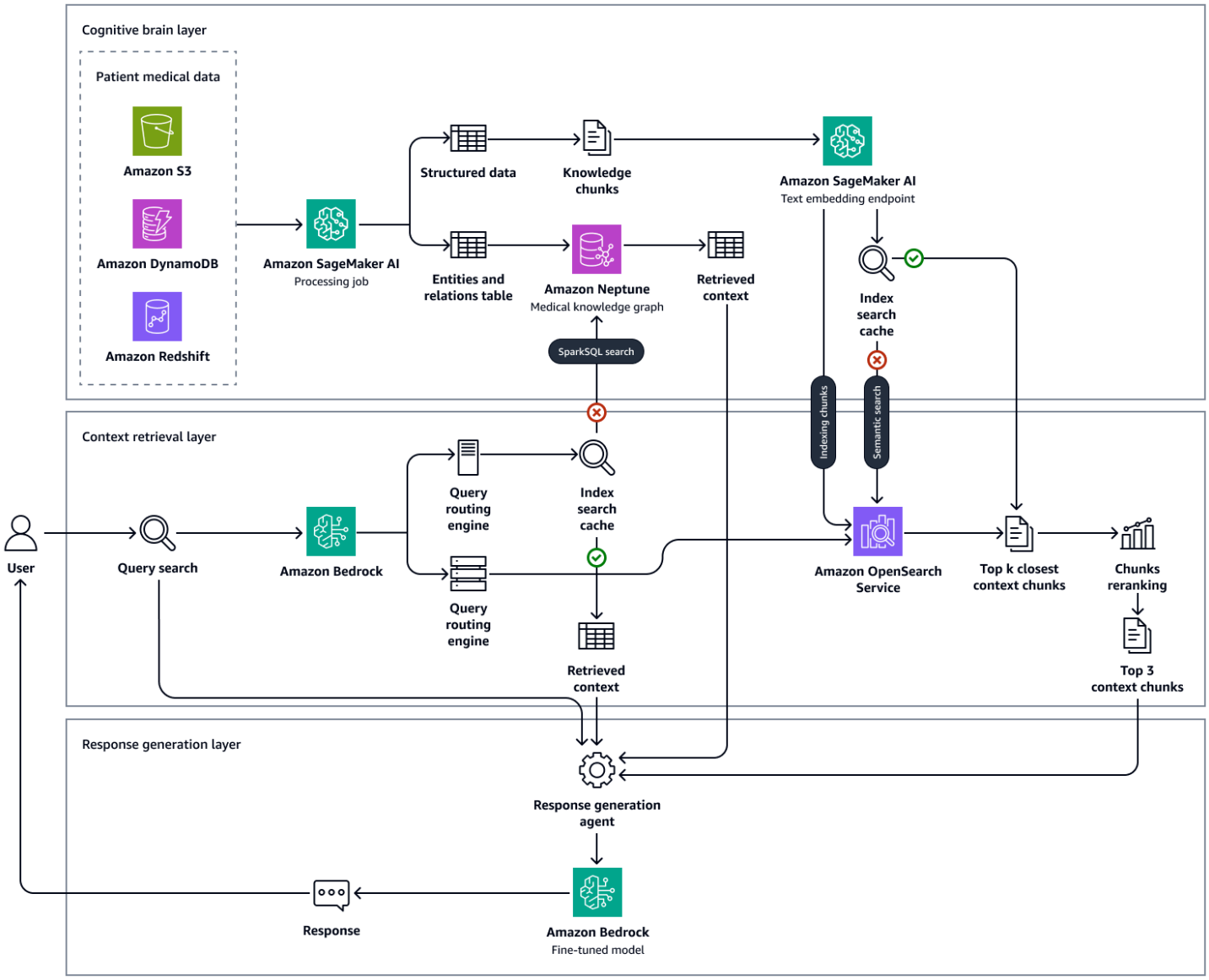
Ao combinar modelos generativos de IA com gráficos de conhecimento, você pode automatizar a organização cronológica dos registros eletrônicos dos pacientes. Isso ajuda você a integrar dados em tempo real de interações médico-paciente, sintomas, diagnósticos, resultados de laboratório e análise de imagens. Isso equipa o médico com dados abrangentes do paciente. Esses dados ajudam o médico a tomar decisões médicas mais precisas e oportunas, melhorando os resultados dos pacientes e a produtividade do profissional de saúde.

Visão geral da solução

A IA pode capacitar médicos e clínicos sintetizando dados de pacientes e conhecimento médico para fornecer informações valiosas. Essa solução Retrieval Augmented Generation (RAG) é um mecanismo de inteligência médica que consome um conjunto abrangente de dados e conhecimentos de pacientes de milhões de interações clínicas. Ele aproveita o poder da IA generativa para criar insights baseados em evidências para melhorar o atendimento ao paciente. Ele foi projetado para aprimorar os fluxos de trabalho clínicos, reduzir erros e melhorar os resultados dos pacientes.

A solução inclui um recurso automatizado de processamento de imagens que é alimentado por. LLMs Esse recurso reduz a quantidade de tempo que a equipe médica deve gastar procurando manualmente imagens de diagnóstico semelhantes e analisando os resultados do diagnóstico.

A imagem a seguir mostra end-to-end-workflow o desta solução. Ele usa o Amazon Neptune, o SageMaker Amazon AI, o OpenSearch Amazon Service e um modelo básico no Amazon Bedrock. Para o agente de recuperação de contexto que interage com o gráfico de conhecimento médico em Neptune, você pode escolher entre um agente Amazon Bedrock e um LangChain agente.



Em nossos experimentos com exemplos de perguntas médicas, observamos que as respostas finais geradas por nossa abordagem usando o gráfico de conhecimento mantido em Neptune OpenSearch, banco de dados vetoriais que abriga a base de conhecimento clínico, e o Amazon LLMs Bedrock foram baseadas na factualidade e são muito mais precisas ao reduzir os falsos positivos e aumentar os verdadeiros positivos. Essa solução pode gerar insights baseados em evidências sobre o estado de saúde do paciente e visa aprimorar os fluxos de trabalho clínicos, reduzir erros e melhorar os resultados dos pacientes.

A criação dessa solução consiste nas seguintes etapas:

- Etapa 1: Descobrindo dados

- Etapa 2: Construindo um gráfico de conhecimento médico
- Etapa 3: Construindo agentes de recuperação de contexto para consultar o gráfico de conhecimento médico
- Etapa 4: Criar uma base de conhecimento de dados descritivos em tempo real
- Etapa 5: Usar LLMs para responder perguntas médicas

Etapa 1: Descobrindo dados

Há muitos conjuntos de dados médicos de código aberto que você pode usar para apoiar o desenvolvimento de uma solução baseada em IA na área de saúde. Um desses conjuntos de dados é o conjunto de dados MIMIC-IV, que é um conjunto de dados de registro eletrônico de saúde (EHR) disponível ao público que é amplamente usado na comunidade de pesquisa em saúde. O MIMIC-IV contém informações clínicas detalhadas, incluindo notas de alta em texto livre dos prontuários dos pacientes. Você pode usar esses registros para experimentar técnicas de soma de texto e extração de entidades. Essas técnicas ajudam a extrair informações médicas (como sintomas do paciente, medicamentos administrados e tratamentos prescritos) de texto não estruturado.

Você também pode usar um conjunto de dados que forneça resumos de alta de pacientes anotados e não identificados, selecionados especificamente para fins de pesquisa. Um conjunto de dados resumidos de alta pode ajudá-lo a fazer experiências com a extração de entidades, permitindo identificar as principais entidades médicas (como condições, procedimentos e medicamentos) a partir do texto. Etapa 2: Construindo um gráfico de conhecimento médiconeste guia, descrevemos como você pode usar os dados estruturados extraídos do MIMIC-IV e dos conjuntos de dados resumidos de alta para criar um gráfico de conhecimento médico. Esse gráfico de conhecimento médico serve como base para sistemas avançados de consulta e suporte à decisão para profissionais de saúde.

Além dos conjuntos de dados baseados em texto, você pode usar conjuntos de dados de imagem. Por exemplo, o conjunto de dados Musculoskeletal Radiographs (MURA), que é um banco de dados abrangente de imagens radiográficas de ossos com várias visualizações. Use esses conjuntos de dados de imagem para experimentar a avaliação diagnóstica por meio de técnicas de decodificação de imagens médicas. Essas técnicas de decodificação são cruciais para o diagnóstico precoce de doenças, como doenças musculoesqueléticas, doenças cardiovasculares e osteoporose. Ao ajustar os modelos básicos de visão e linguagem no conjunto de dados de imagens médicas, você pode detectar anormalidades nas imagens de diagnóstico. Isso ajuda o sistema a fornecer aos médicos informações diagnósticas precoces e precisas. Ao usar conjuntos de dados de imagem e texto, você

pode criar um aplicativo de saúde baseado em IA capaz de processar dados de texto e imagem para melhorar o atendimento ao paciente.

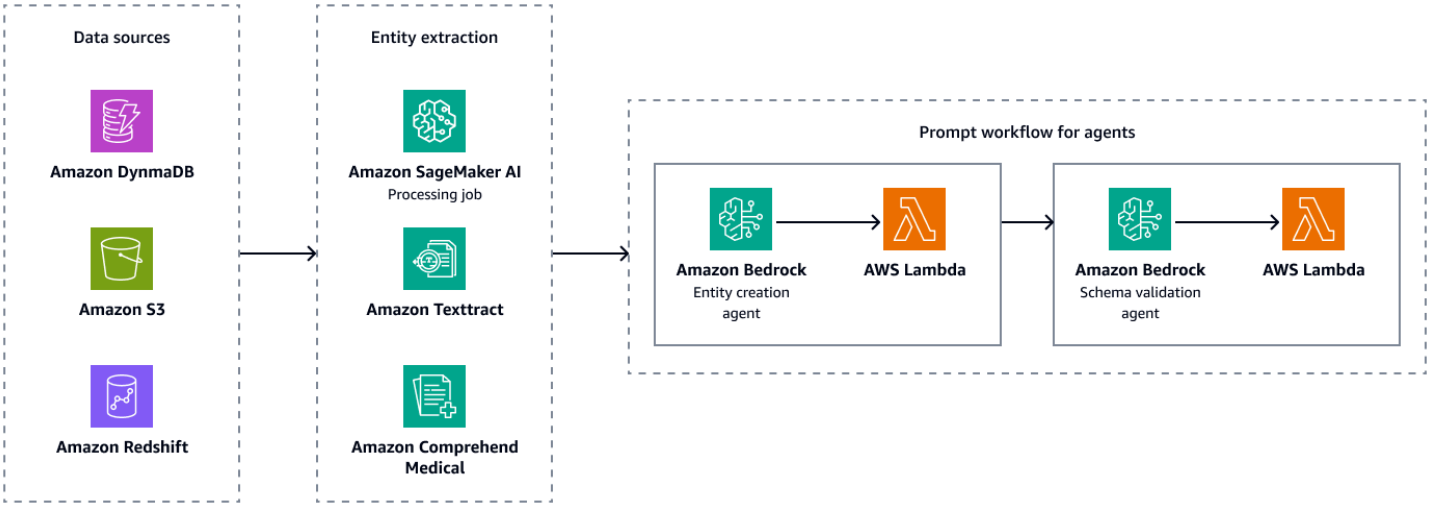
Etapa 2: Construindo um gráfico de conhecimento médico

Para qualquer organização de saúde que queira criar um sistema de apoio à decisão com base em uma enorme base de conhecimento, um dos principais desafios é localizar e extrair as entidades médicas que estão presentes nas notas clínicas, nos periódicos médicos, nos resumos de alta e em outras fontes de dados. Você também precisa capturar as relações temporais, os assuntos e as avaliações de certeza desses registros médicos para usar com eficácia as entidades, atributos e relacionamentos extraídos.

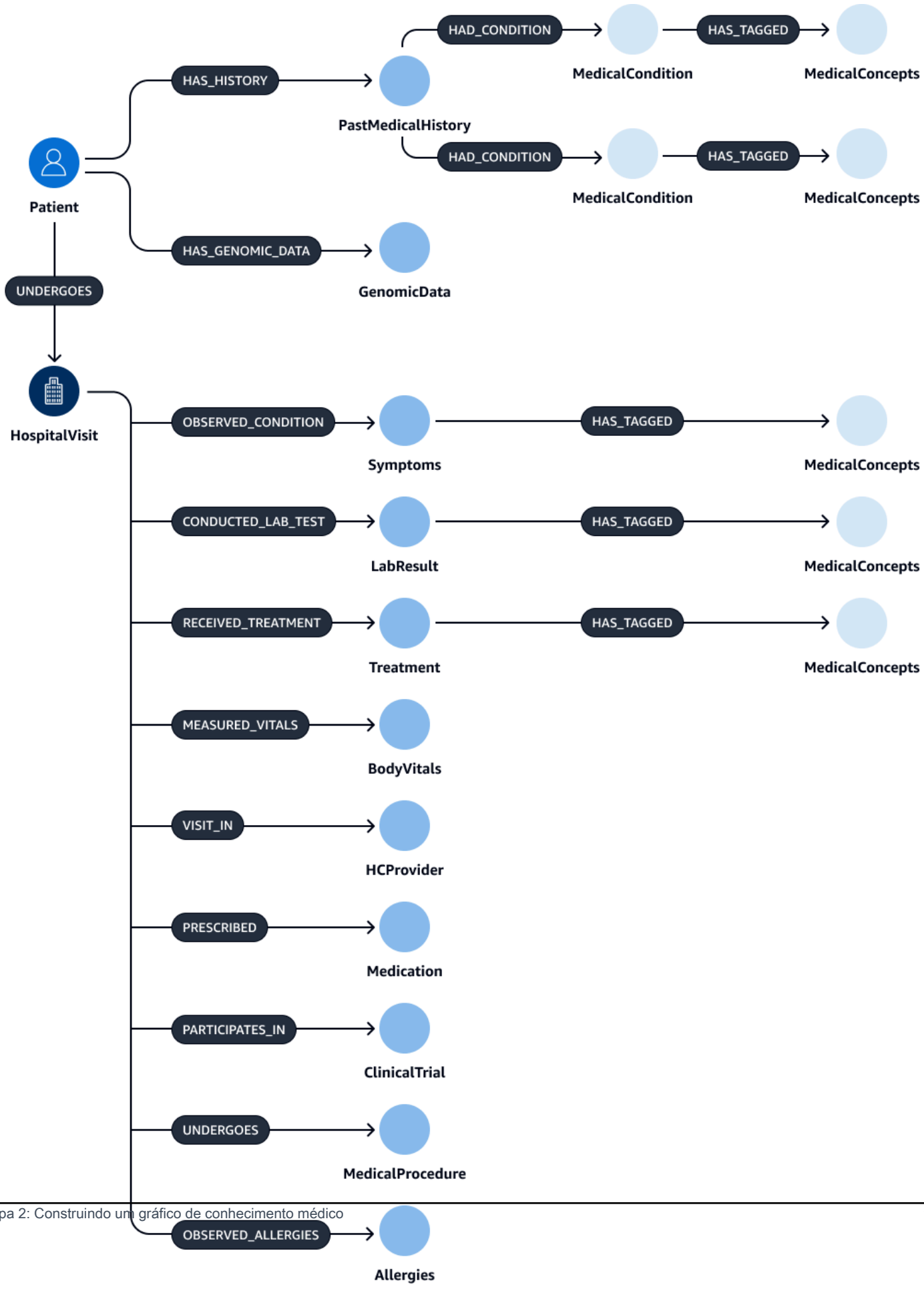
A primeira etapa é extrair conceitos médicos do texto médico não estruturado usando um prompt de poucos passos para um modelo básico, como o Llama 3 no Amazon Bedrock. A solicitação rápida ocorre quando você fornece a um LLM um pequeno número de exemplos que demonstram a tarefa e a saída desejada antes de solicitar que ele execute uma tarefa semelhante. Usando um extrator de entidades médicas baseado em LLM, você pode analisar o texto médico não estruturado e, em seguida, gerar uma representação de dados estruturada das entidades de conhecimento médico. Você também pode armazenar os atributos do paciente para análise e automação posteriores. O processo de extração da entidade inclui as seguintes ações:

- Extraia informações sobre conceitos médicos, como doenças, medicamentos, dispositivos médicos, dosagem, frequência do medicamento, duração do medicamento, sintomas, procedimentos médicos e seus atributos clinicamente relevantes.
- Capture características funcionais, como relações temporais entre entidades extraídas, assuntos e avaliações de certeza.
- Expanda os vocabulários médicos padrão, como os seguintes:
  - Identificadores de conceito (RxCUI) do banco de dados RxNorm
  - Códigos da Classificação Internacional de Doenças, 10ª Revisão, Modificação Clínica (CID-10-CM)
  - Termos do Medical Subject Headings (MeSH)
  - Conceitos da Nomenclatura Sistematizada da Medicina, Termos Clínicos (SNOMED CT)
  - Códigos do Sistema Unificado de Linguagem Médica (UMLS)
- Resuma as notas de alta e obtenha informações médicas a partir das transcrições.

A figura a seguir mostra as etapas de extração de entidades e validação do esquema para criar combinações emparelhadas válidas de entidades, atributos e relacionamentos. Você pode armazenar dados não estruturados, como resumos de alta ou notas de pacientes, no Amazon Simple Storage Service (Amazon S3). Você pode armazenar dados estruturados, como dados de planejamento de recursos corporativos (ERP), registros eletrônicos de pacientes e sistemas de informações laboratoriais, no Amazon Redshift e no Amazon DynamoDB. Você pode criar um agente de criação de entidades Amazon Bedrock. Esse agente pode integrar serviços, como os pipelines de extração de dados de SageMaker IA da Amazon, o Amazon Textract e o Amazon Comprehend Medical, para extrair entidades, relacionamentos e atributos das fontes de dados estruturadas e não estruturadas. Por fim, você usa um agente de validação de esquema do Amazon Bedrock para garantir que as entidades e relacionamentos extraídos estejam em conformidade com o esquema gráfico predefinido e mantenham a integridade das conexões de borda do nó e das propriedades associadas.



Depois de extrair e validar as entidades, relações e atributos, você pode vinculá-los para criar um subject-object-predicate trio. Você ingere esses dados em um banco de dados gráfico do Amazon Neptune, conforme mostrado na figura a seguir. Os bancos de dados gráficos são otimizados para armazenar e consultar as relações entre os itens de dados.


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

Você pode criar um gráfico de conhecimento abrangente com esses dados. Um gráfico de conhecimento ajuda você a organizar e consultar todos os tipos de informações conectadas. Por exemplo, você pode criar um gráfico de conhecimento que tenha os seguintes nós principais: `HospitalVisit` `PastMedicalHistorySymptoms`,`Medication`,`MedicalProcedures`,, `Treatment` e.

As tabelas a seguir listam as entidades e seus atributos que você pode extrair das notas de descarga.

Entidade	Atributos
`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`

Entidade	Atributos
`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`

A tabela a seguir lista os relacionamentos que as entidades podem ter e seus atributos correspondentes. Por exemplo, a `Patient` entidade pode se conectar à `HospitalVisit` entidade com o `[UNDERGOES]` relacionamento. O atributo desse relacionamento é`VisitDate`.

Entidade sujeita	Relacionamento	Entidade de objeto	Atributos
`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`

Entidade sujeita	Relacionamento	Entidade de objeto	Atributos
HospitalVisit	[PRESCRIBED]	Medication	Duration, Dosage, Adherence , VisitDate
Patient	[HAS_HISTORY]	PastMedicalHistory	Nenhum
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	Nenhum
LabResult	[HAS_TAGGED]	MedicalConcepts	Nenhum
Treatment	[HAS_TAGGED]	MedicalConcepts	Nenhum
Symptoms	[HAS_TAGGED]	MedicalConcepts	Nenhum

Etapa 3: Construindo agentes de recuperação de contexto para consultar o gráfico de conhecimento médico

Depois de criar o banco de dados de gráficos médicos, a próxima etapa é criar agentes para interação gráfica. Os agentes recuperam o contexto correto e necessário para a consulta inserida por um médico ou clínico. Há várias opções para configurar esses agentes que recuperam o contexto do gráfico de conhecimento:

- Agentes Amazon Bedrock
- LangChain agentes

Agentes Amazon Bedrock para interação gráfica

Os agentes do Amazon Bedrock funcionam perfeitamente com os bancos de dados gráficos do Amazon Neptune. Você pode realizar interações avançadas por meio dos grupos de ação do Amazon Bedrock. O grupo de ação inicia o processo chamando uma AWS Lambda função, que executa consultas OpenCypher do Neptune.

Para consultar um gráfico de conhecimento, você pode usar duas abordagens distintas: execução direta de consultas ou consultas com incorporação de contexto. Essas abordagens podem ser aplicadas de forma independente ou combinada, dependendo do seu caso de uso específico e dos critérios de classificação. Ao combinar as duas abordagens, você pode fornecer um contexto mais abrangente ao LLM, o que pode melhorar os resultados. A seguir estão as duas abordagens de execução de consultas:

- Execução direta de consultas do Cypher sem incorporações — A função Lambda executa consultas diretamente no Neptune sem nenhuma pesquisa baseada em incorporações. Veja a seguir um exemplo dessa abordagem:

```
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
```

- Execução direta da consulta Cypher usando a pesquisa incorporada — A função Lambda usa a pesquisa incorporada para aprimorar os resultados da consulta. Essa abordagem aprimora a execução da consulta incorporando incorporações, que são representações vetoriais densas de dados. As incorporações são particularmente úteis quando a consulta exige semelhança

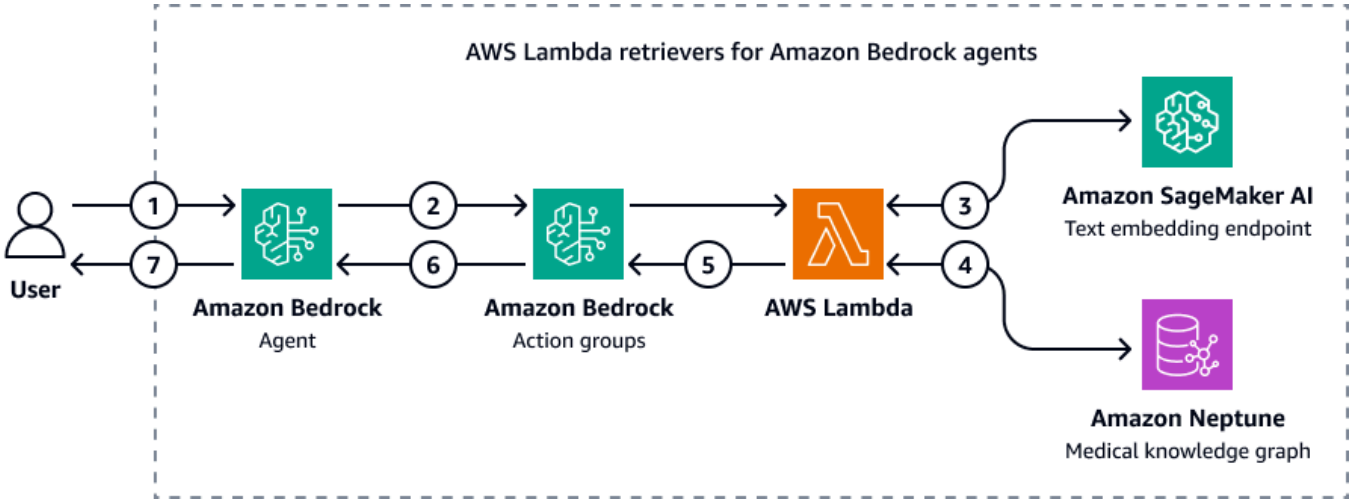
semântica ou compreensão mais ampla além das correspondências exatas. Você pode usar modelos pré-treinados ou personalizados para gerar incorporações para cada condição médica. Veja a seguir um exemplo dessa abordagem:

```
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
```

Neste exemplo, a `search_embedding("Acute Diabetes")` função recupera condições semanticamente próximas de “Diabetes Agudo”. Isso ajuda a consulta a encontrar também pacientes com doenças como pré-diabetes ou síndrome metabólica.

A imagem a seguir mostra como os agentes do Amazon Bedrock interagem com o Amazon Neptune para realizar uma consulta Cypher de um gráfico de conhecimento médico.



O diagrama mostra o seguinte fluxo de trabalho:

1. O usuário envia uma pergunta para o agente Amazon Bedrock.
2. O agente do Amazon Bedrock passa as variáveis de pergunta e filtro de entrada para os grupos de ação do Amazon Bedrock. Esses grupos de ação contêm uma AWS Lambda função que interage com o endpoint de incorporação de texto do Amazon SageMaker AI e o gráfico de conhecimento médico do Amazon Neptune.

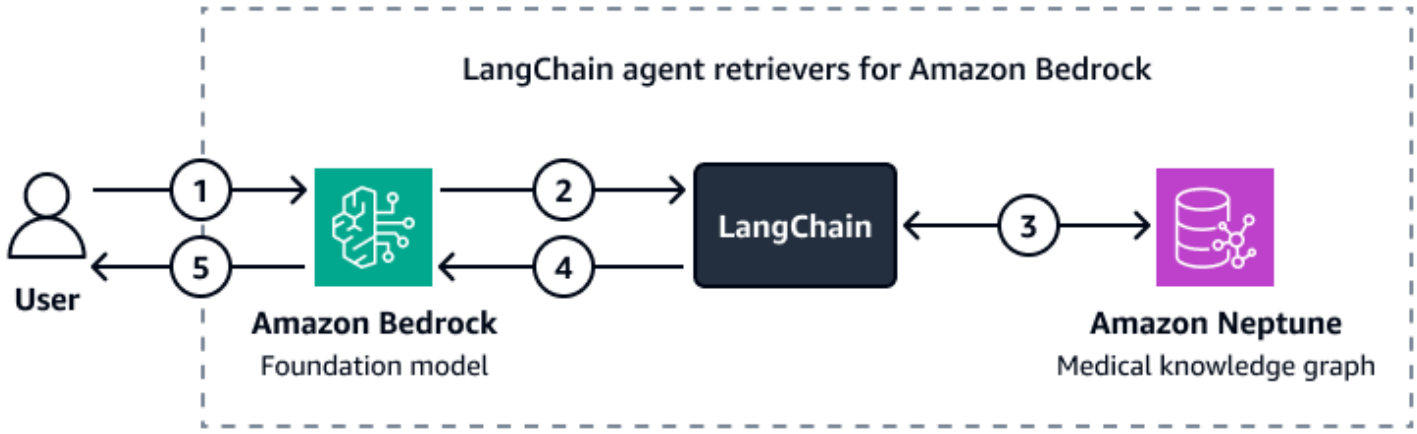
3. A função Lambda se integra ao endpoint de incorporação de texto de SageMaker IA para realizar uma pesquisa semântica na consulta OpenCypher. Ele converte a consulta de linguagem natural em uma consulta OpenCypher usando subjacente LangChain agentes.
4. A função Lambda consulta o gráfico de conhecimento médico de Neptune para obter o conjunto de dados correto e recebe a saída do gráfico de conhecimento médico de Neptune.
5. A função Lambda retorna os resultados do Neptune para os grupos de ação do Amazon Bedrock.
6. Os grupos de ação do Amazon Bedrock enviam o contexto recuperado para o agente do Amazon Bedrock.
7. O agente Amazon Bedrock gera a resposta usando a consulta original do usuário e o contexto recuperado do gráfico de conhecimento.

LangChain agentes para interação gráfica

Você pode integrar LangChain com o Neptune para permitir consultas e recuperações baseadas em gráficos. Essa abordagem pode aprimorar os fluxos de trabalho orientados por IA usando os recursos de banco de dados gráficos do Neptune. O costume LangChain o retriever atua como intermediário. O modelo básico no Amazon Bedrock pode interagir com o Neptune usando consultas diretas do Cypher e algoritmos gráficos mais complexos.

Você pode usar o recuperador personalizado para refinar como LangChain o agente interage com os algoritmos do gráfico de Neptune. Por exemplo, você pode usar a solicitação de algumas etapas, que ajuda a personalizar as respostas do modelo básico com base em padrões ou exemplos específicos. Você também pode aplicar filtros identificados pelo LLM para refinar o contexto e melhorar a precisão das respostas. Isso pode melhorar a eficiência e a precisão do processo geral de recuperação ao interagir com dados gráficos complexos.

A imagem a seguir mostra como um personalizado LangChain O agente orquestra a interação entre um modelo da Amazon Bedrock Foundation e um gráfico de conhecimento médico do Amazon Neptune.



O diagrama mostra o seguinte fluxo de trabalho:

1. Um usuário envia uma pergunta ao Amazon Bedrock e ao LangChain agente.
2. O modelo da Amazon Bedrock Foundation usa o esquema Neptune, que é fornecido pelo LangChain agente, para gerar uma consulta para a pergunta do usuário.
3. A ferramenta LangChain O agente executa a consulta no gráfico de conhecimento médico do Amazon Neptune.
4. A ferramenta LangChain O agente envia o contexto recuperado para o modelo da fundação Amazon Bedrock.
5. O modelo da Amazon Bedrock Foundation usa o contexto recuperado para gerar uma resposta à pergunta do usuário.

Etapa 4: Criar uma base de conhecimento de dados descritivos em tempo real

Em seguida, você cria uma base de conhecimento de notas descritivas de interação médico-paciente em tempo real, avaliações de diagnóstico por imagem e relatórios de análises laboratoriais. Essa base de conhecimento é um banco de dados vetoriais. Ao usar um banco de dados vetorial, que pode armazenar conhecimento médico descritivo de forma indexada e vetorizada, os profissionais de saúde podem consultar e acessar com eficiência informações relevantes de um vasto repositório. Essas representações vetorizadas ajudam você a recuperar dados semanticamente semelhantes. Os profissionais de saúde podem navegar rapidamente pelas notas clínicas, imagens médicas e resultados laboratoriais. Isso acelera a tomada de decisão informada, oferecendo acesso

instantâneo a informações contextualmente relevantes, aumentando a precisão e a velocidade dos diagnósticos e planos de tratamento.

Usando uma base de conhecimento médico de OpenSearch serviços

O Amazon OpenSearch Service pode gerenciar grandes volumes de dados médicos de alta dimensão. É um serviço gerenciado que facilita a pesquisa de alto desempenho e a análise em tempo real. É adequado como banco de dados vetoriais para aplicações RAG. OpenSearch O serviço atua como uma ferramenta de back-end para gerenciar grandes quantidades de dados não estruturados ou semiestruturados, como registros médicos, artigos de pesquisa e notas clínicas. Seus recursos avançados de pesquisa semântica ajudam você a recuperar informações contextualmente relevantes. Isso o torna particularmente útil em aplicações como sistemas de apoio à decisão clínica, ferramentas de resolução de consultas de pacientes e sistemas de gerenciamento de conhecimento em saúde. Por exemplo, um médico pode encontrar rapidamente dados relevantes de pacientes ou estudos de pesquisa que correspondam a sintomas ou protocolos de tratamento específicos. Isso ajuda os médicos a tomar decisões baseadas nas informações mais up-to-date relevantes.

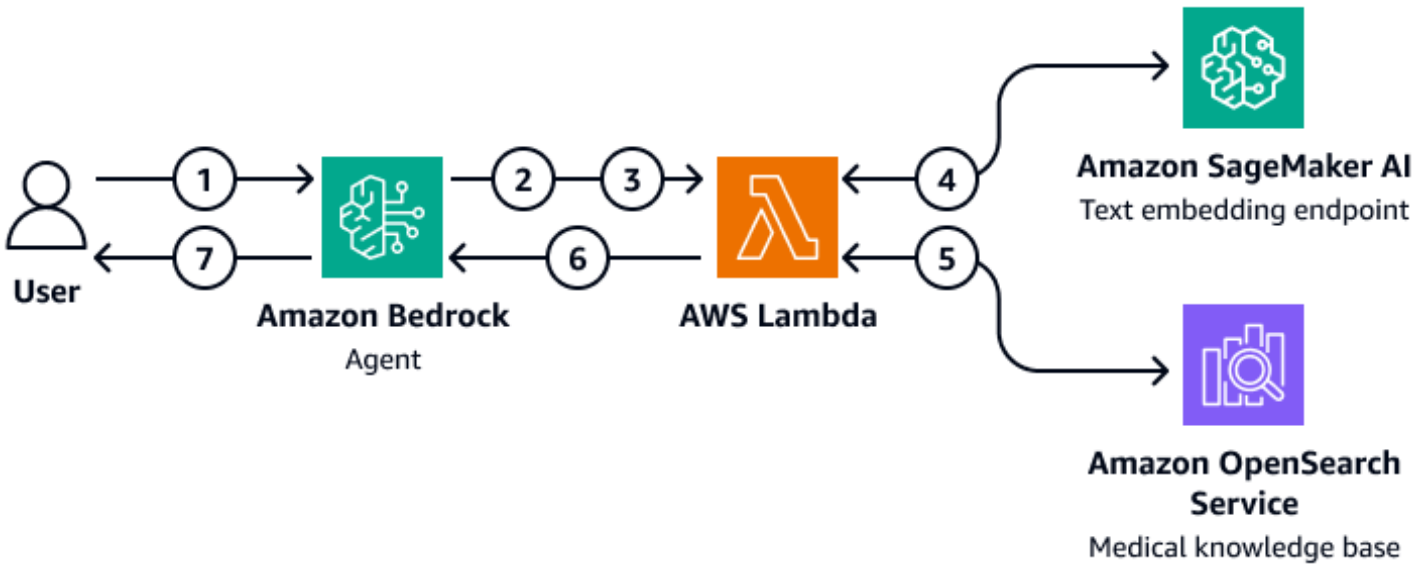
OpenSearch O serviço pode escalar e lidar com indexação e consulta de dados em tempo real. Isso o torna ideal para ambientes dinâmicos de assistência médica, onde o acesso oportuno a informações precisas é fundamental. Além disso, possui recursos de pesquisa multimodais que são ideais para pesquisas que exigem várias entradas, como imagens médicas e anotações médicas. Ao implementar o OpenSearch Service for Healthcare Applications, é fundamental que você defina campos e mapeamentos precisos para otimizar a indexação e a recuperação de dados. Os campos representam os dados individuais, como registros de pacientes, históricos médicos e códigos de diagnóstico. Os mapeamentos definem como esses campos são armazenados (na forma incorporada ou na forma original) e consultados. Para aplicativos de saúde, é essencial estabelecer mapeamentos que acomodem vários tipos de dados, incluindo dados estruturados (como resultados numéricos de testes), dados semiestruturados (como anotações de pacientes) e dados não estruturados (como imagens médicas)

No OpenSearch Service, você pode realizar consultas de pesquisa neural em texto completo por meio de instruções selecionadas para pesquisar registros médicos, notas clínicas ou trabalhos de pesquisa para encontrar rapidamente informações relevantes sobre sintomas, tratamentos ou históricos de pacientes específicos. As consultas de pesquisa neural lidam automaticamente com a incorporação do prompt de entrada e das imagens usando modelos de rede neural integrados. Isso ajuda a entender e capturar as relações semânticas mais profundas em dados multimodais,

oferecendo resultados de pesquisa mais precisos e sensíveis ao contexto em comparação com outros algoritmos de consulta de pesquisa, como a pesquisa k-Nearest Neighbor (k-NN).

Criando uma arquitetura RAG

Você pode implantar uma solução RAG personalizada que usa agentes do Amazon Bedrock para consultar uma base de conhecimento médico no OpenSearch Service. Para fazer isso, você cria uma AWS Lambda função que pode interagir e consultar o OpenSearch Serviço. A função Lambda incorpora a pergunta de entrada do usuário acessando um endpoint de incorporação de texto de SageMaker IA. O agente Amazon Bedrock passa parâmetros de consulta adicionais como entradas para a função Lambda. A função consulta a base de conhecimento médico no OpenSearch Service, que retorna o conteúdo médico relevante. Depois de configurar a função Lambda, adicione-a como um grupo de ação dentro do agente Amazon Bedrock. O agente Amazon Bedrock pega a entrada do usuário, identifica as variáveis necessárias, passa as variáveis e a pergunta para a função Lambda e, em seguida, inicia a função. A função retorna um contexto que ajuda o modelo básico a fornecer uma resposta mais precisa à pergunta do usuário.

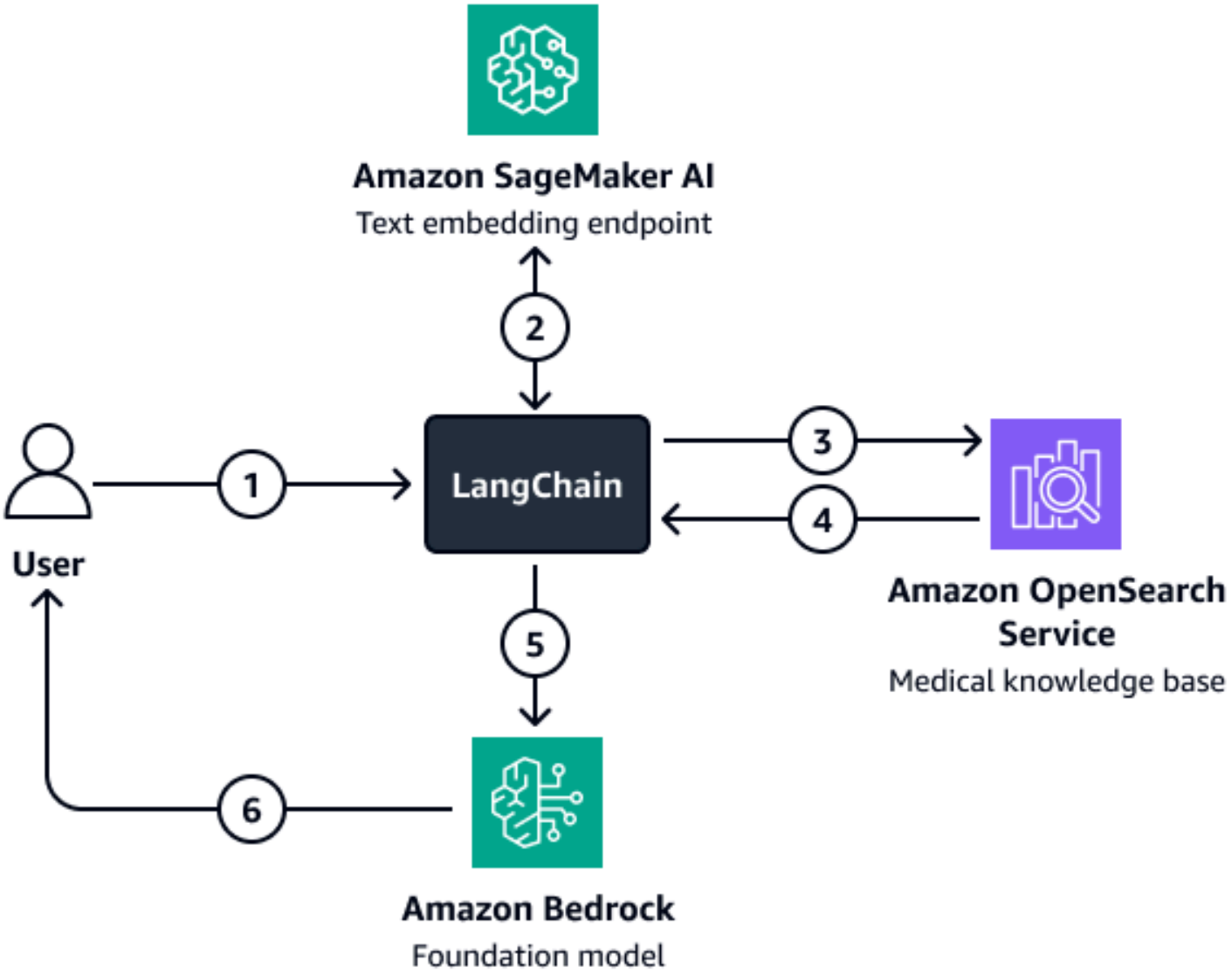


O diagrama mostra o seguinte fluxo de trabalho:

1. Um usuário envia uma pergunta para o agente Amazon Bedrock.
2. O agente Amazon Bedrock seleciona qual grupo de ação iniciar.
3. O agente Amazon Bedrock inicia uma AWS Lambda função e passa parâmetros para ela.
4. A função Lambda inicia o modelo de incorporação de texto do Amazon SageMaker AI para incorporar a pergunta do usuário.

5. A função Lambda passa o texto incorporado e os parâmetros e filtros adicionais para o Amazon OpenSearch Service. O Amazon OpenSearch Service consulta a base de conhecimento médico e retorna os resultados para a função Lambda.
6. A função Lambda repassa os resultados para o agente Amazon Bedrock.
7. O modelo básico no agente Amazon Bedrock gera uma resposta com base nos resultados e retorna a resposta ao usuário.

Para situações em que uma filtragem mais complexa está envolvida, você pode usar um personalizado LangChain retriever. Crie esse recuperador configurando um cliente de pesquisa vetorial de OpenSearch serviço que é carregado diretamente no LangChain. Essa arquitetura permite que você passe mais variáveis para criar os parâmetros do filtro. Depois que o recuperador estiver configurado, use o modelo Amazon Bedrock e o recuperador para configurar uma cadeia de respostas a perguntas de recuperação. Essa cadeia orquestra a interação entre o modelo e o recuperador passando a entrada do usuário e os filtros potenciais para o recuperador. O recuperador retorna o contexto relevante que ajuda o modelo básico a responder à pergunta do usuário.



O diagrama mostra o seguinte fluxo de trabalho:

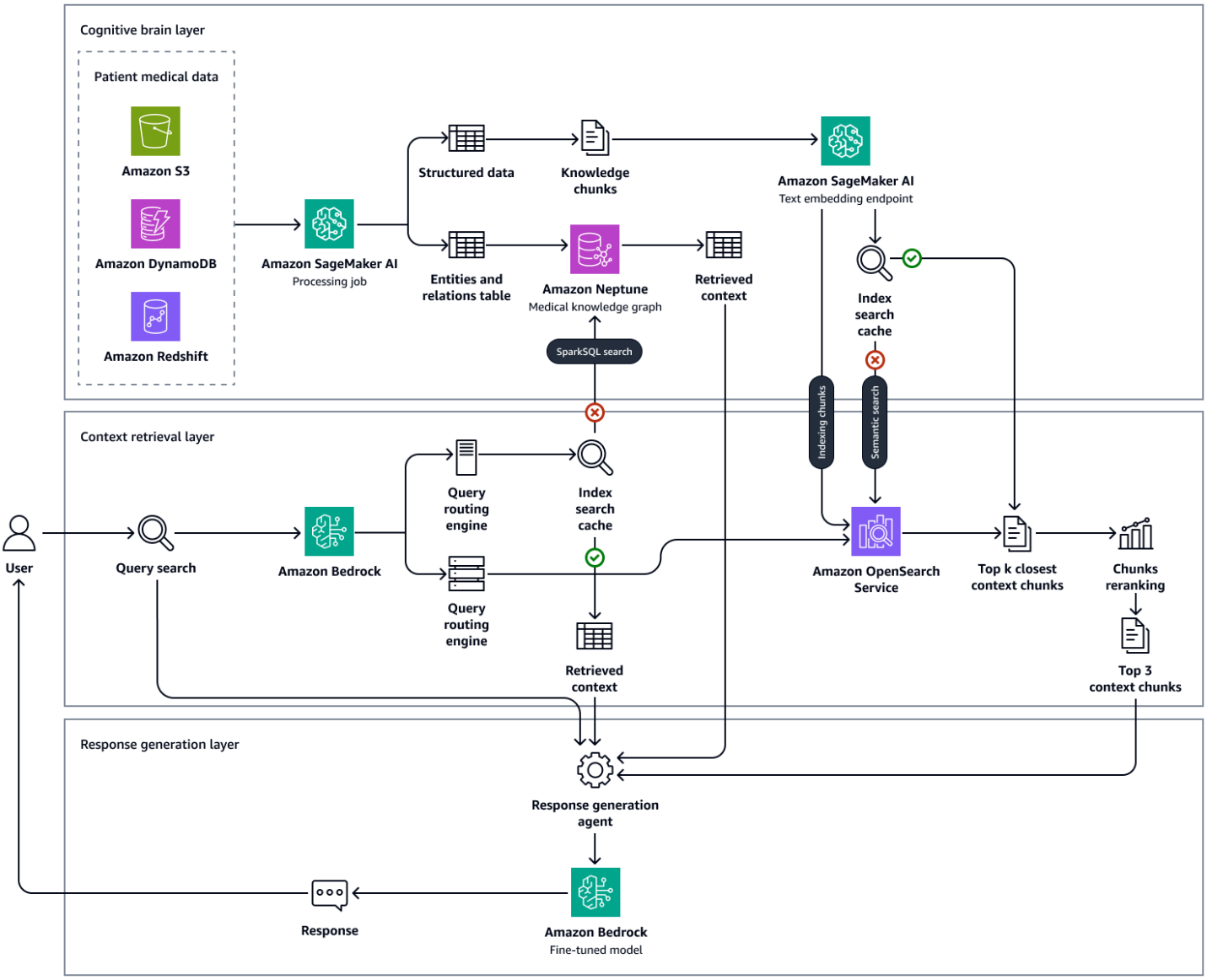
1. Um usuário envia uma pergunta para o LangChain agente retriever.
2. A ferramenta LangChain O agente retriever envia a pergunta para o endpoint de incorporação de texto do Amazon SageMaker AI para incorporar a pergunta.
3. A ferramenta LangChain o agente retriever passa o texto incorporado para o Amazon OpenSearch Service.
4. O Amazon OpenSearch Service retorna os documentos recuperados para o LangChain agente retriever.
5. A ferramenta LangChain O agente retriever passa a pergunta do usuário e o contexto recuperado para o modelo da Amazon Bedrock Foundation.

6. O modelo básico gera uma resposta e a envia ao usuário.

Etapa 5: Usar LLMs para responder perguntas médicas

As etapas anteriores ajudam você a criar um aplicativo de inteligência médica que pode buscar os registros médicos de um paciente e resumir medicamentos relevantes e possíveis diagnósticos. Agora, você constrói a camada de geração. Essa camada usa os recursos generativos de um LLM no Amazon Bedrock, como o Llama 3, para aumentar a saída do aplicativo.

Quando um médico insere uma consulta, a camada de recuperação de contexto do aplicativo executa o processo de recuperação a partir do gráfico de conhecimento e retorna os principais registros relacionados ao histórico, dados demográficos, sintomas, diagnóstico e resultados do paciente. Do banco de dados vetoriais, ele também recupera notas descritivas de interação médico-paciente em tempo real, insights de avaliação de imagens diagnósticas, resumos de relatórios de análises laboratoriais e insights de um grande corpus de pesquisas médicas e livros acadêmicos. Esses principais resultados recuperados, a consulta do médico e os prompts (que são personalizados para selecionar respostas com base na natureza da consulta) são então passados para o modelo básico no Amazon Bedrock. Essa é a camada de geração de resposta. O LLM usa o contexto recuperado para gerar uma resposta à consulta do médico. A figura a seguir mostra o end-to-end fluxo de trabalho das etapas dessa solução.



Você pode usar um modelo básico pré-treinado no Amazon Bedrock, como o Llama 3, para uma variedade de casos de uso que o aplicativo de inteligência médica precisa lidar. O LLM mais eficaz para uma determinada tarefa varia de acordo com o caso de uso. Por exemplo, um modelo pré-treinado pode ser suficiente para resumir conversas médico-paciente, pesquisar medicamentos e históricos de pacientes e recuperar insights de conjuntos de dados médicos internos e corpos de conhecimento científico. No entanto, um LLM aperfeiçoado pode ser necessário para outros casos de uso complexos, como avaliações laboratoriais em tempo real, recomendações de procedimentos médicos e previsões de resultados de pacientes. Você pode ajustar um LLM treinando-o em conjuntos de dados do domínio médico. Requisitos específicos ou complexos de saúde e ciências biológicas impulsionam o desenvolvimento desses modelos aperfeiçoados.

Para obter mais informações sobre como ajustar um LLM ou escolher um LLM existente que tenha sido treinado em dados do domínio médico, consulte Usando grandes modelos de linguagem para casos de uso de saúde e ciências biológicas.

Alinhamento com o AWS Well-Architected Framework

A solução se alinha com todos os seis pilares do AWS Well-Architected Framework da seguinte forma:

- Excelência operacional — A arquitetura é dissociada para monitoramento e atualizações eficientes. Agentes do Amazon Bedrock AWS Lambda ajudam você a implantar e reverter ferramentas rapidamente.
- Segurança — Essa solução foi projetada para estar em conformidade com os regulamentos de saúde, como o HIPAA. Você também pode implementar criptografia, controle de acesso refinado e grades de proteção Amazon Bedrock para ajudar a proteger os dados dos pacientes.
- Confiabilidade — serviços AWS gerenciados, como o Amazon OpenSearch Service e o Amazon Bedrock, fornecem a infraestrutura para a interação contínua do modelo.
- Eficiência de desempenho — A solução RAG recupera dados relevantes rapidamente usando pesquisa semântica otimizada e consultas Cypher, enquanto um roteador agente identifica as rotas ideais para as consultas do usuário.
- Otimização de custos — O pay-per-token modelo na arquitetura Amazon Bedrock e RAG reduz os custos de inferência e pré-treinamento.
- Sustentabilidade — Usar infraestrutura e pay-per-token computação sem servidor minimiza o uso de recursos e aumenta a sustentabilidade.

Caso de uso: previsão de resultados de pacientes e taxas de readmissão

A análise preditiva baseada em IA oferece mais benefícios ao prever os resultados dos pacientes e permitir planos de tratamento personalizados. Isso pode melhorar a satisfação do paciente e os resultados de saúde. Ao integrar esses recursos de IA com o Amazon Bedrock e outras tecnologias, os profissionais de saúde podem obter ganhos significativos de produtividade, reduzir custos e elevar a qualidade geral do atendimento ao paciente.

Você pode armazenar dados médicos, como históricos de pacientes, notas clínicas, medicamentos e tratamentos, em um gráfico de conhecimento. Ao combinar a profunda compreensão contextual LLMs com os dados estruturados e temporais em um gráfico de conhecimento médico, os profissionais de saúde podem obter informações adicionais sobre os padrões individuais dos pacientes. Usando a análise preditiva, você pode identificar precocemente possíveis complicações de não adesão ou tratamento e gerar pontuações personalizadas de propensão à readmissão.

Essa solução ajuda você a prever a probabilidade de uma readmissão. Essas previsões podem melhorar os resultados dos pacientes e reduzir os custos de saúde. Essa solução também pode ajudar os médicos e administradores do hospital a concentrarem sua atenção nos pacientes com maior risco de readmissão. Também os ajuda a iniciar intervenções proativas com esses pacientes por meio de alertas, autoatendimento e ações baseadas em dados.

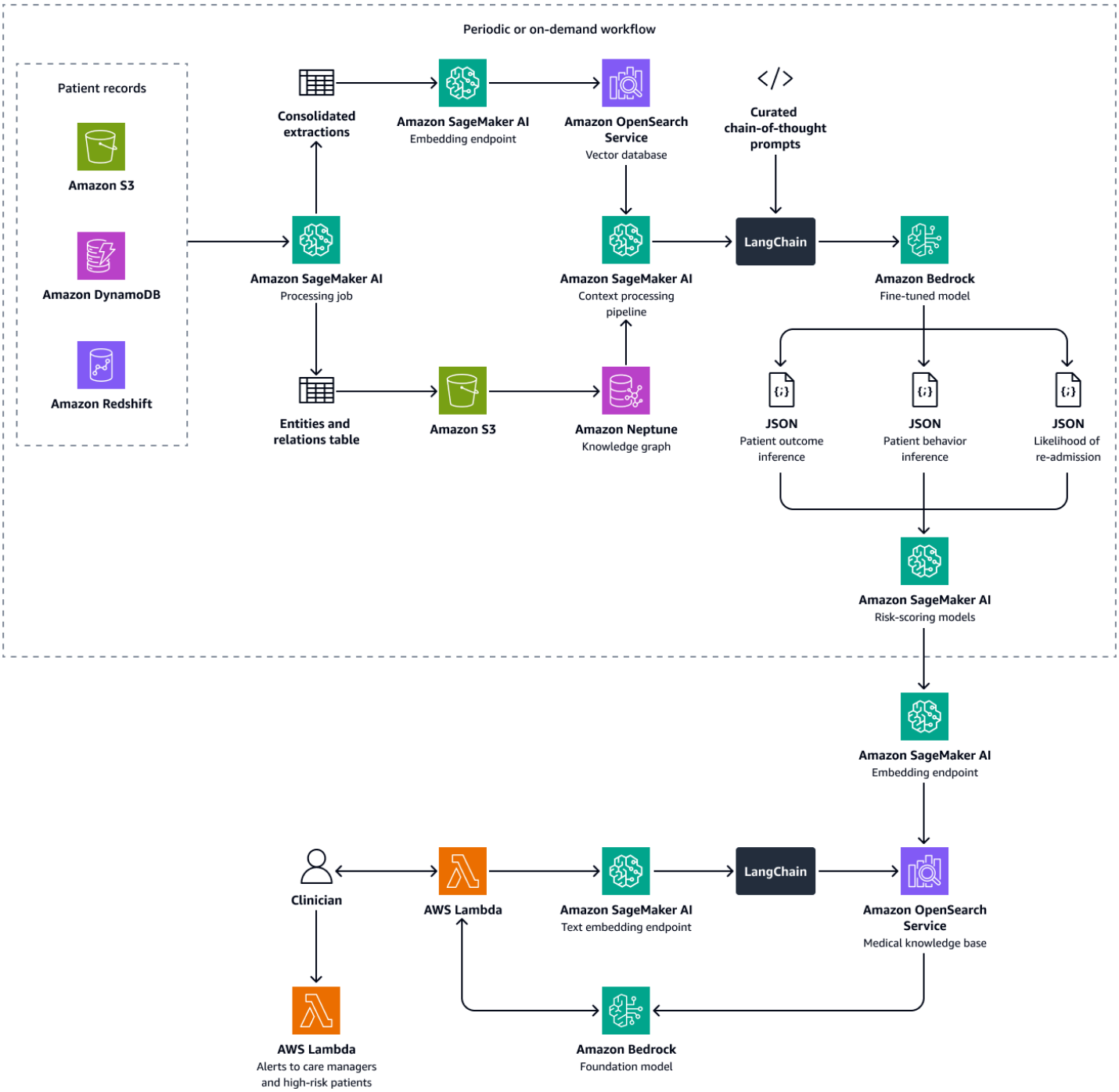
Visão geral da solução

Essa solução usa uma estrutura de geração aumentada de recuperação múltipla (RAG) para analisar os dados do paciente. Ele prevê a probabilidade de readmissão hospitalar para pacientes individuais e ajuda a calcular uma pontuação de propensão à readmissão em nível hospitalar. Essa solução integra os seguintes recursos:

- Gráfico de conhecimento — armazena dados estruturados e cronológicos do paciente, como consultas hospitalares, readmissões anteriores, sintomas, resultados laboratoriais, tratamentos prescritos e histórico de adesão à medicação
- Banco de dados vetorial — armazena dados clínicos não estruturados, como resumos de alta, anotações médicas e registros de consultas perdidas ou efeitos colaterais de medicamentos relatados

- LLM ajustado — Consome dados estruturados do gráfico de conhecimento e dados não estruturados do banco de dados vetoriais para gerar inferências sobre o comportamento do paciente, a adesão ao tratamento e a probabilidade de readmissão

Os modelos de pontuação de risco quantificam as inferências do LLM em pontuações numéricas. Você pode agregar as pontuações em uma pontuação de propensão à readmissão em nível hospitalar. Essa pontuação define a exposição ao risco de cada paciente e você pode calculá-la periodicamente ou conforme necessário. Todas as inferências e pontuações de risco são indexadas e armazenadas no Amazon OpenSearch Service para que os gerentes de atendimento e médicos possam recuperá-las. Ao integrar um agente de IA conversacional a esse banco de dados vetorial, médicos e gerentes de atendimento podem extrair facilmente insights em nível de paciente individual, de toda a instalação ou por especialidade médica. Você também pode configurar alertas automatizados com base nas pontuações de risco, o que incentiva intervenções proativas.



A criação dessa solução consiste nas seguintes etapas:

- Etapa 1: Prever os resultados dos pacientes usando um gráfico de conhecimento médico
- Etapa 2: Prever o comportamento do paciente em relação aos medicamentos ou tratamentos prescritos
- Etapa 3: Prever a probabilidade de readmissão do paciente

Etapa 1: Prever os resultados dos pacientes usando um gráfico de conhecimento médico

No Amazon Neptune, você pode usar um gráfico de conhecimento para armazenar conhecimento temporal sobre visitas de pacientes e resultados ao longo do tempo. A maneira mais eficaz de criar e armazenar um gráfico de conhecimento é usar um modelo gráfico e um banco de dados gráfico. Os bancos de dados gráficos são criados especificamente para armazenar e navegar por relacionamentos. Os bancos de dados gráficos facilitam a modelagem e o gerenciamento de dados altamente conectados e têm esquemas flexíveis.

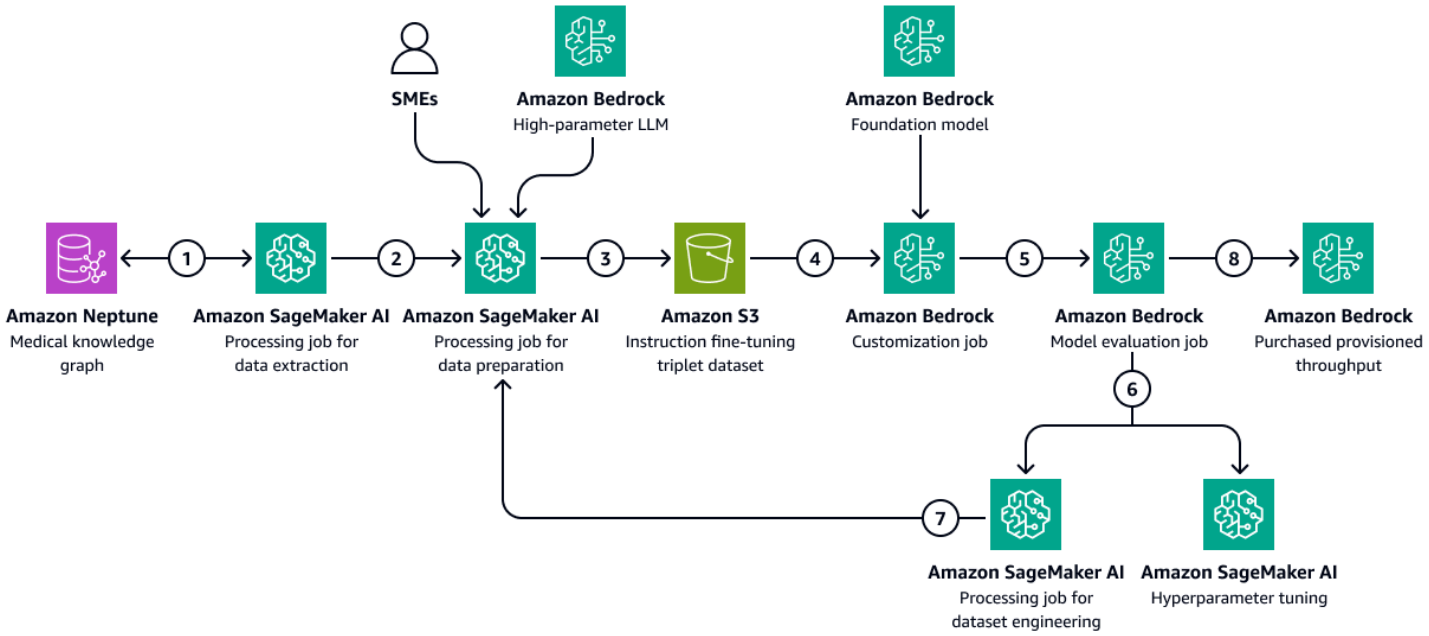
O gráfico de conhecimento ajuda você a realizar análises de séries temporais. A seguir estão os principais elementos do banco de dados gráfico que são usados para predição temporal dos resultados dos pacientes:

- Dados históricos — Diagnósticos anteriores, medicamentos contínuos, medicamentos usados anteriormente e resultados laboratoriais para o paciente
- Visitas ao paciente (cronológicas) — Datas das visitas, sintomas, alergias observadas, notas clínicas, diagnósticos, procedimentos, tratamentos, medicamentos prescritos e resultados laboratoriais
- Sintomas e parâmetros clínicos — Informações clínicas e baseadas em sintomas, incluindo gravidade, padrões de progressão e a resposta do paciente ao medicamento

Você pode usar os insights do gráfico de conhecimento médico para ajustar um LLM no Amazon Bedrock, como o Llama 3. Você ajusta o LLM com dados sequenciais do paciente sobre a resposta do paciente a um conjunto de medicamentos ou tratamentos ao longo do tempo. Use um conjunto de dados rotulado que classifique um conjunto de medicamentos ou tratamentos e dados de interação paciente-clínica em categorias predefinidas que indicam o estado de saúde de um paciente. Exemplos dessas categorias são deterioração da saúde, melhora ou progresso estável. Quando o médico insere um novo contexto sobre o paciente e seus sintomas, o LLM ajustado pode usar os padrões do conjunto de dados de treinamento para prever o resultado potencial do paciente.

A imagem a seguir mostra as etapas sequenciais envolvidas no ajuste fino de um LLM no Amazon Bedrock usando um conjunto de dados de treinamento específico da área de saúde. Esses dados podem incluir condições médicas do paciente e respostas aos tratamentos ao longo do tempo.

Esse conjunto de dados de treinamento ajudaria o modelo a fazer previsões generalizadas sobre os resultados dos pacientes.



O diagrama mostra o seguinte fluxo de trabalho:

1. O trabalho de extração de dados da Amazon SageMaker AI consulta o gráfico de conhecimento para recuperar dados cronológicos sobre as respostas de diferentes pacientes a um conjunto de medicamentos ou tratamentos ao longo do tempo.
2. O trabalho de preparação de dados de SageMaker IA integra um Amazon Bedrock LLM e contribuições de especialistas no assunto (). SMEs O trabalho classifica os dados recuperados do gráfico de conhecimento em categorias predefinidas (como deterioração da saúde, melhora ou progresso estável) que indicam o estado de saúde de cada paciente.
3. O trabalho cria um conjunto de dados de ajuste fino que inclui as informações extraídas do gráfico de conhecimento, as chain-of-thought solicitações e a categoria de resultados do paciente. Ele carrega esse conjunto de dados de treinamento em um bucket do Amazon S3.
4. Um trabalho de personalização do Amazon Bedrock usa esse conjunto de dados de treinamento para ajustar um LLM.
5. O trabalho de personalização do Amazon Bedrock integra o modelo básico preferido do Amazon Bedrock ao ambiente de treinamento. Ele inicia o trabalho de ajuste fino e usa o conjunto de dados de treinamento e os hiperparâmetros de treinamento que você configura.
6. Um trabalho de avaliação do Amazon Bedrock avalia o modelo ajustado usando uma estrutura de avaliação de modelo pré-projetada.

7. Se o modelo precisar ser aprimorado, o trabalho de treinamento será executado novamente com mais dados após uma análise cuidadosa do conjunto de dados de treinamento. Se o modelo não demonstrar melhoria incremental no desempenho, considere também a modificação dos hiperparâmetros de treinamento.
8. Depois que a avaliação do modelo atender aos padrões definidos pelas partes interessadas da empresa, você hospeda o modelo ajustado para a taxa de transferência provisionada do Amazon Bedrock.

Etapa 2: Prever o comportamento do paciente em relação aos medicamentos ou tratamentos prescritos

O Fine-Tuned LLMs pode processar notas clínicas, resumos de alta e outros documentos específicos do paciente a partir do gráfico de conhecimento médico temporal. Eles podem avaliar se é provável que o paciente siga os medicamentos ou tratamentos prescritos.

Essa etapa usa o gráfico de conhecimento criado emEtapa 1: Prever os resultados dos pacientes usando um gráfico de conhecimento médico. O gráfico de conhecimento contém dados do perfil do paciente, incluindo a adesão histórica do paciente como nódulo. Também inclui casos de não adesão a medicamentos ou tratamentos, efeitos colaterais aos medicamentos, falta de acesso ou barreiras de custo aos medicamentos ou regimes de dosagem complexos como atributos desses linfonodos.

O Fine-Tuned LLMs pode consumir dados anteriores de atendimento de prescrições do gráfico de conhecimento médico e resumos descritivos das notas clínicas de um banco de dados vetoriais do Amazon Service. OpenSearch Essas notas clínicas podem mencionar consultas frequentemente perdidas ou não adesão aos tratamentos. O LLM pode usar essas notas para prever a probabilidade de não adesão futura.

1. Prepare os dados de entrada da seguinte forma:
   - Dados estruturados — Extraia dados recentes do paciente, como as últimas três visitas e os resultados do laboratório, do gráfico de conhecimento médico.
   - Dados não estruturados — Recupere as notas clínicas recentes do banco de dados vetoriais do Amazon OpenSearch Service.
2. Crie um prompt de entrada que inclua o histórico do paciente e o contexto atual. Veja a seguir um exemplo de prompt:

```
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. Passe o prompt para o LLM ajustado. O LLM processa a solicitação e prevê o resultado. Veja a seguir um exemplo de resposta do 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. Analise a resposta do modelo para extrair a categoria de resultado previsto. Por exemplo, a categoria do exemplo de resposta na etapa anterior pode ser a de alta probabilidade de não adesão.
5. (Opcional) Use logits de modelo ou métodos adicionais para atribuir pontuações de confiança. Logits são as probabilidades não normalizadas do item pertencer a uma determinada classe ou categoria.

Etapa 3: Prever a probabilidade de readmissão do paciente

As readmissões hospitalares são uma grande preocupação devido ao alto custo da administração da saúde e ao seu impacto no bem-estar do paciente. O cálculo das taxas de readmissão hospitalar é uma forma de medir a qualidade do atendimento ao paciente e o desempenho de um profissional de saúde.

Para calcular a taxa de readmissão, você definiu um indicador, como uma taxa de readmissão de 7 dias. Esse indicador é a porcentagem de pacientes internados que retornam ao hospital para uma visita não planejada dentro de sete dias após a alta. Para prever a chance de readmissão de um paciente, um LLM ajustado pode consumir dados temporais do gráfico de conhecimento médico que você criou em. Etapa 1: Prever os resultados dos pacientes usando um gráfico de conhecimento médico Esse gráfico de conhecimento mantém registros cronológicos de encontros com pacientes, procedimentos, medicamentos e sintomas. Esses registros de dados contêm o seguinte:

- Duração do tempo desde a última alta do paciente
- A resposta do paciente aos tratamentos e medicamentos anteriores

- A progressão dos sintomas ou condições ao longo do tempo

Você pode processar esses eventos de séries temporais para prever a probabilidade de readmissão de um paciente por meio de um prompt de sistema organizado. O prompt transmite a lógica de previsão ao LLM ajustado.

1. Prepare os dados de entrada da seguinte forma:
   - Histórico de adesão — Extraia datas de coleta de medicamentos, frequências de recarga de medicamentos, detalhes de diagnóstico e medicação, histórico médico cronológico e outras informações do gráfico de conhecimento médico.
   - Indicadores comportamentais — Recupere e inclua notas clínicas sobre consultas perdidas e efeitos colaterais relatados pelo paciente.
2. Crie um prompt de entrada que inclua o histórico de adesão e os indicadores comportamentais. Veja a seguir um exemplo de prompt:

```
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. Passe o prompt para o LLM ajustado. O LLM processa a solicitação e prevê a probabilidade e os motivos da readmissão. Veja a seguir um exemplo de resposta do 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. Categorize a previsão em uma escala padronizada, como baixa, média ou alta.
5. Analise o raciocínio fornecido pelo LLM e identifique os principais fatores que contribuem para a previsão.
6. Mapeie os resultados qualitativos para pontuações quantitativas. Por exemplo, muito alto pode ser igual a uma probabilidade de 0,9.
```

7. Use conjuntos de dados de validação para calibrar as saídas do modelo em relação às taxas reais de readmissão.

Etapa 4: Calcular a pontuação de propensão à readmissão hospitalar

Em seguida, você calcula uma pontuação de propensão à readmissão hospitalar por paciente. Essa pontuação reflete o impacto líquido das três análises realizadas nas etapas anteriores: resultados potenciais do paciente, comportamento do paciente em relação a medicamentos e tratamentos e probabilidade de readmissão do paciente. Ao agregar a pontuação de propensão à readmissão em nível de paciente ao nível de especialidade e, em seguida, ao nível hospitalar, você pode obter informações para médicos, gerentes de cuidados e administradores. A pontuação de propensão à readmissão hospitalar ajuda você a avaliar o desempenho geral por instalação, especialidade ou condição. Em seguida, você pode usar essa pontuação para implementar intervenções proativas.

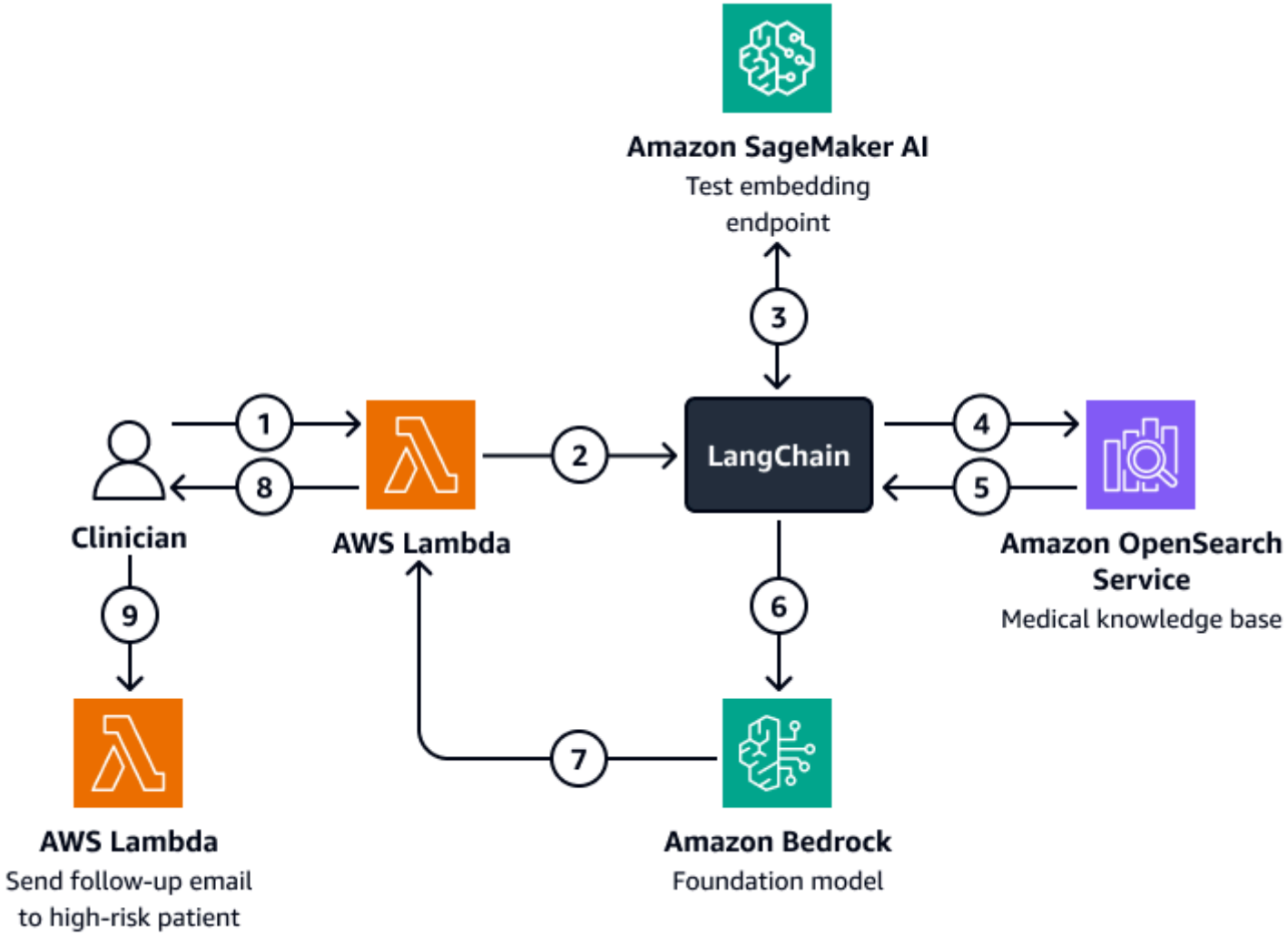
1. Atribua pesos a cada um dos diferentes fatores (previsão de resultados, probabilidade de adesão, readmissão). A seguir estão exemplos de pesos:
   - Peso da previsão de resultado: 0,4
   - Peso da previsão de aderência: 0,3
   - Peso da probabilidade de readmissão: 0,3
2. Use o cálculo a seguir para calcular a pontuação composta:

   `ReadadmissionPropensityScore` = (`OutcomeScore` × `OutcomeWeight`) + (`AdherenceScore` × `AdherenceWeight`) + (`ReadmissionLikelihoodScore` × `ReadmissionLikelihoodWeight`)
3. Certifique-se de que todas as pontuações individuais estejam na mesma escala, como 0 a 1.
4. Defina os limites para a ação. Por exemplo, pontuações acima de 0,7 iniciam alertas.

Com base nas análises acima e na pontuação de propensão à readmissão de um paciente, médicos ou gerentes de atendimento podem configurar alertas para monitorar seus pacientes individuais com base na pontuação computada. Se estiver acima de um limite predefinido, eles serão notificados quando esse limite for atingido. Isso ajuda os gerentes de atendimento a serem proativos em vez de reativos ao criar planos de cuidados de alta para seus pacientes. Salve os resultados, o comportamento e as pontuações de propensão à readmissão do paciente em um

formato indexado em um banco de dados vetorial do Amazon OpenSearch Service para que os gerentes de atendimento possam recuperá-los facilmente usando um agente conversacional de IA.

O diagrama a seguir mostra o fluxo de trabalho de um agente conversacional de IA que um médico ou gerente de atendimento pode usar para obter informações sobre os resultados dos pacientes, o comportamento esperado e a propensão à readmissão. Os usuários podem recuperar insights no nível do paciente, do departamento ou do hospital. O agente de IA recupera esses insights, que são armazenados em um formato indexado em um banco de dados vetorial do Amazon OpenSearch Service. O agente usa a consulta para recuperar dados relevantes e fornece respostas personalizadas, incluindo ações sugeridas para pacientes com alto risco de readmissão. Com base no nível de risco, o agente também pode configurar lembretes para pacientes e cuidadores.



O diagrama mostra o seguinte fluxo de trabalho:

1. O médico faz uma pergunta a um agente conversacional de IA, que abriga uma AWS Lambda função.
2. A função Lambda inicia um LangChain agente.
3. A ferramenta LangChain O agente envia a pergunta do usuário para um endpoint de incorporação de texto do Amazon SageMaker AI. O endpoint incorpora a pergunta.
4. A ferramenta LangChain o agente passa a pergunta incorporada para uma base de conhecimento médico no Amazon OpenSearch Service.
5. O Amazon OpenSearch Service retorna os insights específicos que são mais relevantes para a consulta do usuário para o LangChain agente.
6. A ferramenta LangChain os agentes enviam a consulta e o contexto recuperado da base de conhecimento para um modelo da Amazon Bedrock Foundation.
7. O modelo Amazon Bedrock Foundation gera uma resposta e a envia para a função Lambda.
8. A função Lambda retorna a resposta ao médico.
9. O médico inicia uma função Lambda que envia um e-mail de acompanhamento a um paciente com alto risco de readmissão.

Alinhamento com o AWS Well-Architected Framework

A arquitetura para monitorar o comportamento do paciente e prever as taxas de readmissão hospitalar integra Serviços da AWS gráficos de conhecimento médico e melhorar os resultados de saúde, alinhando-se LLMs aos seis pilares do Well-Architected Framework:AWS

- Excelência operacional — A solução é um sistema automatizado e desacoplado que usa o Amazon Bedrock e AWS Lambda para alertas em tempo real.
- Segurança — Essa solução foi projetada para estar em conformidade com os regulamentos de saúde, como o HIPAA. Você também pode implementar criptografia, controle de acesso refinado e grades de proteção Amazon Bedrock para ajudar a proteger os dados dos pacientes.
- Confiabilidade — A arquitetura usa tolerância a falhas e sem servidor. Serviços da AWS
- Eficiência de desempenho — o Amazon OpenSearch Service e o fine-tuned LLMs podem fornecer previsões rápidas e precisas.
- Otimização de custos — tecnologias e pay-per-inference modelos sem servidor ajudam a minimizar os custos. Embora o uso de LLM ajustado possa incorrer em custos extras, o modelo usa uma abordagem RAG que reduz os dados e o tempo computacional necessários para o processo de ajuste fino.

- Sustentabilidade — A arquitetura minimiza o consumo de recursos por meio do uso da infraestrutura sem servidor. Ele também oferece suporte a operações de saúde eficientes e escaláveis.

Caso de uso: Gerenciando e aprimorando sua equipe de saúde

A implementação de estratégias de transformação e aprimoramento de talentos ajuda as forças de trabalho a permanecerem hábeis no uso de novas tecnologias e práticas em serviços médicos e de saúde. Iniciativas proativas de aprimoramento de habilidades garantem que os profissionais de saúde possam oferecer atendimento de alta qualidade aos pacientes, otimizar a eficiência operacional e manter a conformidade com os padrões regulatórios. Além disso, a transformação de talentos promove uma cultura de aprendizado contínuo. Isso é fundamental para se adaptar às mudanças no cenário da saúde e enfrentar os desafios emergentes da saúde pública. As abordagens tradicionais de treinamento, como o treinamento em sala de aula e os módulos estáticos de aprendizado, oferecem conteúdo uniforme para um público amplo. Eles geralmente carecem de caminhos de aprendizagem personalizados, que são essenciais para atender às necessidades específicas e aos níveis de proficiência de cada profissional. Essa one-size-fits-all estratégia pode resultar em desengajamento e retenção de conhecimento abaixo do ideal.

Consequentemente, as organizações de saúde devem adotar soluções inovadoras, escaláveis e orientadas pela tecnologia que possam determinar a lacuna de cada um de seus funcionários em seu estado atual e em seu estado futuro potencial. Essas soluções devem recomendar caminhos de aprendizagem hiperpersonalizados e o conjunto certo de conteúdo de aprendizagem. Isso prepara efetivamente a força de trabalho para o futuro da saúde.

No setor de saúde, você pode aplicar a IA generativa para ajudá-lo a entender e aprimorar sua força de trabalho. Por meio da conexão de grandes modelos de linguagem (LLMs) e recuperadores avançados, as organizações podem entender quais habilidades possuem atualmente e identificar as principais habilidades que podem ser necessárias no futuro. Essas informações ajudam você a preencher a lacuna contratando novos trabalhadores e aprimorando a força de trabalho atual. Usando o Amazon Bedrock e os gráficos de conhecimento, as organizações de saúde podem desenvolver aplicativos específicos de domínio que facilitam o aprendizado contínuo e o desenvolvimento de habilidades.

O conhecimento fornecido por essa solução ajuda você a gerenciar talentos de forma eficaz, otimizar o desempenho da força de trabalho, impulsionar o sucesso organizacional, identificar as habilidades existentes e criar uma estratégia de talentos. Essa solução pode ajudá-lo a realizar essas tarefas em semanas, em vez de meses.

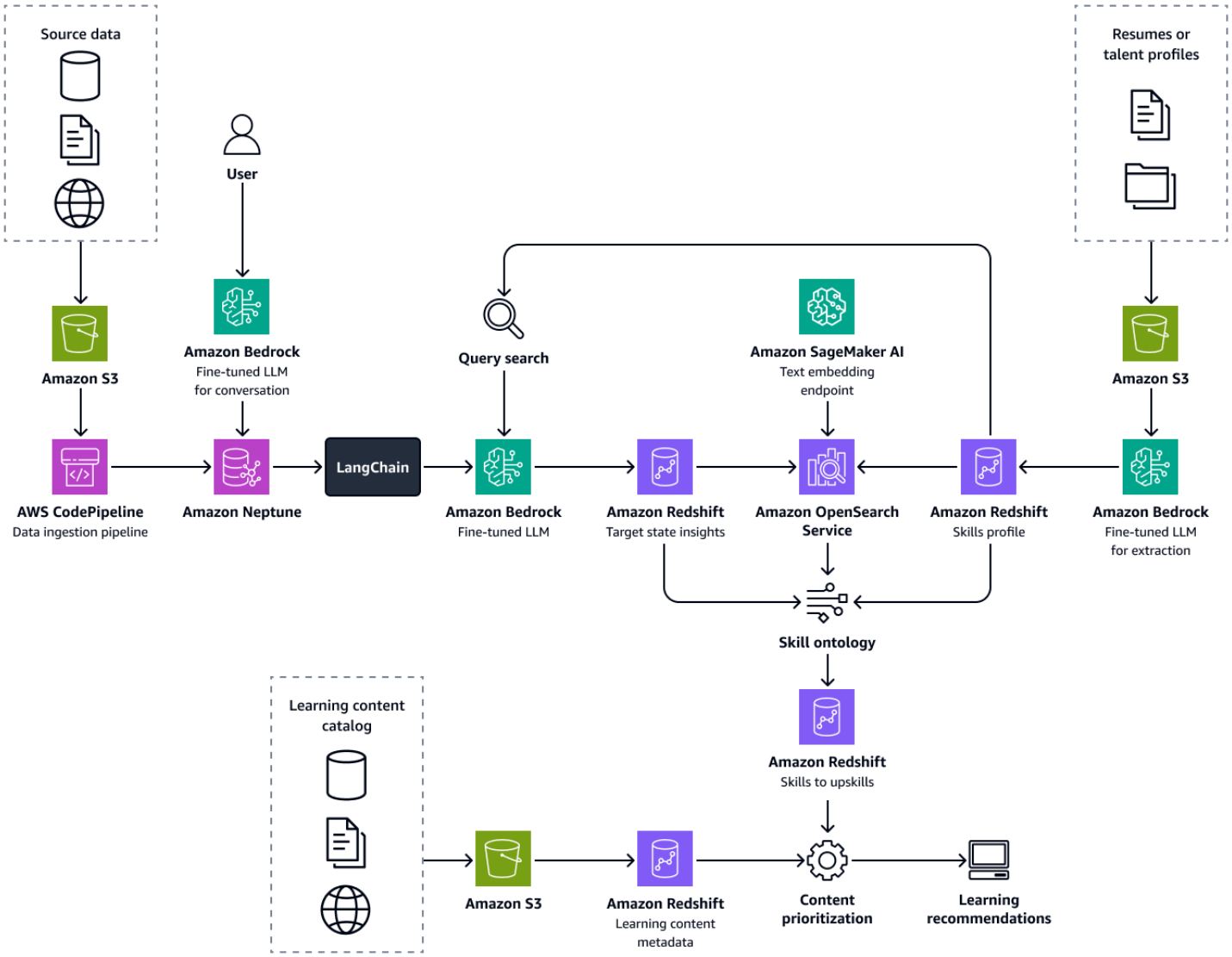
Visão geral da solução

Essa solução é uma estrutura de transformação de talentos da área de saúde que consiste nos seguintes componentes:

- Analisador inteligente de currículos — Esse componente pode ler o currículo de um candidato e extrair com precisão as informações do candidato, incluindo habilidades. Solução inteligente de extração de informações criada usando o modelo Llama 2 ajustado no Amazon Bedrock em um conjunto de dados de treinamento proprietário que abrange currículos e perfis de talentos de mais de 19 setores. Esse processo baseado em LLM economiza centenas de horas automatizando o processo de revisão manual de currículos e combinando os melhores candidatos com as vagas abertas.
- Gráfico de conhecimento — Um gráfico de conhecimento criado no Amazon Neptune, um repositório unificado de informações sobre talentos, incluindo a taxonomia de funções e habilidades da organização e do setor, capturando a semântica dos talentos da área de saúde usando definições de habilidades, funções e suas propriedades, relações e restrições lógicas.
- Ontologia de habilidades — A descoberta da proximidade de habilidades entre as habilidades do candidato e as habilidades ideais do estado atual ou do futuro (recuperadas usando um gráfico de conhecimento) é obtida por meio de algoritmos de ontologia que medem a semelhança semântica entre as habilidades do candidato e as habilidades do estado-alvo.
- Caminho e conteúdo de aprendizado — Esse componente é um mecanismo de recomendação de aprendizado que pode recomendar o conteúdo de aprendizado correto a partir de um catálogo de materiais didáticos de qualquer fornecedor, com base nas lacunas de habilidades identificadas. Identificar os melhores caminhos de aprimoramento de habilidades para cada candidato, analisando as lacunas de habilidades e recomendando conteúdo de aprendizagem priorizado, para permitir um desenvolvimento profissional contínuo e contínuo para cada candidato durante a transição para uma nova função.

Essa solução automatizada baseada em nuvem é alimentada por serviços de aprendizado de máquina LLMs, gráficos de conhecimento e Geração Aumentada de Recuperação (RAG). Ele pode ser escalado para processar dezenas ou milhares de currículos em um período mínimo de tempo, criar perfis instantâneos de candidatos, identificar lacunas em seu estado futuro atual ou potencial e, em seguida, recomendar com eficiência o conteúdo de aprendizagem certo para preencher essas lacunas.

A imagem a seguir mostra o end-to-end fluxo da estrutura. A solução foi desenvolvida e aperfeiçoada no LLMs Amazon Bedrock. Eles LLMs recuperam dados da base de conhecimento de talentos da área de saúde no Amazon Neptune. Algoritmos baseados em dados fazem recomendações para caminhos de aprendizagem ideais para cada candidato.



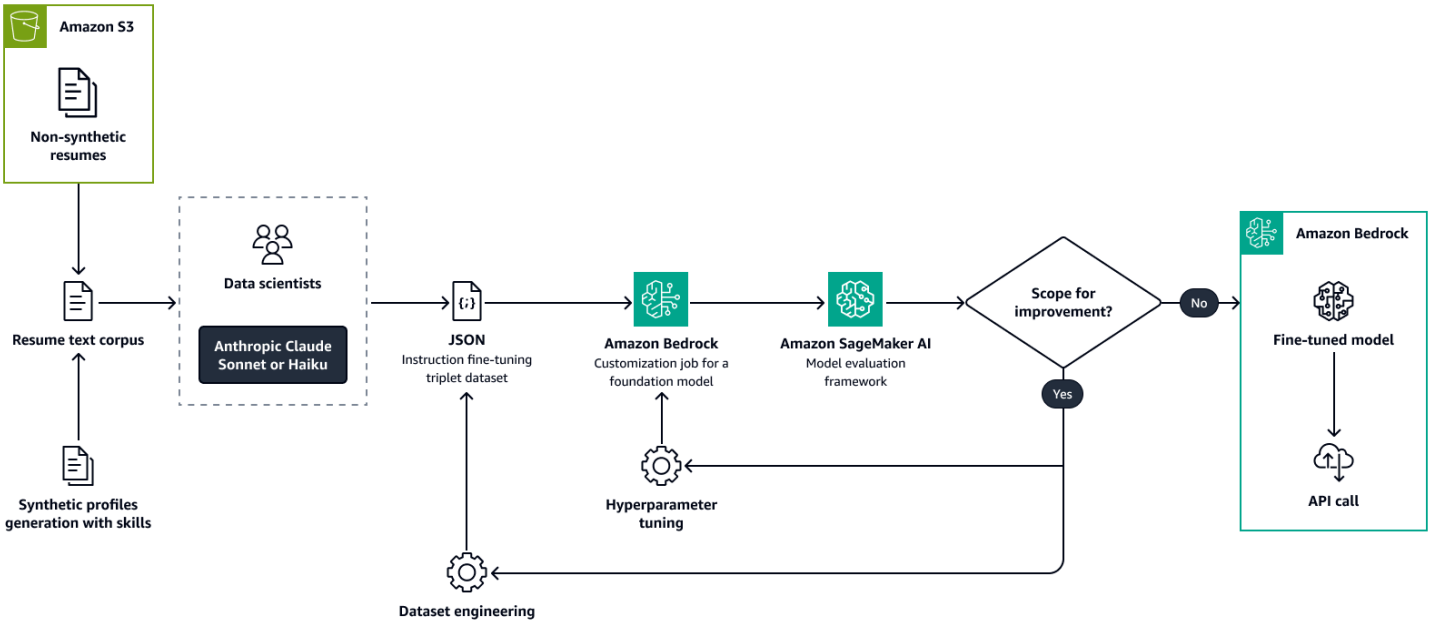
A criação dessa solução consiste nas seguintes etapas:

- Etapa 1: Extrair informações sobre talentos e criar um perfil de habilidades
- Etapa 2: Descobrir a role-to-skill relevância em um gráfico de conhecimento
- Etapa 3: Identificar lacunas de habilidades e recomendar treinamento

Etapa 1: Extrair informações sobre talentos e criar um perfil de habilidades

Primeiro, você ajusta um grande modelo de linguagem, como o Llama 2, no Amazon Bedrock com um conjunto de dados personalizado. Isso adapta o LLM para o caso de uso. Durante o treinamento, você extrai de forma precisa e consistente os principais atributos de talento dos currículos dos candidatos ou perfis de talentos semelhantes. Esses atributos de talento incluem habilidades, cargo atual, títulos de experiência com datas, educação e certificações. Para obter mais informações, consulte Personalizar seu modelo para melhorar seu desempenho para seu caso de uso na documentação do Amazon Bedrock.

A imagem a seguir mostra o processo para ajustar um modelo de análise de currículo usando o Amazon Bedrock. Os currículos reais e criados sinteticamente são passados para um LLM para extrair informações importantes. Um grupo de cientistas de dados valida as informações extraídas em relação ao texto original e bruto. As informações extraídas são então concatenadas usando a chain-of-thoughtsolicitação e o texto original para derivar um conjunto de dados de treinamento para ajuste fino. Esse conjunto de dados é então passado para um trabalho de personalização do Amazon Bedrock, que ajusta o modelo. Um trabalho em lote do Amazon SageMaker AI executa uma estrutura de avaliação de modelo que avalia o modelo ajustado. Se o modelo precisar ser aprimorado, o trabalho será executado novamente com mais dados ou hiperparâmetros diferentes. Depois que a avaliação atender aos padrões, você hospeda o modelo personalizado por meio da taxa de transferência provisionada do Amazon Bedrock.



Etapa 2: Descobrir a role-to-skill relevância em um gráfico de conhecimento

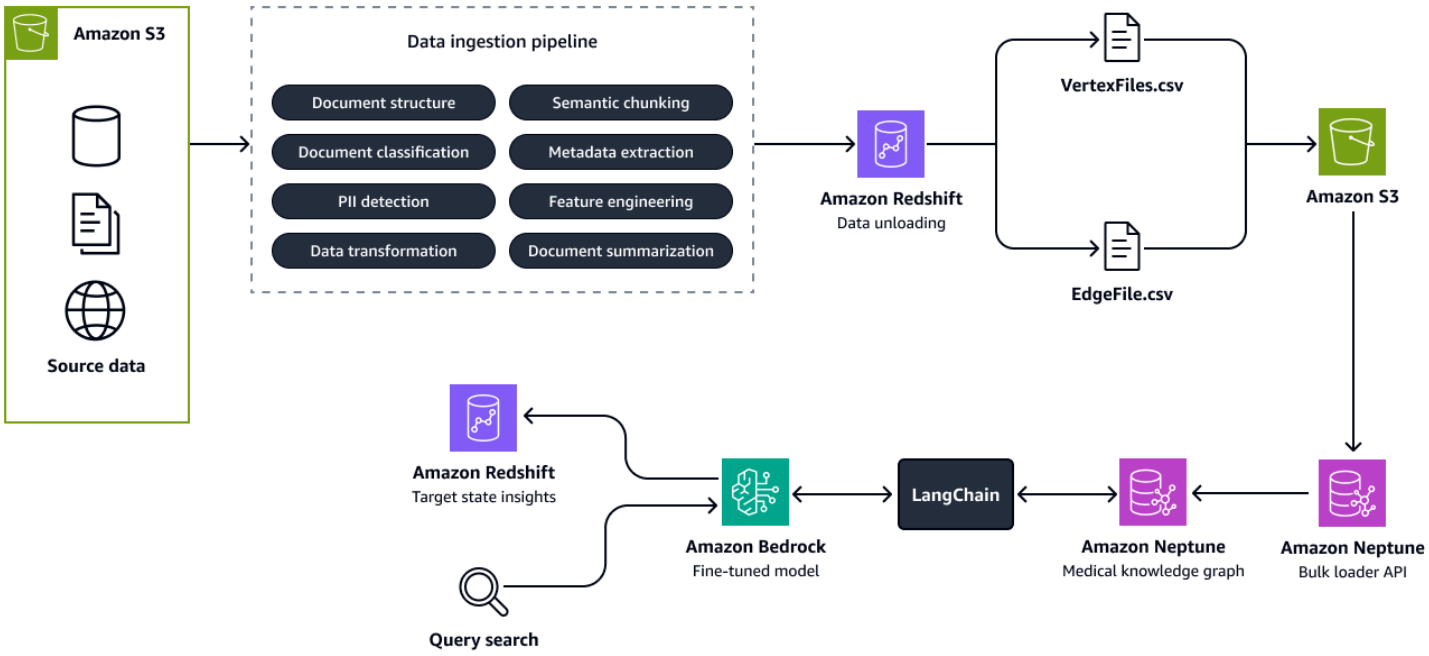
Em seguida, você cria um gráfico de conhecimento que encapsula as habilidades e a taxonomia de funções de sua organização e de outras organizações do setor de saúde. Essa base de conhecimento enriquecida é obtida a partir de dados agregados de talentos e organizações no Amazon Redshift. Você pode coletar dados de talentos de uma variedade de provedores de dados do mercado de trabalho e de fontes de dados estruturadas e não estruturadas específicas da organização, como sistemas de planejamento de recursos corporativos (ERP), um sistema de informações de recursos humanos (HRIS), currículos de funcionários, descrições de cargos e documentos de arquitetura de talentos.

Crie o gráfico de conhecimento no Amazon Neptune. Os nós representam habilidades e funções, e as bordas representam as relações entre eles. Enriqueça esse gráfico com metadados para incluir detalhes como nome da organização, setor, família de cargos, tipo de habilidade, tipo de função e etiquetas do setor.

Em seguida, você desenvolve um aplicativo Graph Retrieval Augmented Generation (Graph RAG). O Graph RAG é uma abordagem RAG que recupera dados de um banco de dados gráfico. A seguir estão os componentes do aplicativo Graph RAG:

- Integração com um LLM no Amazon Bedrock — O aplicativo usa um LLM no Amazon Bedrock para compreensão da linguagem natural e geração de consultas. Os usuários podem interagir com o sistema usando linguagem natural. Isso o torna acessível a partes interessadas não técnicas.
- Orquestração e recuperação de informações — Uso ou LlamaIndexLangChainorquestradores para facilitar a integração entre o LLM e o gráfico de conhecimento de Neptune. Eles gerenciam o processo de conversão de consultas de linguagem natural em consultas OpenCypher. Em seguida, eles executam as consultas no gráfico de conhecimento. Use engenharia imediata para instruir o LLM sobre as melhores práticas para criar consultas OpenCypher. Isso ajuda a otimizar as consultas para recuperar o subgráfico relevante, que contém todas as entidades e relacionamentos pertinentes sobre as funções e habilidades consultadas.
- Geração de insights — O LLM no Amazon Bedrock processa os dados gráficos recuperados. Ele gera insights detalhados sobre o estado atual e projeta os estados futuros da função consultada e das habilidades associadas.

A imagem a seguir mostra as etapas para criar um gráfico de conhecimento a partir dos dados de origem. Você passa os dados de origem estruturados e não estruturados para o pipeline de ingestão de dados. O pipeline extrai e transforma as informações em uma formação de carga em massa CSV compatível com o Amazon Neptune. A API do carregador em massa carrega os arquivos CSV que estão armazenados em um bucket do Amazon S3 para o gráfico de conhecimento do Neptune. Para consultas de usuários relacionadas a talentos, futuro estado, funções ou habilidades relevantes, o LLM aperfeiçoado no Amazon Bedrock interage com o gráfico de conhecimento por meio de um LangChain orquestrador. O orquestrador recupera o contexto relevante do gráfico de conhecimento e envia as respostas para a tabela de insights no Amazon Redshift. A ferramenta LangChain um orquestrador, como o Graph QAChain, converte a consulta de linguagem natural do usuário em uma consulta OpenCypher para consultar o gráfico de conhecimento. O modelo ajustado do Amazon Bedrock gera uma resposta com base no contexto recuperado.



Etapa 3: Identificar lacunas de habilidades e recomendar treinamento

Nesta etapa, você calcula com precisão a proximidade entre o estado atual de um profissional de saúde e as possíveis funções do future state. Para fazer isso, você realiza uma análise de afinidade de habilidades comparando os conjuntos de habilidades do indivíduo com a função profissional. Em um banco de dados vetorial do Amazon OpenSearch Service, você armazena informações de taxonomia de habilidades e metadados de habilidades, como a descrição da habilidade, o tipo

de habilidade e os grupos de habilidades. Use um modelo de incorporação do Amazon Bedrock, como os modelos Amazon Titan Text Embeddings, para incorporar a habilidade-chave identificada aos vetores. Por meio de uma pesquisa vetorial, você recupera as descrições das habilidades do estado atual e das habilidades do estado alvo e realiza uma análise ontológica. A análise fornece pontuações de proximidade entre os pares de habilidades do estado atual e do estado alvo. Para cada par, você usa as pontuações computadas da ontologia para identificar as lacunas nas afinidades de habilidades. Em seguida, você recomenda o caminho ideal para o aprimoramento de habilidades, que o candidato pode considerar durante as transições de funções.

Para cada função, recomendar o conteúdo de aprendizagem correto para aprimoramento ou requalificação envolve uma abordagem sistemática que começa com a criação de um catálogo abrangente de conteúdo de aprendizagem. Esse catálogo, que você armazena em um banco de dados do Amazon Redshift, agrega conteúdo de vários provedores e inclui metadados, como duração do conteúdo, nível de dificuldade e modo de aprendizado. A próxima etapa é extrair as principais habilidades oferecidas por cada conteúdo e, em seguida, mapeá-las de acordo com as habilidades individuais necessárias para a função alvo. Você consegue esse mapeamento analisando a cobertura fornecida pelo conteúdo por meio de uma análise de proximidade de habilidades. Essa análise avalia até que ponto as habilidades ensinadas pelo conteúdo se alinham às habilidades desejadas para a função. Os metadados desempenham um papel fundamental na seleção do conteúdo mais adequado para cada habilidade, garantindo que os alunos recebam recomendações personalizadas que atendam às suas necessidades de aprendizado. Use LLMs no Amazon Bedrock para extrair habilidades dos metadados do conteúdo, realizar engenharia de recursos e validar as recomendações de conteúdo. Isso melhora a precisão e a relevância no processo de aprimoramento ou requalificação.

Alinhamento com o AWS Well-Architected Framework

A solução está alinhada com todos os seis pilares do Well-Architected AWS Framework:

- Excelência operacional — Um pipeline modular e automatizado aumenta a excelência operacional. Os principais componentes do pipeline são desacoplados e automatizados, permitindo atualizações mais rápidas do modelo e monitoramento mais fácil. Além disso, os canais de treinamento automatizados oferecem suporte a lançamentos mais rápidos de modelos ajustados.
- Segurança — Essa solução processa informações confidenciais e de identificação pessoal (PII), como dados em currículos e perfis de talentos. Em AWS Identity and Access Management (IAM), implemente políticas de controle de acesso refinadas e garanta que somente funcionários autorizados tenham acesso a esses dados.

- Confiabilidade — A solução usa Serviços da AWS, como Neptune, Amazon Bedrock e Service OpenSearch , que fornecem tolerância a falhas, alta disponibilidade e acesso ininterrupto a insights, mesmo durante alta demanda.
- Eficiência de desempenho — Os bancos de dados vetoriais aprimorados do LLMs Amazon Bedrock and OpenSearch Service foram projetados para processar grandes conjuntos de dados com rapidez e precisão, a fim de fornecer recomendações de aprendizado personalizadas e oportunas.
- Otimização de custos — Essa solução usa uma abordagem RAG, que reduz a necessidade de pré-treinamento contínuo dos modelos. Em vez de ajustar todo o modelo repetidamente, o sistema ajusta apenas processos específicos, como extrair informações de currículos e estruturar resultados. Isso resulta em economias de custo significativas. Ao minimizar a frequência e a escala do treinamento de modelos com uso intensivo de recursos e ao usar serviços em pay-per-use nuvem, as organizações de saúde podem otimizar seus custos operacionais e, ao mesmo tempo, manter o alto desempenho.
- Sustentabilidade — Essa solução usa serviços escaláveis e nativos da nuvem que alocam recursos de computação dinamicamente. Isso reduz o consumo de energia e o impacto ambiental, ao mesmo tempo em que apoia iniciativas de transformação de talentos em grande escala e com uso intensivo de dados.

Desenvolvendo e orquestrando soluções generativas de IA para assistência médica

Para criar as soluções deste guia, você deve criar uma arquitetura RAG que seja ajustada LLMs para fornecer dados aumentados do paciente, insights clínicos e de diagnóstico e resultados previstos dos pacientes aos profissionais de saúde. Isso requer a integração de várias Serviços da AWS ferramentas para criar um fluxo de trabalho coeso e eficiente. Esta seção aborda o seguinte:

- Amazon Q Developer— Use o Amazon Q Developer para resolver questões de engenharia e erros de código durante o processo de desenvolvimento.
- Design RAG com vários recuperadores— Projete e implemente soluções RAG que usam vários recuperadores para obter o contexto médico correto para a pergunta do usuário.
- ReAct agentes— Implemente agentes que combinem raciocínio com ação dinâmica.

Amazon Q Developer

Ao criar uma solução generativa de IA, pode ser difícil criar agentes de IA e conectar os principais serviços. No entanto, o Amazon Q Developer ajuda cientistas de dados e engenheiros de IA fornecendo acesso a um assistente avançado de IA generativa. O Amazon Q pode resolver com rapidez e precisão as perguntas dos usuários e os erros de código, o que pode ajudar você a otimizar o processo de desenvolvimento do LLM. O Amazon Q oferece vantagens significativas para desenvolvedores que criam aplicativos que usam os modelos básicos do Amazon Bedrock. Ele pode agilizar os fluxos de trabalho e melhorar a qualidade do código. Ele automatiza a geração de scripts Python e configurações de infraestrutura como código (IaC), reduzindo significativamente o tempo e o esforço de desenvolvimento. Por meio de recursos avançados de refatoração, o Amazon Q pode melhorar o desempenho do código, identificar vulnerabilidades de segurança e garantir que os desenvolvedores sigam as melhores práticas. Além disso, facilita o aprendizado e a adoção para iniciantes, fornecendo sugestões e explicações contextuais, tornando as tarefas complexas de codificação mais acessíveis e eficientes.

Design RAG com vários recuperadores

Em um aplicativo generativo de IA, um pipeline RAG com vários recuperadores pode recuperar com eficiência informações de várias fontes de dados para ajudar profissionais de saúde e médicos

a responder perguntas médicas. Esse pipeline usa diferentes tipos de recuperadores para extrair dados relevantes de diferentes bases de conhecimento. Cada recuperador é especializado em buscar um tipo específico de informação, como históricos de pacientes, informações diagnósticas, notas clínicas ou conteúdo de pesquisas médicas e textos acadêmicos.

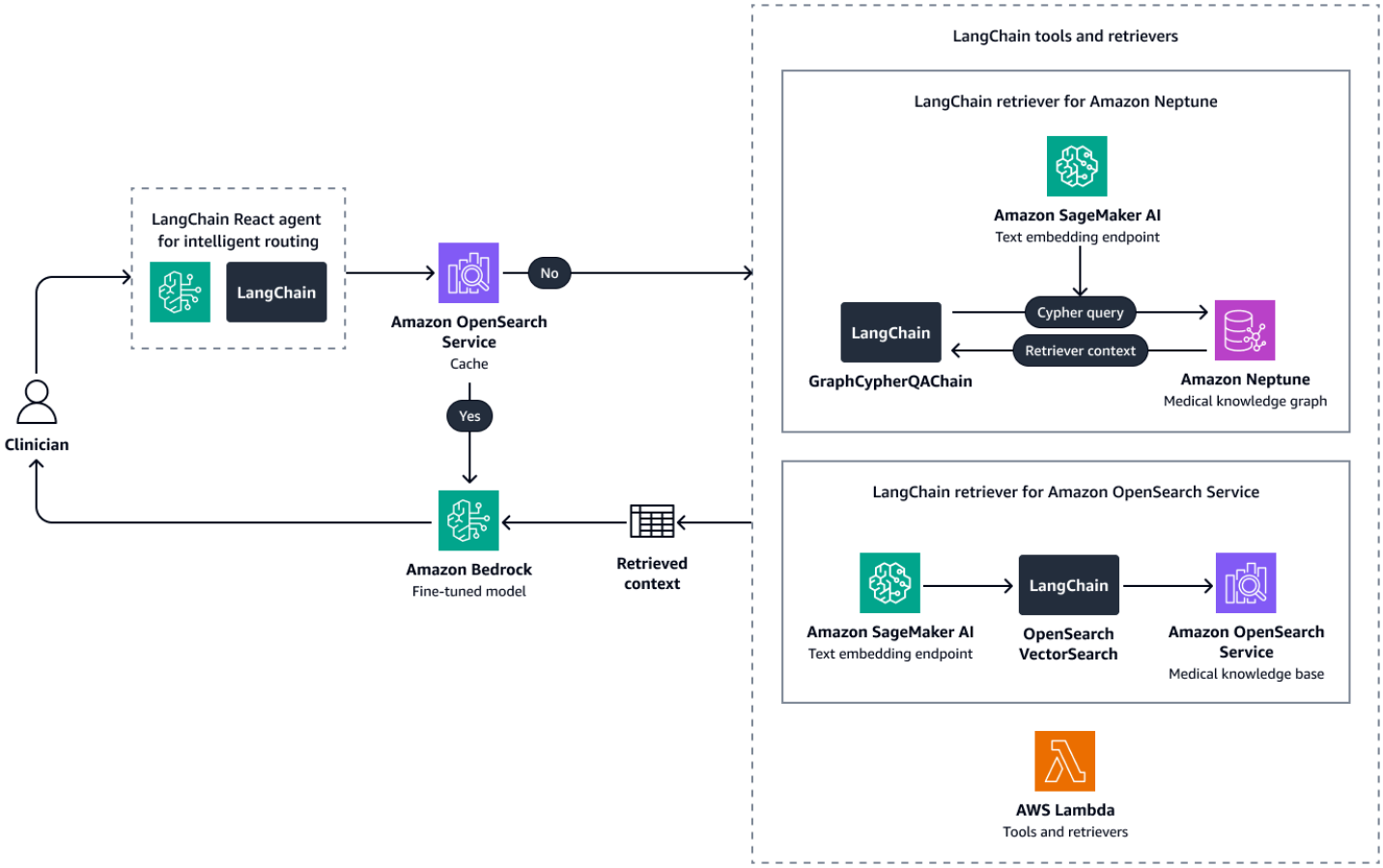
Use a natureza dos dados e os requisitos específicos do aplicativo para determinar qual é a base de conhecimento de back-end correta para seu caso de uso. Um banco de dados vetorial do Amazon OpenSearch Service é adequado para grandes volumes de dados de saúde não estruturados ou semiestruturados, incluindo resumos de avaliação de diagnóstico por imagem, resumos de alta, relatórios clínicos, pesquisas médicas e conteúdo de texto acadêmico. Por outro lado, um serviço de banco de dados gráfico, como o Amazon Neptune, pode ser ideal para casos de uso de assistência médica que exijam uma exploração profunda das relações temporais entre entidades, como paciente, histórico do paciente, profissional de saúde, medicamentos, sintomas e tratamentos.

Um componente essencial desse pipeline é a previsão da intenção de consulta do usuário. Isso garante que o sistema encaminhe a consulta para a cadeia de recuperação correta. Por exemplo, se um médico perguntar sobre o histórico de tratamento, os sintomas, a interação com o hospital, a probabilidade de readmissão hospitalar ou os possíveis resultados do paciente, o módulo de predição da intenção de consulta identifica essa intenção. Ele direciona a solicitação para a cadeia de recuperadores, que pode buscar registros de pacientes ou dados cronológicos de tratamento a partir do gráfico de conhecimento médico. Como alternativa, se a pergunta for sobre a descoberta de doenças, avaliações diagnósticas específicas ou detalhes de procedimentos clínicos específicos de livros acadêmicos, a consulta será encaminhada para a cadeia de recuperadores que pode buscar essas informações do banco de dados vetoriais do Service. OpenSearch Você pode usar a funcionalidade de chamada de ferramentas do LangChain para vincular uma ferramenta personalizada ao Amazon Bedrock LLM que possa classificar uma pergunta do usuário em intenções predefinidas.

Este sistema RAG multi-retriever inclui LangChain agentes projetados para gerenciar o acesso à base de conhecimento específica. Você pode usar: LangChain para orquestrar a interação entre o Amazon Bedrock LLM, os diferentes retrievers e ferramentas. LangChain inclui uma classe de chamada de ferramentas que ajuda você a criar ferramentas personalizadas, como um classificador de intenção, um recuperador para Neptune, um recuperador para OpenSearch Serviço ou qualquer outra ferramenta que possa ser desenvolvida para classificar a intenção do usuário e acessar dados de uma base de conhecimento específica em um formato estruturado. Em seguida, você fornece essas ferramentas à classe para criar um agente de Raciocínio e Ação (ReAct). O ReAct agente processa a pergunta do usuário, planeja as etapas sequenciais para responder à pergunta e, em

seguida, executa iterativamente as ferramentas disponíveis e processa as respostas da ferramenta para finalmente responder à consulta do usuário.

A imagem a seguir mostra como funciona um sistema RAG com vários recuperadores projetado para recuperação eficiente de conhecimento e resolução inteligente de consultas. A LangChain ReAct o agente analisa a intenção do usuário, formula um plano estruturado para execução e seleciona as ferramentas de recuperação mais relevantes. O sistema consulta um cache de perguntas anteriores e verifica consultas semelhantes com base nos principais atributos, como ID do paciente, condição médica e data da visita. Se uma pergunta muito semelhante for encontrada, a resposta correspondente será recuperada diretamente. Caso contrário, o agente executa o recuperador apropriado. Para recuperar informações centradas no paciente, como histórico de tratamento, sintomas, interações hospitalares ou probabilidade de readmissão, o sistema usa um recuperador gráfico. Para avaliações diagnósticas, procedimentos clínicos e achados médicos estruturados, o agente emprega um recuperador de banco de dados vetoriais. Em cenários que exigem uma combinação de conhecimento contextual de ambos os armazenamentos de dados, para gerar uma resposta abrangente, o sistema usa uma estratégia de recuperação híbrida que integra os resultados do gráfico de conhecimento e do banco de dados vetorial.



ReAct agentes

Os agentes de Raciocínio e Ação (ReAct) são projetados para aplicações RAG multifacetadas. Esses agentes fornecem uma combinação poderosa de raciocínio e ação dinâmica, especialmente para aplicativos complexos que envolvem step-by-step fluxos de trabalho lógicos de recuperação de informações. Para obter mais informações, consulte ReAct: Sinergizando o raciocínio e a ação em modelos de linguagem.

Em contextos médicos e de saúde, as consultas de um clínico ou médico geralmente são multifacetadas. Por exemplo, um médico pode perguntar “Quais tratamentos foram dados a pacientes semelhantes com hipertensão e diabetes tipo 2?” Depois de identificar a intenção do usuário, que é buscar os tratamentos para hipertensão e diabetes tipo 2, o agente de IA precisa dividir essa consulta em subtarefas e, em seguida, escolher a estratégia de recuperação mais eficiente. Nesse caso, o agente de IA deve identificar os nós mais relevantes (como idade, sexo, condições, tratamentos e medicamentos do paciente) e, em seguida, consultar o gráfico dessas entidades e seus atributos e relacionamentos. ReAct os agentes são muito úteis porque combinam a capacidade de raciocínio (inferência lógica) de um LLM com uma ação (consultar ou interagir com recursos externos ou bases de conhecimento).

Para responder à pergunta do usuário “Quais tratamentos foram dados a pacientes semelhantes com hipertensão e diabetes tipo 2?” , o exemplo a seguir ilustra como um ReAct agente funciona:

1. Raciocínio do agente — O ReAct agente infere que a questão envolve a recuperação de informações sobre condições (diabetes e hipertensão). Ele considera a idade do paciente, os tratamentos, os medicamentos e o período de análise.
2. Ação do agente — O agente usa o OpenCypher para consultar o gráfico de conhecimento de tratamentos específicos para diabetes tipo 2 e hipertensão. Ele também recupera medicamentos administrados, datas das visitas ao hospital, efeitos colaterais dos medicamentos, resultados conhecidos dos pacientes e dados de referência cruzada para pacientes semelhantes (como pacientes do mesmo sexo e idade).
3. Observação do agente — A partir do gráfico de conhecimento, o agente recupera os seis meses mais recentes de dados tabulares sobre tratamentos administrados a pacientes com hipertensão e diabetes tipo 2.
4. Raciocínio do agente — Para classificar os resultados dos registros recuperados, o agente identifica atributos importantes, como recência, efeitos colaterais dos medicamentos ou resultados conhecidos do paciente.

5. Ação do agente — O agente reclassifica os registros com base nos atributos identificados e na lógica predefinida que é transmitida por meio do prompt do sistema.
6. Geração de respostas — O LLM no Amazon Bedrock gera uma resposta com base no contexto que o ReAct agente preparou.

Avaliação de soluções generativas de IA para assistência médica

Avaliar as soluções de IA de saúde que você cria é fundamental para garantir que elas sejam eficazes, confiáveis e escaláveis em ambientes médicos do mundo real. Use uma abordagem sistemática para avaliar o desempenho de cada componente da solução. A seguir está um resumo das metodologias e métricas que você pode usar para avaliar sua solução.

Tópicos

Avaliando a extração de informações

Avalie o desempenho das soluções de extração de informações, como o analisador inteligente de currículos e o extrator de entidades personalizado. Você pode medir o alinhamento das respostas dessas soluções usando um conjunto de dados de teste. Se você não tiver um conjunto de dados que abranja perfis versáteis de talentos da área de saúde e registros médicos de pacientes, você pode criar um conjunto de dados de teste personalizado usando a capacidade de raciocínio de um LLM. Por exemplo, você pode usar um modelo de parâmetros grandes, como Anthropic Claude modelos, para gerar um conjunto de dados de teste.

A seguir estão três métricas principais que você pode usar para avaliar os modelos de extração de informações:

- Precisão e integridade — Essas métricas avaliam até que ponto a saída capturou as informações corretas e completas presentes nos dados reais básicos. Isso envolve verificar a exatidão das informações extraídas e a presença de todos os detalhes relevantes nas informações extraídas.
- Similaridade e relevância — Essas métricas avaliam as semelhanças semânticas, estruturais e contextuais entre a saída e os dados verdadeiros fundamentais (a semelhança) e o grau em que a saída se alinha e aborda o conteúdo, o contexto e a intenção dos dados verídicos básicos (a relevância).

- Taxa ajustada de recall ou captura — Essas taxas determinam empiricamente quantos dos valores presentes nos dados reais fundamentais foram identificados corretamente pelo modelo. A taxa deve incluir uma penalização para todos os valores falsos que o modelo extrai.
- Pontuação de precisão — A pontuação de precisão ajuda a determinar quantos falsos positivos estão presentes nas previsões, em comparação com os verdadeiros positivos. Por exemplo, você pode usar métricas de precisão para medir a exatidão da proficiência de habilidade extraída.

Avaliando soluções RAG com vários recuperadores

Para avaliar o quão bem o sistema recupera as informações relevantes e com que eficácia ele usa essas informações para gerar respostas precisas e contextualmente apropriadas, você pode usar as seguintes métricas:

- Relevância da resposta — meça a relevância da resposta gerada, que usa o contexto recuperado, em relação à consulta original.
- Precisão do contexto — do total de resultados recuperados, avalie a proporção de documentos ou trechos recuperados que são relevantes para a consulta. Uma maior precisão de contexto indica que o mecanismo de recuperação é eficaz na seleção de informações relevantes.
- Fidelidade — avalia a precisão com que a resposta gerada reflete as informações no contexto recuperado. Em outras palavras, meça se a resposta permanece fiel às informações de origem.

Avaliando uma solução usando um LLM

Você pode usar uma técnica chamada LLM- as-a-judge para avaliar as respostas de texto da sua solução generativa de IA. Envolve o uso LLMs para avaliar e avaliar o desempenho dos resultados do modelo. Essa técnica usa os recursos do Amazon Bedrock para fornecer julgamentos sobre vários atributos, como qualidade da resposta, coerência, aderência, precisão e integridade das preferências humanas ou dados reais fundamentais. Você usa chain-of-thought (CoT) e algumas técnicas de solicitação para uma avaliação abrangente. O prompt instrui o LLM a avaliar a resposta gerada com a rubrica de pontuação, e as poucas amostras no prompt demonstram o processo de avaliação real. O aviso também inclui diretrizes para o avaliador do LLM seguir. Por exemplo, você pode considerar o uso de uma ou mais das seguintes técnicas de avaliação que usam um LLM para avaliar as respostas geradas:

- Comparação em pares — Dê ao avaliador do LLM uma pergunta médica e várias respostas que foram geradas por versões diferentes e iterativas dos sistemas RAG que você criou. Peça ao avaliador do LLM que determine a melhor resposta com base na qualidade da resposta, coerência e adesão à pergunta original.
- Classificação de resposta única — Essa técnica é adequada para casos de uso em que você precisa avaliar a precisão da categorização, como classificação dos resultados do paciente, categorização do comportamento do paciente, probabilidade de readmissão do paciente e categorização do risco. Use o avaliador LLM para analisar a categorização ou classificação individual isoladamente e avaliar o raciocínio fornecido com base em dados verídicos.
- Classificação guiada por referência — Forneça ao avaliador do LLM uma série de perguntas médicas que exigem respostas descritivas. Crie exemplos de respostas para essas perguntas, como respostas de referência ou respostas ideais. Solicite ao avaliador do LLM que compare a resposta gerada pelo LLM com as respostas de referência ou respostas ideais e solicite que o avaliador do LLM classifique a resposta gerada quanto à precisão, integridade, semelhança, relevância ou outros atributos. Essa técnica ajuda a avaliar se as respostas geradas estão alinhadas com uma resposta padrão ou exemplar bem definida.

Recursos

AWS documentação

- Documentação do Amazon Bedrock
- Documentação do Amazon Neptune
- Documentação OpenSearch do Amazon Service
- Aplicação do AWS Well-Architected Framework para o Amazon Neptune (orientação prescritiva)AWS
- Melhores práticas operacionais para o Amazon OpenSearch Service (documentação OpenSearch do serviço)
- Usando o Amazon Comprehend Medical e LLMs para saúde e ciências biológicas (orientação prescritiva)AWS

AWS postagens no blog

- Crie aplicativos de IA generativa baseados em agentes e RAG com o novo modelo Amazon Titan Text Premier, disponível no Amazon Bedrock
- Complemente a inteligência comercial criando um gráfico de conhecimento a partir de um data warehouse com o Amazon Neptune
- Usando gráficos de conhecimento para criar aplicativos GraphRag com o Amazon Bedrock e o Amazon Neptune

Outros recursos

- Integrando a geração aumentada de recuperação com grandes modelos de linguagem em nefrologia: aplicações práticas avançadas (Central, National Library of Medicine) PubMed
- Introdução ao LangChain (LangChain documentação)

Colaboradores

Autoria

- Nitu Nivedita, Diretora Executiva — Líder de Inteligência Artificial, Dados e IA, Accenture
- Manoj Appully, fundador e CTO da Cadiem
- Conor Folan, consultor — Dados e IA, Accenture
- Deepak Krishna AR, consultor — Dados e IA, Accenture
- Almore Cato, gerente de dados e IA, Accenture
- Soonam Kurian, arquiteta principal de soluções, AWS

Analisando

- Sally Lin, gerente sênior de ciência de dados — Dados e IA, Accenture
- Terry Huang, gerente de ciência de dados — Dados e IA, Accenture
- William Lorenz, arquiteto de soluções da Partners, AWS

Redação técnica

- Lilly AbouHarb, redatora técnica sênior, AWS

Histórico do documento

A tabela a seguir descreve alterações significativas feitas neste guia. Se desejar receber notificações sobre futuras atualizações, inscreva-se em um feed RSS.

Alteração	Descrição	Data
Publicação inicial	—	14 de março de 2025

AWS Glossário de orientação prescritiva

A seguir estão os termos comumente usados em estratégias, guias e padrões fornecidos pela Orientação AWS Prescritiva. Para sugerir entradas, use o link Fornecer feedback no final do glossário.

Números

7 Rs

Sete estratégias comuns de migração para mover aplicações para a nuvem. Essas estratégias baseiam-se nos 5 Rs identificados pela Gartner em 2011 e consistem em:

- Refatorar/rearquitetar: mova uma aplicação e modifique sua arquitetura aproveitando ao máximo os recursos nativos de nuvem para melhorar a agilidade, a performance e a escalabilidade. Isso normalmente envolve a portabilidade do sistema operacional e do banco de dados. Exemplo: migre seu banco de dados Oracle local para a edição compatível com o Amazon Aurora PostgreSQL.
- Redefinir a plataforma (mover e redefinir [mover e redefinir (lift-and-reshape]): mova uma aplicação para a nuvem e introduza algum nível de otimização a fim de aproveitar os recursos da nuvem. Exemplo: Migre seu banco de dados Oracle local para o Amazon Relational Database Service (Amazon RDS) for Oracle no. Nuvem AWS
- Recomprar (drop and shop): mude para um produto diferente, normalmente migrando de uma licença tradicional para um modelo SaaS. Exemplo: migre seu sistema de gerenciamento de relacionamento com o cliente (CRM) para a Salesforce.com.
- Redefinir a hospedagem (mover sem alterações [lift-and-shift])mover uma aplicação para a nuvem sem fazer nenhuma alteração a fim de aproveitar os recursos da nuvem. Exemplo: Migre seu banco de dados Oracle local para o Oracle em uma EC2 instância no. Nuvem AWS
- Realocar (mover o hipervisor sem alterações [hypervisor-level lift-and-shift]): mover a infraestrutura para a nuvem sem comprar novo hardware, reescrever aplicações ou modificar suas operações existentes. Você migra servidores de uma plataforma local para um serviço em nuvem para a mesma plataforma. Exemplo: Migrar um Microsoft Hyper-V aplicativo para o. AWS
- Reter (revisitar): mantenha as aplicações em seu ambiente de origem. Isso pode incluir aplicações que exigem grande refatoração, e você deseja adiar esse trabalho para um

momento posterior, e aplicações antigas que você deseja manter porque não há justificativa comercial para migrá-las.

- Retirar: desative ou remova aplicações que não são mais necessárias em seu ambiente de origem.

A

ABAC

Consulte controle de acesso baseado em atributos.

serviços abstratos

Veja os serviços gerenciados.

ACID

Veja atomicidade, consistência, isolamento, durabilidade.

migração ativa-ativa

Um método de migração de banco de dados no qual os bancos de dados de origem e de destino são mantidos em sincronia (por meio de uma ferramenta de replicação bidirecional ou operações de gravação dupla), e ambos os bancos de dados lidam com transações de aplicações conectadas durante a migração. Esse método oferece suporte à migração em lotes pequenos e controlados, em vez de exigir uma substituição única. É mais flexível, mas exige mais trabalho do que a migração ativa-passiva.

migração ativa-passiva

Um método de migração de banco de dados no qual os bancos de dados de origem e de destino são mantidos em sincronia, mas somente o banco de dados de origem manipula as transações das aplicações conectadas enquanto os dados são replicados no banco de dados de destino. O banco de dados de destino não aceita nenhuma transação durante a migração.

função agregada

Uma função SQL que opera em um grupo de linhas e calcula um único valor de retorno para o grupo. Exemplos de funções agregadas incluem `SUM` e. `MAX`

AI

Veja a inteligência artificial.

AIOps

Veja as operações de inteligência artificial.

anonimização

O processo de excluir permanentemente informações pessoais em um conjunto de dados. A anonimização pode ajudar a proteger a privacidade pessoal. Dados anônimos não são mais considerados dados pessoais.

antipadrões

Uma solução frequentemente usada para um problema recorrente em que a solução é contraproducente, ineficaz ou menos eficaz do que uma alternativa.

controle de aplicativos

Uma abordagem de segurança que permite o uso somente de aplicativos aprovados para ajudar a proteger um sistema contra malware.

portfólio de aplicações

Uma coleção de informações detalhadas sobre cada aplicação usada por uma organização, incluindo o custo para criar e manter a aplicação e seu valor comercial. Essas informações são fundamentais para o processo de descoberta e análise de portfólio e ajudam a identificar e priorizar as aplicações a serem migradas, modernizadas e otimizadas.

inteligência artificial (IA)

O campo da ciência da computação que se dedica ao uso de tecnologias de computação para desempenhar funções cognitivas normalmente associadas aos humanos, como aprender, resolver problemas e reconhecer padrões. Para obter mais informações, consulte O que é inteligência artificial?

operações de inteligência artificial (AIOps)

O processo de usar técnicas de machine learning para resolver problemas operacionais, reduzir incidentes operacionais e intervenção humana e aumentar a qualidade do serviço. Para obter mais informações sobre como AIOps é usado na estratégia de AWS migração, consulte o guia de integração de operações.

criptografia assimétrica

Um algoritmo de criptografia que usa um par de chaves, uma chave pública para criptografia e uma chave privada para descriptografia. É possível compartilhar a chave pública porque ela não é usada na descriptografia, mas o acesso à chave privada deve ser altamente restrito.

atomicidade, consistência, isolamento, durabilidade (ACID)

Um conjunto de propriedades de software que garantem a validade dos dados e a confiabilidade operacional de um banco de dados, mesmo no caso de erros, falhas de energia ou outros problemas.

controle de acesso por atributo (ABAC)

A prática de criar permissões minuciosas com base nos atributos do usuário, como departamento, cargo e nome da equipe. Para obter mais informações, consulte ABAC AWS na documentação AWS Identity and Access Management (IAM).

fonte de dados autorizada

Um local onde você armazena a versão principal dos dados, que é considerada a fonte de informações mais confiável. Você pode copiar dados da fonte de dados autorizada para outros locais com o objetivo de processar ou modificar os dados, como anonimizá-los, redigi-los ou pseudonimizá-los.

Zona de disponibilidade

Um local distinto dentro de um Região da AWS que está isolado de falhas em outras zonas de disponibilidade e fornece conectividade de rede barata e de baixa latência a outras zonas de disponibilidade na mesma região.

AWS Estrutura de adoção da nuvem (AWS CAF)

Uma estrutura de diretrizes e melhores práticas AWS para ajudar as organizações a desenvolver um plano eficiente e eficaz para migrar com sucesso para a nuvem. AWS O CAF organiza a orientação em seis áreas de foco chamadas perspectivas: negócios, pessoas, governança, plataforma, segurança e operações. As perspectivas de negócios, pessoas e governança têm como foco habilidades e processos de negócios; as perspectivas de plataforma, segurança e operações concentram-se em habilidades e processos técnicos. Por exemplo, a perspectiva das pessoas tem como alvo as partes interessadas que lidam com recursos humanos (RH), funções de pessoal e gerenciamento de pessoal. Nessa perspectiva, o AWS CAF fornece orientação para desenvolvimento, treinamento e comunicação de pessoas para ajudar a preparar a organização

para a adoção bem-sucedida da nuvem. Para obter mais informações, consulte o site da AWS CAF e o whitepaper da AWS CAF.

AWS Estrutura de qualificação da carga de trabalho (AWS WQF)

Uma ferramenta que avalia as cargas de trabalho de migração do banco de dados, recomenda estratégias de migração e fornece estimativas de trabalho. AWS O WQF está incluído com AWS Schema Conversion Tool ()AWS SCT. Ela analisa esquemas de banco de dados e objetos de código, código de aplicações, dependências e características de performance, além de fornecer relatórios de avaliação.

B

bot ruim

Um bot destinado a perturbar ou causar danos a indivíduos ou organizações.

BCP

Veja o planejamento de continuidade de negócios.

gráfico de comportamento

Uma visualização unificada e interativa do comportamento e das interações de recursos ao longo do tempo. É possível usar um gráfico de comportamento com o Amazon Detective para examinar tentativas de login malsucedidas, chamadas de API suspeitas e ações similares. Para obter mais informações, consulte Dados em um gráfico de comportamento na documentação do Detective.

sistema big-endian

Um sistema que armazena o byte mais significativo antes. Veja também endianness.

classificação binária

Um processo que prevê um resultado binário (uma de duas classes possíveis). Por exemplo, seu modelo de ML pode precisar prever problemas como “Este e-mail é ou não é spam?” ou “Este produto é um livro ou um carro?”

filtro de bloom

Uma estrutura de dados probabilística e eficiente em termos de memória que é usada para testar se um elemento é membro de um conjunto.

blue/green deployment (implantação azul/verde)

Uma estratégia de implantação em que você cria dois ambientes separados, mas idênticos. Você executa a versão atual do aplicativo em um ambiente (azul) e a nova versão do aplicativo no outro ambiente (verde). Essa estratégia ajuda você a reverter rapidamente com o mínimo de impacto.

bot

Um aplicativo de software que executa tarefas automatizadas pela Internet e simula a atividade ou interação humana. Alguns bots são úteis ou benéficos, como rastreadores da Web que indexam informações na Internet. Alguns outros bots, conhecidos como bots ruins, têm como objetivo perturbar ou causar danos a indivíduos ou organizações.

botnet

Redes de bots infectadas por malware e sob o controle de uma única parte, conhecidas como pastor de bots ou operador de bots. As redes de bots são o mecanismo mais conhecido para escalar bots e seu impacto.

ramo

Uma área contida de um repositório de código. A primeira ramificação criada em um repositório é a ramificação principal. Você pode criar uma nova ramificação a partir de uma ramificação existente e, em seguida, desenvolver recursos ou corrigir bugs na nova ramificação. Uma ramificação que você cria para gerar um recurso é comumente chamada de ramificação de recurso. Quando o recurso estiver pronto para lançamento, você mesclará a ramificação do recurso de volta com a ramificação principal. Para obter mais informações, consulte Sobre filiais (GitHub documentação).

acesso em vidro quebrado

Em circunstâncias excepcionais e por meio de um processo aprovado, um meio rápido para um usuário obter acesso a um Conta da AWS que ele normalmente não tem permissão para acessar. Para obter mais informações, consulte o indicador Implementar procedimentos de quebra de vidro na orientação do Well-Architected AWS .

estratégia brownfield

A infraestrutura existente em seu ambiente. Ao adotar uma estratégia brownfield para uma arquitetura de sistema, você desenvolve a arquitetura de acordo com as restrições dos sistemas e da infraestrutura atuais. Se estiver expandindo a infraestrutura existente, poderá combinar as estratégias brownfield e greenfield.

cache do buffer

A área da memória em que os dados acessados com mais frequência são armazenados.

capacidade de negócios

O que uma empresa faz para gerar valor (por exemplo, vendas, atendimento ao cliente ou marketing). As arquiteturas de microsserviços e as decisões de desenvolvimento podem ser orientadas por recursos de negócios. Para obter mais informações, consulte a seção Organizados de acordo com as capacidades de negócios do whitepaper Executar microsserviços conteinerizados na AWS.

planejamento de continuidade de negócios (BCP)

Um plano que aborda o impacto potencial de um evento disruptivo, como uma migração em grande escala, nas operações e permite que uma empresa retome as operações rapidamente.

C

CAF

Consulte Estrutura de adoção da AWS nuvem.

implantação canária

O lançamento lento e incremental de uma versão para usuários finais. Quando estiver confiante, você implanta a nova versão e substituirá a versão atual em sua totalidade.

CCoE

Veja o Centro de Excelência em Nuvem.

CDC

Veja a captura de dados de alterações.

captura de dados de alterações (CDC)

O processo de rastrear alterações em uma fonte de dados, como uma tabela de banco de dados, e registrar metadados sobre a alteração. É possível usar o CDC para várias finalidades, como auditar ou replicar alterações em um sistema de destino para manter a sincronização.

engenharia do caos

Introduzir intencionalmente falhas ou eventos disruptivos para testar a resiliência de um sistema. Você pode usar AWS Fault Injection Service (AWS FIS) para realizar experimentos que estressam suas AWS cargas de trabalho e avaliar sua resposta.

CI/CD

Veja a integração e a entrega contínuas.

classificação

Um processo de categorização que ajuda a gerar previsões. Os modelos de ML para problemas de classificação predizem um valor discreto. Os valores discretos são sempre diferentes uns dos outros. Por exemplo, um modelo pode precisar avaliar se há ou não um carro em uma imagem.

criptografia no lado do cliente

Criptografia de dados localmente, antes que o alvo os AWS service (Serviço da AWS) receba.

Centro de excelência em nuvem (CCoE)

Uma equipe multidisciplinar que impulsiona os esforços de adoção da nuvem em toda a organização, incluindo o desenvolvimento de práticas recomendas de nuvem, a mobilização de recursos, o estabelecimento de cronogramas de migração e a liderança da organização em transformações em grande escala. Para obter mais informações, consulte as publicações CCo E no Blog de Estratégia Nuvem AWS Empresarial.

computação em nuvem

A tecnologia de nuvem normalmente usada para armazenamento de dados remoto e gerenciamento de dispositivos de IoT. A computação em nuvem geralmente está conectada à tecnologia de computação de ponta.

modelo operacional em nuvem

Em uma organização de TI, o modelo operacional usado para criar, amadurecer e otimizar um ou mais ambientes de nuvem. Para obter mais informações, consulte Criar seu modelo operacional de nuvem.

estágios de adoção da nuvem

As quatro fases pelas quais as organizações normalmente passam quando migram para o Nuvem AWS:

- Projeto: executar alguns projetos relacionados à nuvem para fins de prova de conceito e aprendizado
- Fundação — Fazer investimentos fundamentais para escalar sua adoção da nuvem (por exemplo, criar uma landing zone, definir um CCo E, estabelecer um modelo de operações)
- Migração: migrar aplicações individuais
- Reinvenção: otimizar produtos e serviços e inovar na nuvem

Esses estágios foram definidos por Stephen Orban na postagem do blog The Journey Toward Cloud-First & the Stages of Adoption no blog de estratégia Nuvem AWS empresarial. Para obter informações sobre como eles se relacionam com a estratégia de AWS migração, consulte o guia de preparação para migração.

CMDB

Consulte o banco de dados de gerenciamento de configuração.

repositório de código

Um local onde o código-fonte e outros ativos, como documentação, amostras e scripts, são armazenados e atualizados por meio de processos de controle de versão. Os repositórios de nuvem comuns incluem GitHub ouBitbucket Cloud. Cada versão do código é chamada de ramificação. Em uma estrutura de microsserviços, cada repositório é dedicado a uma única peça de funcionalidade. Um único pipeline de CI/CD pode usar vários repositórios.

cache frio

Um cache de buffer que está vazio, não está bem preenchido ou contém dados obsoletos ou irrelevantes. Isso afeta a performance porque a instância do banco de dados deve ler da memória principal ou do disco, um processo que é mais lento do que a leitura do cache do buffer.

dados frios

Dados que raramente são acessados e geralmente são históricos. Ao consultar esse tipo de dados, consultas lentas geralmente são aceitáveis. Mover esses dados para níveis ou classes de armazenamento de baixo desempenho e menos caros pode reduzir os custos.

visão computacional (CV)

Um campo da IA que usa aprendizado de máquina para analisar e extrair informações de formatos visuais, como imagens e vídeos digitais. Por exemplo, a Amazon SageMaker AI fornece algoritmos de processamento de imagem para CV.

desvio de configuração

Para uma carga de trabalho, uma alteração de configuração em relação ao estado esperado. Isso pode fazer com que a carga de trabalho se torne incompatível e, normalmente, é gradual e não intencional.

banco de dados de gerenciamento de configuração (CMDB)

Um repositório que armazena e gerencia informações sobre um banco de dados e seu ambiente de TI, incluindo componentes de hardware e software e suas configurações. Normalmente, os dados de um CMDB são usados no estágio de descoberta e análise do portfólio da migração.

pacote de conformidade

Um conjunto de AWS Config regras e ações de remediação que você pode montar para personalizar suas verificações de conformidade e segurança. Você pode implantar um pacote de conformidade como uma entidade única em uma Conta da AWS região ou em uma organização usando um modelo YAML. Para obter mais informações, consulte Pacotes de conformidade na documentação. AWS Config

integração contínua e entrega contínua (CI/CD)

O processo de automatizar os estágios de origem, criação, teste, preparação e produção do processo de lançamento do software. CI/CD is commonly described as a pipeline. CI/CDpode ajudá-lo a automatizar processos, melhorar a produtividade, melhorar a qualidade do código e entregar com mais rapidez. Para obter mais informações, consulte Benefícios da entrega contínua. CD também pode significar implantação contínua. Para obter mais informações, consulte Entrega contínua versus implantação contínua.

CV

Veja visão computacional.

D

dados em repouso

Dados estacionários em sua rede, por exemplo, dados que estão em um armazenamento.

classificação de dados

Um processo para identificar e categorizar os dados em sua rede com base em criticalidade e confidencialidade. É um componente crítico de qualquer estratégia de gerenciamento de riscos de

segurança cibernética, pois ajuda a determinar os controles adequados de proteção e retenção para os dados. A classificação de dados é um componente do pilar de segurança no AWS Well-Architected Framework. Para obter mais informações, consulte Classificação de dados.

desvio de dados

Uma variação significativa entre os dados de produção e os dados usados para treinar um modelo de ML ou uma alteração significativa nos dados de entrada ao longo do tempo. O desvio de dados pode reduzir a qualidade geral, a precisão e a imparcialidade das previsões do modelo de ML.

dados em trânsito

Dados que estão se movendo ativamente pela sua rede, como entre os recursos da rede.

malha de dados

Uma estrutura arquitetônica que fornece propriedade de dados distribuída e descentralizada com gerenciamento e governança centralizados.

minimização de dados

O princípio de coletar e processar apenas os dados estritamente necessários. Praticar a minimização de dados no Nuvem AWS pode reduzir os riscos de privacidade, os custos e a pegada de carbono de sua análise.

perímetro de dados

Um conjunto de proteções preventivas em seu AWS ambiente que ajudam a garantir que somente identidades confiáveis acessem recursos confiáveis das redes esperadas. Para obter mais informações, consulte Construindo um perímetro de dados em. AWS

pré-processamento de dados

A transformação de dados brutos em um formato que seja facilmente analisado por seu modelo de ML. O pré-processamento de dados pode significar a remoção de determinadas colunas ou linhas e o tratamento de valores ausentes, inconsistentes ou duplicados.

proveniência dos dados

O processo de rastrear a origem e o histórico dos dados ao longo de seu ciclo de vida, por exemplo, como os dados foram gerados, transmitidos e armazenados.

titular dos dados

Um indivíduo cujos dados estão sendo coletados e processados.

data warehouse

Um sistema de gerenciamento de dados que oferece suporte à inteligência comercial, como análises. Os data warehouses geralmente contêm grandes quantidades de dados históricos e geralmente são usados para consultas e análises.

linguagem de definição de dados (DDL)

Instruções ou comandos para criar ou modificar a estrutura de tabelas e objetos em um banco de dados.

linguagem de manipulação de dados (DML)

Instruções ou comandos para modificar (inserir, atualizar e excluir) informações em um banco de dados.

DDL

Consulte a linguagem de definição de banco de dados.

deep ensemble

A combinação de vários modelos de aprendizado profundo para gerar previsões. Os deep ensembles podem ser usados para produzir uma previsão mais precisa ou para estimar a incerteza nas previsões.

Aprendizado profundo

Um subcampo do ML que usa várias camadas de redes neurais artificiais para identificar o mapeamento entre os dados de entrada e as variáveis-alvo de interesse.

defense-in-depth

Uma abordagem de segurança da informação na qual uma série de mecanismos e controles de segurança são cuidadosamente distribuídos por toda a rede de computadores para proteger a confidencialidade, a integridade e a disponibilidade da rede e dos dados nela contidos. Ao adotar essa estratégia AWS, você adiciona vários controles em diferentes camadas da AWS Organizations estrutura para ajudar a proteger os recursos. Por exemplo, uma defense-in-depth abordagem pode combinar autenticação multifatorial, segmentação de rede e criptografia.

administrador delegado

Em AWS Organizations, um serviço compatível pode registrar uma conta de AWS membro para administrar as contas da organização e gerenciar as permissões desse serviço. Essa conta

é chamada de administrador delegado para esse serviço Para obter mais informações e uma lista de serviços compatíveis, consulte Serviços que funcionam com o AWS Organizations na documentação do AWS Organizations .

implantação

O processo de criar uma aplicação, novos recursos ou correções de código disponíveis no ambiente de destino. A implantação envolve a implementação de mudanças em uma base de código e, em seguida, a criação e execução dessa base de código nos ambientes da aplicação

ambiente de desenvolvimento

Veja o ambiente.

controle detectivo

Um controle de segurança projetado para detectar, registrar e alertar após a ocorrência de um evento. Esses controles são uma segunda linha de defesa, alertando você sobre eventos de segurança que contornaram os controles preventivos em vigor. Para obter mais informações, consulte Controles detectivos em Como implementar controles de segurança na AWS.

mapeamento do fluxo de valor de desenvolvimento (DVSM)

Um processo usado para identificar e priorizar restrições que afetam negativamente a velocidade e a qualidade em um ciclo de vida de desenvolvimento de software. O DVSM estende o processo de mapeamento do fluxo de valor originalmente projetado para práticas de manufatura enxuta. Ele se concentra nas etapas e equipes necessárias para criar e movimentar valor por meio do processo de desenvolvimento de software.

gêmeo digital

Uma representação virtual de um sistema real, como um prédio, fábrica, equipamento industrial ou linha de produção. Os gêmeos digitais oferecem suporte à manutenção preditiva, ao monitoramento remoto e à otimização da produção.

tabela de dimensões

Em um esquema em estrela, uma tabela menor que contém atributos de dados sobre dados quantitativos em uma tabela de fatos. Os atributos da tabela de dimensões geralmente são campos de texto ou números discretos que se comportam como texto. Esses atributos são comumente usados para restringir consultas, filtrar e rotular conjuntos de resultados.

desastre

Um evento que impede que uma workload ou sistema cumpra seus objetivos de negócios em seu local principal de implantação. Esses eventos podem ser desastres naturais, falhas técnicas ou o resultado de ações humanas, como configuração incorreta não intencional ou ataque de malware.

Recuperação de desastres (RD)

A estratégia e o processo que você usa para minimizar o tempo de inatividade e a perda de dados causados por um desastre. Para obter mais informações, consulte Recuperação de desastres de cargas de trabalho em AWS: Recuperação na nuvem no AWS Well-Architected Framework.

DML

Veja a linguagem de manipulação de banco de dados.

design orientado por domínio

Uma abordagem ao desenvolvimento de um sistema de software complexo conectando seus componentes aos domínios em evolução, ou principais metas de negócios, atendidos por cada componente. Esse conceito foi introduzido por Eric Evans em seu livro, Design orientado por domínio: lidando com a complexidade no coração do software (Boston: Addison-Wesley Professional, 2003). Para obter informações sobre como usar o design orientado por domínio com o padrão strangler fig, consulte Modernizar incrementalmente os serviços web herdados do Microsoft ASP.NET (ASMX) usando contêineres e o Amazon API Gateway.

DR

Veja a recuperação de desastres.

detecção de deriva

Rastreando desvios de uma configuração básica. Por exemplo, você pode usar AWS CloudFormation para detectar desvios nos recursos do sistema ou AWS Control Tower para detectar mudanças em seu landing zone que possam afetar a conformidade com os requisitos de governança.

DVSM

Veja o mapeamento do fluxo de valor do desenvolvimento.

E

EDA

Veja a análise exploratória de dados.

EDI

Veja intercâmbio eletrônico de dados.

computação de borda

A tecnologia que aumenta o poder computacional de dispositivos inteligentes nas bordas de uma rede de IoT. Quando comparada à computação em nuvem, a computação de ponta pode reduzir a latência da comunicação e melhorar o tempo de resposta.

intercâmbio eletrônico de dados (EDI)

A troca automatizada de documentos comerciais entre organizações. Para obter mais informações, consulte O que é intercâmbio eletrônico de dados.

Criptografia

Um processo de computação que transforma dados de texto simples, legíveis por humanos, em texto cifrado.

chave de criptografia

Uma sequência criptográfica de bits aleatórios que é gerada por um algoritmo de criptografia. As chaves podem variar em tamanho, e cada chave foi projetada para ser imprevisível e exclusiva.

endianismo

A ordem na qual os bytes são armazenados na memória do computador. Os sistemas big-endian armazenam o byte mais significativo antes. Os sistemas little-endian armazenam o byte menos significativo antes.

endpoint

Veja o endpoint do serviço.

serviço de endpoint

Um serviço que pode ser hospedado em uma nuvem privada virtual (VPC) para ser compartilhado com outros usuários. Você pode criar um serviço de endpoint com AWS PrivateLink e conceder permissões a outros diretores Contas da AWS ou a AWS Identity and Access Management (IAM).

Essas contas ou entidades principais podem se conectar ao serviço de endpoint de maneira privada criando endpoints da VPC de interface. Para obter mais informações, consulte Criar um serviço de endpoint na documentação do Amazon Virtual Private Cloud (Amazon VPC).

planejamento de recursos corporativos (ERP)

Um sistema que automatiza e gerencia os principais processos de negócios (como contabilidade, MES e gerenciamento de projetos) para uma empresa.

criptografia envelopada

O processo de criptografar uma chave de criptografia com outra chave de criptografia. Para obter mais informações, consulte Criptografia de envelope na documentação AWS Key Management Service (AWS KMS).

ambiente

Uma instância de uma aplicação em execução. Estes são tipos comuns de ambientes na computação em nuvem:

- ambiente de desenvolvimento: uma instância de uma aplicação em execução que está disponível somente para a equipe principal responsável pela manutenção da aplicação. Ambientes de desenvolvimento são usados para testar mudanças antes de promovê-las para ambientes superiores. Esse tipo de ambiente às vezes é chamado de ambiente de teste.
- ambientes inferiores: todos os ambientes de desenvolvimento para uma aplicação, como aqueles usados para compilações e testes iniciais.
- ambiente de produção: uma instância de uma aplicação em execução que os usuários finais podem acessar. Em um pipeline de CI/CD, o ambiente de produção é o último ambiente de implantação.
- ambientes superiores: todos os ambientes que podem ser acessados por usuários que não sejam a equipe principal de desenvolvimento. Isso pode incluir um ambiente de produção, ambientes de pré-produção e ambientes para testes de aceitação do usuário.

epic

Em metodologias ágeis, categorias funcionais que ajudam a organizar e priorizar seu trabalho. Os epics fornecem uma descrição de alto nível dos requisitos e das tarefas de implementação. Por exemplo, os épicos de segurança AWS da CAF incluem gerenciamento de identidade e acesso, controles de detetive, segurança de infraestrutura, proteção de dados e resposta a incidentes. Para obter mais informações sobre epics na estratégia de migração da AWS , consulte o guia de implementação do programa.

ERP

Veja o planejamento de recursos corporativos.

análise exploratória de dados (EDA)

O processo de analisar um conjunto de dados para entender suas principais características. Você coleta ou agrega dados e, em seguida, realiza investigações iniciais para encontrar padrões, detectar anomalias e verificar suposições. O EDA é realizado por meio do cálculo de estatísticas resumidas e da criação de visualizações de dados.

F

tabela de fatos

A tabela central em um esquema em estrela. Ele armazena dados quantitativos sobre as operações comerciais. Normalmente, uma tabela de fatos contém dois tipos de colunas: aquelas que contêm medidas e aquelas que contêm uma chave externa para uma tabela de dimensões.

falham rapidamente

Uma filosofia que usa testes frequentes e incrementais para reduzir o ciclo de vida do desenvolvimento. É uma parte essencial de uma abordagem ágil.

limite de isolamento de falhas

No Nuvem AWS, um limite, como uma zona de disponibilidade, Região da AWS um plano de controle ou um plano de dados, que limita o efeito de uma falha e ajuda a melhorar a resiliência das cargas de trabalho. Para obter mais informações, consulte Limites de isolamento de AWS falhas.

ramificação de recursos

Veja a filial.

recursos

Os dados de entrada usados para fazer uma previsão. Por exemplo, em um contexto de manufatura, os recursos podem ser imagens capturadas periodicamente na linha de fabricação.

importância do recurso

O quanto um recurso é importante para as previsões de um modelo. Isso geralmente é expresso como uma pontuação numérica que pode ser calculada por meio de várias técnicas, como

Shapley Additive Explanations (SHAP) e gradientes integrados. Para obter mais informações, consulte Interpretabilidade do modelo de aprendizado de máquina com AWS.

transformação de recursos

O processo de otimizar dados para o processo de ML, incluindo enriquecer dados com fontes adicionais, escalar valores ou extrair vários conjuntos de informações de um único campo de dados. Isso permite que o modelo de ML se beneficie dos dados. Por exemplo, se a data “2021-05-27 00:15:37” for dividida em “2021”, “maio”, “quinta” e “15”, isso poderá ajudar o algoritmo de aprendizado a aprender padrões diferenciados associados a diferentes componentes de dados.

solicitação rápida

Fornecer a um LLM um pequeno número de exemplos que demonstram a tarefa e o resultado desejado antes de solicitar que ele execute uma tarefa semelhante. Essa técnica é uma aplicação do aprendizado contextual, em que os modelos aprendem com exemplos (fotos) incorporados aos prompts. Solicitações rápidas podem ser eficazes para tarefas que exigem formatação, raciocínio ou conhecimento de domínio específicos. Veja também a solicitação zero-shot.

FGAC

Veja o controle de acesso refinado.

Controle de acesso refinado (FGAC)

O uso de várias condições para permitir ou negar uma solicitação de acesso.

migração flash-cut

Um método de migração de banco de dados que usa replicação contínua de dados por meio da captura de dados alterados para migrar dados no menor tempo possível, em vez de usar uma abordagem em fases. O objetivo é reduzir ao mínimo o tempo de inatividade.

FM

Veja o modelo da fundação.

modelo de fundação (FM)

Uma grande rede neural de aprendizado profundo que vem treinando em grandes conjuntos de dados generalizados e não rotulados. FMs são capazes de realizar uma ampla variedade de tarefas gerais, como entender a linguagem, gerar texto e imagens e conversar em linguagem natural. Para obter mais informações, consulte O que são modelos básicos.

G

IA generativa

Um subconjunto de modelos de IA que foram treinados em grandes quantidades de dados e que podem usar uma simples solicitação de texto para criar novos conteúdos e artefatos, como imagens, vídeos, texto e áudio. Para obter mais informações, consulte O que é IA generativa.

bloqueio geográfico

Veja as restrições geográficas.

restrições geográficas (bloqueio geográfico)

Na Amazon CloudFront, uma opção para impedir que usuários em países específicos acessem distribuições de conteúdo. É possível usar uma lista de permissões ou uma lista de bloqueios para especificar países aprovados e banidos. Para obter mais informações, consulte Restringir a distribuição geográfica do seu conteúdo na CloudFront documentação.

Fluxo de trabalho do GitFlow

Uma abordagem na qual ambientes inferiores e superiores usam ramificações diferentes em um repositório de código-fonte. O fluxo de trabalho do Gitflow é considerado legado, e o fluxo de trabalho baseado em troncos é a abordagem moderna e preferida.

imagem dourada

Um instantâneo de um sistema ou software usado como modelo para implantar novas instâncias desse sistema ou software. Por exemplo, na manufatura, uma imagem dourada pode ser usada para provisionar software em vários dispositivos e ajudar a melhorar a velocidade, a escalabilidade e a produtividade nas operações de fabricação de dispositivos.

estratégia greenfield

A ausência de infraestrutura existente em um novo ambiente. Ao adotar uma estratégia greenfield para uma arquitetura de sistema, é possível selecionar todas as novas tecnologias sem a restrição da compatibilidade com a infraestrutura existente, também conhecida como brownfield. Se estiver expandindo a infraestrutura existente, poderá combinar as estratégias brownfield e greenfield.

barreira de proteção

Uma regra de alto nível que ajuda a governar recursos, políticas e conformidade em todas as unidades organizacionais ()OUs. Barreiras de proteção preventivas impõem políticas para

garantir o alinhamento a padrões de conformidade. Elas são implementadas usando políticas de controle de serviço e limites de permissões do IAM. Barreiras de proteção detectivas detectam violações de políticas e problemas de conformidade e geram alertas para remediação. Eles são implementados usando AWS Config, AWS Security Hub, Amazon GuardDuty AWS Trusted Advisor, Amazon Inspector e verificações personalizadas AWS Lambda .

H

HA

Veja a alta disponibilidade.

migração heterogênea de bancos de dados

Migrar seu banco de dados de origem para um banco de dados de destino que usa um mecanismo de banco de dados diferente (por exemplo, Oracle para Amazon Aurora). A migração heterogênea geralmente faz parte de um esforço de redefinição da arquitetura, e converter o esquema pode ser uma tarefa complexa. O AWS fornece o AWS SCT para ajudar nas conversões de esquemas.

alta disponibilidade (HA)

A capacidade de uma workload operar continuamente, sem intervenção, em caso de desafios ou desastres. Os sistemas AH são projetados para realizar o failover automático, oferecer consistentemente desempenho de alta qualidade e lidar com diferentes cargas e falhas com impacto mínimo no desempenho.

modernização de historiador

Uma abordagem usada para modernizar e atualizar os sistemas de tecnologia operacional (OT) para melhor atender às necessidades do setor de manufatura. Um historiador é um tipo de banco de dados usado para coletar e armazenar dados de várias fontes em uma fábrica.

dados de retenção

Uma parte dos dados históricos rotulados que são retidos de um conjunto de dados usado para treinar um modelo de aprendizado de máquina. Você pode usar dados de retenção para avaliar o desempenho do modelo comparando as previsões do modelo com os dados de retenção.

migração homogênea de bancos de dados

Migrar seu banco de dados de origem para um banco de dados de destino que compartilha o mesmo mecanismo de banco de dados (por exemplo, Microsoft SQL Server para Amazon RDS para SQL Server). A migração homogênea geralmente faz parte de um esforço de redefinição da hospedagem ou da plataforma. É possível usar utilitários de banco de dados nativos para migrar o esquema.

dados quentes

Dados acessados com frequência, como dados em tempo real ou dados translacionais recentes. Esses dados normalmente exigem uma camada ou classe de armazenamento de alto desempenho para fornecer respostas rápidas às consultas.

hotfix

Uma correção urgente para um problema crítico em um ambiente de produção. Devido à sua urgência, um hotfix geralmente é feito fora do fluxo de trabalho típico de uma DevOps versão.

período de hipercuidados

Imediatamente após a substituição, o período em que uma equipe de migração gerencia e monitora as aplicações migradas na nuvem para resolver quaisquer problemas. Normalmente, a duração desse período é de 1 a 4 dias. No final do período de hipercuidados, a equipe de migração normalmente transfere a responsabilidade pelas aplicações para a equipe de operações de nuvem.

eu

IaC

Veja a infraestrutura como código.

Política baseada em identidade

Uma política anexada a um ou mais diretores do IAM que define suas permissões no Nuvem AWS ambiente.

aplicação ociosa

Uma aplicação que tem um uso médio de CPU e memória entre 5 e 20% em um período de 90 dias. Em um projeto de migração, é comum retirar essas aplicações ou retê-las on-premises.

IIoT

Veja a Internet das Coisas industrial.

infraestrutura imutável

Um modelo que implanta uma nova infraestrutura para cargas de trabalho de produção em vez de atualizar, corrigir ou modificar a infraestrutura existente. Infraestruturas imutáveis são inerentemente mais consistentes, confiáveis e previsíveis do que infraestruturas mutáveis. Para obter mais informações, consulte as melhores práticas de implantação usando infraestrutura imutável no Well-Architected AWS Framework.

VPC de entrada (admissão)

Em uma arquitetura de AWS várias contas, uma VPC que aceita, inspeciona e roteia conexões de rede de fora de um aplicativo. A Arquitetura de Referência de AWS Segurança recomenda configurar sua conta de rede com entrada, saída e inspeção VPCs para proteger a interface bidirecional entre seu aplicativo e a Internet em geral.

migração incremental

Uma estratégia de substituição na qual você migra a aplicação em pequenas partes, em vez de realizar uma única substituição completa. Por exemplo, é possível mover inicialmente apenas alguns microsserviços ou usuários para o novo sistema. Depois de verificar se tudo está funcionando corretamente, mova os microsserviços ou usuários adicionais de forma incremental até poder descomissionar seu sistema herdado. Essa estratégia reduz os riscos associados a migrações de grande porte.

Indústria 4.0

Um termo que foi introduzido por Klaus Schwab em 2016 para se referir à modernização dos processos de fabricação por meio de avanços em conectividade, dados em tempo real, automação, análise e IA/ML.

infraestrutura

Todos os recursos e ativos contidos no ambiente de uma aplicação.

Infraestrutura como código (IaC)

O processo de provisionamento e gerenciamento da infraestrutura de uma aplicação por meio de um conjunto de arquivos de configuração. A IaC foi projetada para ajudar você a centralizar o gerenciamento da infraestrutura, padronizar recursos e escalar rapidamente para que novos ambientes sejam reproduzíveis, confiáveis e consistentes.

Internet industrial das coisas (IIoT)

O uso de sensores e dispositivos conectados à Internet nos setores industriais, como manufatura, energia, automotivo, saúde, ciências biológicas e agricultura. Para obter mais informações, consulte Criando uma estratégia de transformação digital industrial da Internet das Coisas (IIoT).

VPC de inspeção

Em uma arquitetura de AWS várias contas, uma VPC centralizada que gerencia as inspeções do tráfego de rede entre VPCs (na mesma ou em diferentes Regiões da AWS) a Internet e as redes locais. A Arquitetura de Referência de AWS Segurança recomenda configurar sua conta de rede com entrada, saída e inspeção VPCs para proteger a interface bidirecional entre seu aplicativo e a Internet em geral.

Internet das Coisas (IoT)

A rede de objetos físicos conectados com sensores ou processadores incorporados que se comunicam com outros dispositivos e sistemas pela Internet ou por uma rede de comunicação local. Para obter mais informações, consulte O que é IoT?

interpretabilidade

Uma característica de um modelo de machine learning que descreve o grau em que um ser humano pode entender como as previsões do modelo dependem de suas entradas. Para obter mais informações, consulte Interpretabilidade do modelo de aprendizado de máquina com AWS.

IoT

Consulte Internet das Coisas.

Biblioteca de informações de TI (ITIL)

Um conjunto de práticas recomendadas para fornecer serviços de TI e alinhar esses serviços a requisitos de negócios. A ITIL fornece a base para o ITSM.

Gerenciamento de serviços de TI (ITSM)

Atividades associadas a design, implementação, gerenciamento e suporte de serviços de TI para uma organização. Para obter informações sobre a integração de operações em nuvem com ferramentas de ITSM, consulte o guia de integração de operações.

ITIL

Consulte a biblioteca de informações de TI.

ITSM

Veja o gerenciamento de serviços de TI.

L

controle de acesso baseado em etiqueta (LBAC)

Uma implementação do controle de acesso obrigatório (MAC) em que os usuários e os dados em si recebem explicitamente um valor de etiqueta de segurança. A interseção entre a etiqueta de segurança do usuário e a etiqueta de segurança dos dados determina quais linhas e colunas podem ser vistas pelo usuário.

zona de pouso

Uma landing zone é um AWS ambiente bem arquitetado, com várias contas, escalável e seguro. Um ponto a partir do qual suas organizações podem iniciar e implantar rapidamente workloads e aplicações com confiança em seu ambiente de segurança e infraestrutura. Para obter mais informações sobre zonas de pouso, consulte Configurar um ambiente da AWS com várias contas seguro e escalável.

modelo de linguagem grande (LLM)

Um modelo de IA de aprendizado profundo que é pré-treinado em uma grande quantidade de dados. Um LLM pode realizar várias tarefas, como responder perguntas, resumir documentos, traduzir texto para outros idiomas e completar frases. Para obter mais informações, consulte O que são LLMs.

migração de grande porte

Uma migração de 300 servidores ou mais.

LBAC

Veja controle de acesso baseado em etiquetas.

privilégio mínimo

A prática recomendada de segurança de conceder as permissões mínimas necessárias para executar uma tarefa. Para obter mais informações, consulte Aplicar permissões de privilégios mínimos na documentação do IAM.

mover sem alterações (lift-and-shift)

Veja 7 Rs.

sistema little-endian

Um sistema que armazena o byte menos significativo antes. Veja também endianness.

LLM

Veja um modelo de linguagem grande.

ambientes inferiores

Veja o ambiente.

M

machine learning (ML)

Um tipo de inteligência artificial que usa algoritmos e técnicas para reconhecimento e aprendizado de padrões. O ML analisa e aprende com dados gravados, por exemplo, dados da Internet das Coisas (IoT), para gerar um modelo estatístico baseado em padrões. Para obter mais informações, consulte Machine learning.

ramificação principal

Veja a filial.

malware

Software projetado para comprometer a segurança ou a privacidade do computador. O malware pode interromper os sistemas do computador, vazar informações confidenciais ou obter acesso não autorizado. Exemplos de malware incluem vírus, worms, ransomware, cavalos de Tróia, spyware e keyloggers.

serviços gerenciados

Serviços da AWS para o qual AWS opera a camada de infraestrutura, o sistema operacional e as plataformas, e você acessa os endpoints para armazenar e recuperar dados. O Amazon Simple Storage Service (Amazon S3) e o Amazon DynamoDB são exemplos de serviços gerenciados. Eles também são conhecidos como serviços abstratos.

sistema de execução de manufatura (MES)

Um sistema de software para rastrear, monitorar, documentar e controlar processos de produção que convertem matérias-primas em produtos acabados no chão de fábrica.

MAP

Consulte Migration Acceleration Program.

mecanismo

Um processo completo no qual você cria uma ferramenta, impulsiona a adoção da ferramenta e, em seguida, inspeciona os resultados para fazer ajustes. Um mecanismo é um ciclo que se reforça e se aprimora à medida que opera. Para obter mais informações, consulte Construindo mecanismos no AWS Well-Architected Framework.

conta de membro

Todos, Contas da AWS exceto a conta de gerenciamento, que fazem parte de uma organização em AWS Organizations. Uma conta só pode ser membro de uma organização de cada vez.

MES

Veja o sistema de execução de manufatura.

Transporte de telemetria de enfileiramento de mensagens (MQTT)

Um protocolo de comunicação leve machine-to-machine (M2M), baseado no padrão de publicação/assinatura, para dispositivos de IoT com recursos limitados.

microsserviço

Um serviço pequeno e independente que se comunica de forma bem definida APIs e normalmente é de propriedade de equipes pequenas e independentes. Por exemplo, um sistema de seguradora pode incluir microsserviços que mapeiam as capacidades comerciais, como vendas ou marketing, ou subdomínios, como compras, reclamações ou análises. Os benefícios dos microsserviços incluem agilidade, escalabilidade flexível, fácil implantação, código reutilizável e resiliência. Para obter mais informações, consulte Integração de microsserviços usando serviços sem AWS servidor.

arquitetura de microsserviços

Uma abordagem à criação de aplicações com componentes independentes que executam cada processo de aplicação como um microsserviço. Esses microsserviços se comunicam por meio

de uma interface bem definida usando leveza. APIs Cada microsserviço nessa arquitetura pode ser atualizado, implantado e escalado para atender à demanda por funções específicas de uma aplicação. Para obter mais informações, consulte Implementação de microsserviços em. AWS

Programa de Aceleração da Migração (MAP)

Um AWS programa que fornece suporte de consultoria, treinamento e serviços para ajudar as organizações a criar uma base operacional sólida para migrar para a nuvem e ajudar a compensar o custo inicial das migrações. O MAP inclui uma metodologia de migração para executar migrações legadas de forma metódica e um conjunto de ferramentas para automatizar e acelerar cenários comuns de migração.

migração em escala

O processo de mover a maior parte do portfólio de aplicações para a nuvem em ondas, com mais aplicações sendo movidas em um ritmo mais rápido a cada onda. Essa fase usa as práticas recomendadas e lições aprendidas nas fases anteriores para implementar uma fábrica de migração de equipes, ferramentas e processos para agilizar a migração de workloads por meio de automação e entrega ágeis. Esta é a terceira fase da estratégia de migração para a AWS.

fábrica de migração

Equipes multifuncionais que simplificam a migração de workloads por meio de abordagens automatizadas e ágeis. As equipes da fábrica de migração geralmente incluem operações, analistas e proprietários de negócios, engenheiros de migração, desenvolvedores e DevOps profissionais que trabalham em sprints. Entre 20 e 50% de um portfólio de aplicações corporativas consiste em padrões repetidos que podem ser otimizados por meio de uma abordagem de fábrica. Para obter mais informações, consulte discussão sobre fábricas de migração e o guia do Cloud Migration Factory neste conjunto de conteúdo.

metadados de migração

As informações sobre a aplicação e o servidor necessárias para concluir a migração. Cada padrão de migração exige um conjunto de metadados de migração diferente. Exemplos de metadados de migração incluem a sub-rede, o grupo de segurança e AWS a conta de destino.

padrão de migração

Uma tarefa de migração repetível que detalha a estratégia de migração, o destino da migração e a aplicação ou o serviço de migração usado. Exemplo: rehospede a migração para a Amazon EC2 com o AWS Application Migration Service.

Avaliação de Portfólio para Migração (MPA)

Uma ferramenta on-line que fornece informações para validar o caso de negócios para migrar para o. Nuvem AWS O MPA fornece avaliação detalhada do portfólio (dimensionamento correto do servidor, preços, comparações de TCO, análise de custos de migração), bem como planejamento de migração (análise e coleta de dados de aplicações, agrupamento de aplicações, priorização de migração e planejamento de ondas). A ferramenta MPA (requer login) está disponível gratuitamente para todos os AWS consultores e consultores parceiros da APN.

Avaliação de Preparação para Migração (MRA)

O processo de obter insights sobre o status de prontidão de uma organização para a nuvem, identificar pontos fortes e fracos e criar um plano de ação para fechar as lacunas identificadas, usando o CAF. AWS Para mais informações, consulte o guia de preparação para migração. A MRA é a primeira fase da estratégia de migração para a AWS.

estratégia de migração

A abordagem usada para migrar uma carga de trabalho para o. Nuvem AWS Para obter mais informações, consulte a entrada de 7 Rs neste glossário e consulte Mobilize sua organização para acelerar migrações em grande escala.

ML

Veja o aprendizado de máquina.

modernização

Transformar uma aplicação desatualizada (herdada ou monolítica) e sua infraestrutura em um sistema ágil, elástico e altamente disponível na nuvem para reduzir custos, ganhar eficiência e aproveitar as inovações. Para obter mais informações, consulte Estratégia para modernizar aplicativos no Nuvem AWS.

avaliação de preparação para modernização

Uma avaliação que ajuda a determinar a preparação para modernização das aplicações de uma organização. Ela identifica benefícios, riscos e dependências e determina o quão bem a organização pode acomodar o estado futuro dessas aplicações. O resultado da avaliação é um esquema da arquitetura de destino, um roteiro que detalha as fases de desenvolvimento e os marcos do processo de modernização e um plano de ação para abordar as lacunas identificadas. Para obter mais informações, consulte Avaliação da prontidão para modernização de aplicativos no. Nuvem AWS

aplicações monolíticas (monólitos)

Aplicações que são executadas como um único serviço com processos fortemente acoplados. As aplicações monolíticas apresentam várias desvantagens. Se um recurso da aplicação apresentar um aumento na demanda, toda a arquitetura deverá ser escalada. Adicionar ou melhorar os recursos de uma aplicação monolítica também se torna mais complexo quando a base de código cresce. Para resolver esses problemas, é possível criar uma arquitetura de microsserviços. Para obter mais informações, consulte Decompor monólitos em microsserviços.

MAPA

Consulte Avaliação do portfólio de migração.

MQTT

Consulte Transporte de telemetria de enfileiramento de mensagens.

classificação multiclasse

Um processo que ajuda a gerar previsões para várias classes (prevendo um ou mais de dois resultados). Por exemplo, um modelo de ML pode perguntar “Este produto é um livro, um carro ou um telefone?” ou “Qual categoria de produtos é mais interessante para este cliente?”

infraestrutura mutável

Um modelo que atualiza e modifica a infraestrutura existente para cargas de trabalho de produção. Para melhorar a consistência, confiabilidade e previsibilidade, o AWS Well-Architected Framework recomenda o uso de infraestrutura imutável como uma prática recomendada.

O

OAC

Veja o controle de acesso de origem.

CARVALHO

Veja a identidade de acesso de origem.

OCM

Veja o gerenciamento de mudanças organizacionais.

migração offline

Um método de migração no qual a workload de origem é desativada durante o processo de migração. Esse método envolve tempo de inatividade prolongado e geralmente é usado para workloads pequenas e não críticas.

OI

Veja a integração de operações.

OLA

Veja o contrato em nível operacional.

migração online

Um método de migração no qual a workload de origem é copiada para o sistema de destino sem ser colocada offline. As aplicações conectadas à workload podem continuar funcionando durante a migração. Esse método envolve um tempo de inatividade nulo ou mínimo e normalmente é usado para workloads essenciais para a produção.

OPC-UA

Consulte Comunicação de processo aberto — Arquitetura unificada.

Comunicação de processo aberto — Arquitetura unificada (OPC-UA)

Um protocolo de comunicação machine-to-machine (M2M) para automação industrial. O OPC-UA fornece um padrão de interoperabilidade com esquemas de criptografia, autenticação e autorização de dados.

acordo de nível operacional (OLA)

Um acordo que esclarece o que os grupos funcionais de TI prometem oferecer uns aos outros para apoiar um acordo de serviço (SLA).

análise de prontidão operacional (ORR)

Uma lista de verificação de perguntas e melhores práticas associadas que ajudam você a entender, avaliar, prevenir ou reduzir o escopo de incidentes e possíveis falhas. Para obter mais informações, consulte Operational Readiness Reviews (ORR) no Well-Architected AWS Framework.

tecnologia operacional (OT)

Sistemas de hardware e software que funcionam com o ambiente físico para controlar operações, equipamentos e infraestrutura industriais. Na manufatura, a integração dos sistemas OT e de tecnologia da informação (TI) é o foco principal das transformações da Indústria 4.0.

integração de operações (OI)

O processo de modernização das operações na nuvem, que envolve planejamento de preparação, automação e integração. Para obter mais informações, consulte o guia de integração de operações.

trilha organizacional

Uma trilha criada por ela AWS CloudTrail registra todos os eventos de todos Contas da AWS em uma organização em AWS Organizations. Essa trilha é criada em cada Conta da AWS que faz parte da organização e monitora a atividade em cada conta. Para obter mais informações, consulte Criação de uma trilha para uma organização na CloudTrail documentação.

gerenciamento de alterações organizacionais (OCM)

Uma estrutura para gerenciar grandes transformações de negócios disruptivas de uma perspectiva de pessoas, cultura e liderança. O OCM ajuda as organizações a se prepararem e fazerem a transição para novos sistemas e estratégias, acelerando a adoção de alterações, abordando questões de transição e promovendo mudanças culturais e organizacionais. Na estratégia de AWS migração, essa estrutura é chamada de aceleração de pessoas, devido à velocidade de mudança exigida nos projetos de adoção da nuvem. Para obter mais informações, consulte o guia do OCM.

controle de acesso de origem (OAC)

Em CloudFront, uma opção aprimorada para restringir o acesso para proteger seu conteúdo do Amazon Simple Storage Service (Amazon S3). O OAC oferece suporte a todos os buckets S3 Regiões da AWS, criptografia do lado do servidor com AWS KMS (SSE-KMS) e solicitações dinâmicas ao bucket S3. `PUT` `DELETE`

Identidade do acesso de origem (OAI)

Em CloudFront, uma opção para restringir o acesso para proteger seu conteúdo do Amazon S3. Quando você usa o OAI, CloudFront cria um principal com o qual o Amazon S3 pode se autenticar. Os diretores autenticados podem acessar o conteúdo em um bucket do S3 somente por meio de uma distribuição específica. CloudFront Veja também OAC, que fornece um controle de acesso mais granular e aprimorado.

ORR

Veja a análise de prontidão operacional.

OT

Veja a tecnologia operacional.

VPC de saída (egresso)

Em uma arquitetura de AWS várias contas, uma VPC que gerencia conexões de rede que são iniciadas de dentro de um aplicativo. A Arquitetura de Referência de AWS Segurança recomenda configurar sua conta de rede com entrada, saída e inspeção VPCs para proteger a interface bidirecional entre seu aplicativo e a Internet em geral.

P

limite de permissões

Uma política de gerenciamento do IAM anexada a entidades principais do IAM para definir as permissões máximas que o usuário ou perfil podem ter. Para obter mais informações, consulte Limites de permissões na documentação do IAM.

Informações de identificação pessoal (PII)

Informações que, quando visualizadas diretamente ou combinadas com outros dados relacionados, podem ser usadas para inferir razoavelmente a identidade de um indivíduo. Exemplos de PII incluem nomes, endereços e informações de contato.

PII

Veja as informações de identificação pessoal.

manual

Um conjunto de etapas predefinidas que capturam o trabalho associado às migrações, como a entrega das principais funções operacionais na nuvem. Um manual pode assumir a forma de scripts, runbooks automatizados ou um resumo dos processos ou etapas necessários para operar seu ambiente modernizado.

PLC

Consulte controlador lógico programável.

AMEIXA

Veja o gerenciamento do ciclo de vida do produto.

política

Um objeto que pode definir permissões (consulte a política baseada em identidade), especificar as condições de acesso (consulte a política baseada em recursos) ou definir as permissões máximas para todas as contas em uma organização em AWS Organizations (consulte a política de controle de serviços).

persistência poliglota

Escolher de forma independente a tecnologia de armazenamento de dados de um microsserviço com base em padrões de acesso a dados e outros requisitos. Se seus microsserviços tiverem a mesma tecnologia de armazenamento de dados, eles poderão enfrentar desafios de implementação ou apresentar baixa performance. Os microsserviços serão implementados com mais facilidade e alcançarão performance e escalabilidade melhores se usarem o armazenamento de dados mais bem adaptado às suas necessidades. Para obter mais informações, consulte Habilitar a persistência de dados em microsserviços.

avaliação do portfólio

Um processo de descobrir, analisar e priorizar o portfólio de aplicações para planejar a migração. Para obter mais informações, consulte Avaliar a preparação para a migração.

predicado

Uma condição de consulta que retorna `true` ou`false`, normalmente localizada em uma `WHERE` cláusula.

pressão de predicados

Uma técnica de otimização de consulta de banco de dados que filtra os dados na consulta antes da transferência. Isso reduz a quantidade de dados que devem ser recuperados e processados do banco de dados relacional e melhora o desempenho das consultas.

controle preventivo

Um controle de segurança projetado para evitar que um evento ocorra. Esses controles são a primeira linha de defesa para ajudar a evitar acesso não autorizado ou alterações indesejadas em sua rede. Para obter mais informações, consulte Controles preventivos em Como implementar controles de segurança na AWS.

principal (entidade principal)

Uma entidade AWS que pode realizar ações e acessar recursos. Essa entidade geralmente é um usuário raiz para um Conta da AWS, uma função do IAM ou um usuário. Para obter mais informações, consulte Entidade principal em Termos e conceitos de perfis na documentação do IAM.

privacidade por design

Uma abordagem de engenharia de sistema que leva em consideração a privacidade em todo o processo de desenvolvimento.

zonas hospedadas privadas

Um contêiner que contém informações sobre como você deseja que o Amazon Route 53 responda às consultas de DNS para um domínio e seus subdomínios em um ou mais. VPCs Para obter mais informações, consulte Como trabalhar com zonas hospedadas privadas na documentação do Route 53.

controle proativo

Um controle de segurança projetado para impedir a implantação de recursos não compatíveis. Esses controles examinam os recursos antes de serem provisionados. Se o recurso não estiver em conformidade com o controle, ele não será provisionado. Para obter mais informações, consulte o guia de referência de controles na AWS Control Tower documentação e consulte Controles proativos em Implementação de controles de segurança em AWS.

gerenciamento do ciclo de vida do produto (PLM)

O gerenciamento de dados e processos de um produto em todo o seu ciclo de vida, desde o design, desenvolvimento e lançamento, passando pelo crescimento e maturidade, até o declínio e a remoção.

ambiente de produção

Veja o ambiente.

controlador lógico programável (PLC)

Na fabricação, um computador altamente confiável e adaptável que monitora as máquinas e automatiza os processos de fabricação.

encadeamento imediato

Usando a saída de um prompt do LLM como entrada para o próximo prompt para gerar respostas melhores. Essa técnica é usada para dividir uma tarefa complexa em subtarefas ou para refinar ou expandir iterativamente uma resposta preliminar. Isso ajuda a melhorar a precisão e a relevância das respostas de um modelo e permite resultados mais granulares e personalizados.

pseudonimização

O processo de substituir identificadores pessoais em um conjunto de dados por valores de espaço reservado. A pseudonimização pode ajudar a proteger a privacidade pessoal. Os dados pseudonimizados ainda são considerados dados pessoais.

publish/subscribe (pub/sub)

Um padrão que permite comunicações assíncronas entre microsserviços para melhorar a escalabilidade e a capacidade de resposta. Por exemplo, em um MES baseado em microsserviços, um microsserviço pode publicar mensagens de eventos em um canal no qual outros microsserviços possam se inscrever. O sistema pode adicionar novos microsserviços sem alterar o serviço de publicação.

Q

plano de consulta

Uma série de etapas, como instruções, usadas para acessar os dados em um sistema de banco de dados relacional SQL.

regressão de planos de consultas

Quando um otimizador de serviço de banco de dados escolhe um plano menos adequado do que escolhia antes de uma determinada alteração no ambiente de banco de dados ocorrer. Isso pode ser causado por alterações em estatísticas, restrições, configurações do ambiente, associações de parâmetros de consulta e atualizações do mecanismo de banco de dados.

R

Matriz RACI

Veja responsável, responsável, consultado, informado (RACI).

RAG

Consulte Geração Aumentada de Recuperação.

ransomware

Um software mal-intencionado desenvolvido para bloquear o acesso a um sistema ou dados de computador até que um pagamento seja feito.

Matriz RASCI

Veja responsável, responsável, consultado, informado (RACI).

RCAC

Veja o controle de acesso por linha e coluna.

réplica de leitura

Uma cópia de um banco de dados usada somente para leitura. É possível encaminhar consultas para a réplica de leitura e reduzir a carga no banco de dados principal.

rearquiteta

Veja 7 Rs.

objetivo de ponto de recuperação (RPO).

O máximo período de tempo aceitável desde o último ponto de recuperação de dados. Isso determina o que é considerado uma perda aceitável de dados entre o último ponto de recuperação e a interrupção do serviço.

objetivo de tempo de recuperação (RTO)

O máximo atraso aceitável entre a interrupção e a restauração do serviço.

refatorar

Veja 7 Rs.

Região

Uma coleção de AWS recursos em uma área geográfica. Cada um Região da AWS é isolado e independente dos outros para fornecer tolerância a falhas, estabilidade e resiliência. Para obter mais informações, consulte Especificar o que Regiões da AWS sua conta pode usar.

regressão

Uma técnica de ML que prevê um valor numérico. Por exemplo, para resolver o problema de “Por qual preço esta casa será vendida?” um modelo de ML pode usar um modelo de regressão linear para prever o preço de venda de uma casa com base em fatos conhecidos sobre a casa (por exemplo, a metragem quadrada).

redefinir a hospedagem

Veja 7 Rs.

versão

Em um processo de implantação, o ato de promover mudanças em um ambiente de produção.

realocar

Veja 7 Rs.

redefinir a plataforma

Veja 7 Rs.

recomprar

Veja 7 Rs.

resiliência

A capacidade de um aplicativo de resistir ou se recuperar de interrupções. Alta disponibilidade e recuperação de desastres são considerações comuns ao planejar a resiliência no. Nuvem AWS Para obter mais informações, consulte Nuvem AWS Resiliência.

política baseada em recurso

Uma política associada a um recurso, como um bucket do Amazon S3, um endpoint ou uma chave de criptografia. Esse tipo de política especifica quais entidades principais têm acesso permitido, ações válidas e quaisquer outras condições que devem ser atendidas.

matriz responsável, accountable, consultada, informada (RACI)

Uma matriz que define as funções e responsabilidades de todas as partes envolvidas nas atividades de migração e nas operações de nuvem. O nome da matriz é derivado dos tipos de responsabilidade definidos na matriz: responsável (R), responsabilizável (A), consultado (C) e informado (I). O tipo de suporte (S) é opcional. Se você incluir suporte, a matriz será chamada de matriz RASCI e, se excluir, será chamada de matriz RACI.

controle responsivo

Um controle de segurança desenvolvido para conduzir a remediação de eventos adversos ou desvios em relação à linha de base de segurança. Para obter mais informações, consulte Controles responsivos em Como implementar controles de segurança na AWS.

reter

Veja 7 Rs.

aposentar-se

Veja 7 Rs.

Geração Aumentada de Recuperação (RAG)

Uma tecnologia de IA generativa na qual um LLM faz referência a uma fonte de dados autorizada que está fora de suas fontes de dados de treinamento antes de gerar uma resposta. Por exemplo, um modelo RAG pode realizar uma pesquisa semântica na base de conhecimento ou nos dados personalizados de uma organização. Para obter mais informações, consulte O que é RAG.

alternância

O processo de atualizar periodicamente um segredo para dificultar o acesso das credenciais por um invasor.

controle de acesso por linha e coluna (RCAC)

O uso de expressões SQL básicas e flexíveis que tenham regras de acesso definidas. O RCAC consiste em permissões de linha e máscaras de coluna.

RPO

Veja o objetivo do ponto de recuperação.

RTO

Veja o objetivo do tempo de recuperação.

runbook

Um conjunto de procedimentos manuais ou automatizados necessários para realizar uma tarefa específica. Eles são normalmente criados para agilizar operações ou procedimentos repetitivos com altas taxas de erro.

S

SAML 2.0

Um padrão aberto que muitos provedores de identidade (IdPs) usam. Esse recurso permite o login único federado (SSO), para que os usuários possam fazer login AWS Management Console ou chamar as operações da AWS API sem que você precise criar um usuário no IAM para todos em sua organização. Para obter mais informações sobre a federação baseada em SAML 2.0, consulte Sobre a federação baseada em SAML 2.0 na documentação do IAM.

SCADA

Veja controle de supervisão e aquisição de dados.

SCP

Veja a política de controle de serviços.

secret

Em AWS Secrets Manager, informações confidenciais ou restritas, como uma senha ou credenciais de usuário, que você armazena de forma criptografada. Ele consiste no valor secreto e em seus metadados. O valor secreto pode ser binário, uma única string ou várias strings. Para obter mais informações, consulte O que há em um segredo do Secrets Manager? na documentação do Secrets Manager.

segurança por design

Uma abordagem de engenharia de sistemas que leva em conta a segurança em todo o processo de desenvolvimento.

controle de segurança

Uma barreira de proteção técnica ou administrativa que impede, detecta ou reduz a capacidade de uma ameaça explorar uma vulnerabilidade de segurança. Existem quatro tipos principais de controles de segurança: preventivos, detectivos, responsivos e proativos.

fortalecimento da segurança

O processo de reduzir a superfície de ataque para torná-la mais resistente a ataques. Isso pode incluir ações como remover recursos que não são mais necessários, implementar a prática recomendada de segurança de conceder privilégios mínimos ou desativar recursos desnecessários em arquivos de configuração.

sistema de gerenciamento de eventos e informações de segurança (SIEM)

Ferramentas e serviços que combinam sistemas de gerenciamento de informações de segurança (SIM) e gerenciamento de eventos de segurança (SEM). Um sistema SIEM coleta, monitora e analisa dados de servidores, redes, dispositivos e outras fontes para detectar ameaças e violações de segurança e gerar alertas.

automação de resposta de segurança

Uma ação predefinida e programada projetada para responder ou remediar automaticamente um evento de segurança. Essas automações servem como controles de segurança responsivos ou detectivos que ajudam você a implementar as melhores práticas AWS de segurança. Exemplos de ações de resposta automatizada incluem a modificação de um grupo de segurança da VPC, a correção de uma instância EC2 da Amazon ou a rotação de credenciais.

Criptografia do lado do servidor

Criptografia dos dados em seu destino, por AWS service (Serviço da AWS) quem os recebe.

política de controle de serviços (SCP)

Uma política que fornece controle centralizado sobre as permissões de todas as contas em uma organização em AWS Organizations. SCPs defina barreiras ou estabeleça limites nas ações que um administrador pode delegar a usuários ou funções. Você pode usar SCPs como listas de permissão ou listas de negação para especificar quais serviços ou ações são permitidos ou proibidos. Para obter mais informações, consulte Políticas de controle de serviço na AWS Organizations documentação.

service endpoint (endpoint de serviço)

O URL do ponto de entrada para um AWS service (Serviço da AWS). Você pode usar o endpoint para se conectar programaticamente ao serviço de destino. Para obter mais informações, consulte Endpoints do AWS service (Serviço da AWS) na Referência geral da AWS.

acordo de serviço (SLA)

Um acordo que esclarece o que uma equipe de TI promete fornecer aos clientes, como tempo de atividade e performance do serviço.

indicador de nível de serviço (SLI)

Uma medida de um aspecto de desempenho de um serviço, como taxa de erro, disponibilidade ou taxa de transferência.

objetivo de nível de serviço (SLO)

Uma métrica alvo que representa a integridade de um serviço, conforme medida por um indicador de nível de serviço.

modelo de responsabilidade compartilhada

Um modelo que descreve a responsabilidade com a qual você compartilha AWS pela segurança e conformidade na nuvem. AWS é responsável pela segurança da nuvem, enquanto você é responsável pela segurança na nuvem. Para obter mais informações, consulte o Modelo de responsabilidade compartilhada.

SIEM

Veja informações de segurança e sistema de gerenciamento de eventos.

ponto único de falha (SPOF)

Uma falha em um único componente crítico de um aplicativo que pode interromper o sistema.

SLA

Veja o contrato de nível de serviço.

ESGUIO

Veja o indicador de nível de serviço.

SLO

Veja o objetivo do nível de serviço.

split-and-seed modelo

Um padrão para escalar e acelerar projetos de modernização. À medida que novos recursos e lançamentos de produtos são definidos, a equipe principal se divide para criar novas equipes de produtos. Isso ajuda a escalar os recursos e os serviços da sua organização, melhora a produtividade do desenvolvedor e possibilita inovações rápidas. Para obter mais informações, consulte Abordagem em fases para modernizar aplicativos no. Nuvem AWS

CUSPE

Veja um único ponto de falha.

esquema de estrelas

Uma estrutura organizacional de banco de dados que usa uma grande tabela de fatos para armazenar dados transacionais ou medidos e usa uma ou mais tabelas dimensionais menores

para armazenar atributos de dados. Essa estrutura foi projetada para uso em um data warehouse ou para fins de inteligência comercial.

padrão strangler fig

Uma abordagem à modernização de sistemas monolíticos que consiste em reescrever e substituir incrementalmente a funcionalidade do sistema até que o sistema herdado possa ser desativado. Esse padrão usa a analogia de uma videira que cresce e se torna uma árvore estabelecida e, eventualmente, supera e substitui sua hospedeira. O padrão foi apresentado por Martin Fowler como forma de gerenciar riscos ao reescrever sistemas monolíticos. Para ver um exemplo de como aplicar esse padrão, consulte Modernizar incrementalmente os serviços Web herdados do Microsoft ASP.NET (ASMX) usando contêineres e o Amazon API Gateway.

sub-rede

Um intervalo de endereços IP na VPC. Cada sub-rede fica alocada em uma única zona de disponibilidade.

controle de supervisão e aquisição de dados (SCADA)

Na manufatura, um sistema que usa hardware e software para monitorar ativos físicos e operações de produção.

symmetric encryption (criptografia simétrica)

Um algoritmo de criptografia que usa a mesma chave para criptografar e descriptografar dados.

testes sintéticos

Testar um sistema de forma que simule as interações do usuário para detectar possíveis problemas ou monitorar o desempenho. Você pode usar o Amazon CloudWatch Synthetics para criar esses testes.

prompt do sistema

Uma técnica para fornecer contexto, instruções ou diretrizes a um LLM para direcionar seu comportamento. Os prompts do sistema ajudam a definir o contexto e estabelecer regras para interações com os usuários.

T

tags

Pares de valores-chave que atuam como metadados para organizar seus recursos. AWS As tags podem ajudar você a gerenciar, identificar, organizar, pesquisar e filtrar recursos. Para obter mais informações, consulte Marcar seus recursos do AWS.

variável-alvo

O valor que você está tentando prever no ML supervisionado. Ela também é conhecida como variável de resultado. Por exemplo, em uma configuração de fabricação, a variável-alvo pode ser um defeito do produto.

lista de tarefas

Uma ferramenta usada para monitorar o progresso por meio de um runbook. Uma lista de tarefas contém uma visão geral do runbook e uma lista de tarefas gerais a serem concluídas. Para cada tarefa geral, ela inclui o tempo estimado necessário, o proprietário e o progresso.

ambiente de teste

Veja o ambiente.

treinamento

O processo de fornecer dados para que seu modelo de ML aprenda. Os dados de treinamento devem conter a resposta correta. O algoritmo de aprendizado descobre padrões nos dados de treinamento que mapeiam os atributos dos dados de entrada no destino (a resposta que você deseja prever). Ele gera um modelo de ML que captura esses padrões. Você pode usar o modelo de ML para obter previsões de novos dados cujo destino você não conhece.

gateway de trânsito

Um hub de trânsito de rede que você pode usar para interconectar sua rede com VPCs a rede local. Para obter mais informações, consulte O que é um gateway de trânsito na AWS Transit Gateway documentação.

fluxo de trabalho baseado em troncos

Uma abordagem na qual os desenvolvedores criam e testam recursos localmente em uma ramificação de recursos e, em seguida, mesclam essas alterações na ramificação principal. A

ramificação principal é então criada para os ambientes de desenvolvimento, pré-produção e produção, sequencialmente.

Acesso confiável

Conceder permissões a um serviço que você especifica para realizar tarefas em sua organização AWS Organizations e em suas contas em seu nome. O serviço confiável cria um perfil vinculado ao serviço em cada conta, quando esse perfil é necessário, para realizar tarefas de gerenciamento para você. Para obter mais informações, consulte Usando AWS Organizations com outros AWS serviços na AWS Organizations documentação.

tuning (ajustar)

Alterar aspectos do processo de treinamento para melhorar a precisão do modelo de ML. Por exemplo, você pode treinar o modelo de ML gerando um conjunto de rótulos, adicionando rótulos e repetindo essas etapas várias vezes em configurações diferentes para otimizar o modelo.

equipe de duas pizzas

Uma pequena DevOps equipe que você pode alimentar com duas pizzas. Uma equipe de duas pizzas garante a melhor oportunidade possível de colaboração no desenvolvimento de software.

U

incerteza

Um conceito que se refere a informações imprecisas, incompletas ou desconhecidas que podem minar a confiabilidade dos modelos preditivos de ML. Há dois tipos de incertezas: a incerteza epistêmica é causada por dados limitados e incompletos, enquanto a incerteza aleatória é causada pelo ruído e pela aleatoriedade inerentes aos dados. Para obter mais informações, consulte o guia Como quantificar a incerteza em sistemas de aprendizado profundo.

tarefas indiferenciadas

Também conhecido como trabalho pesado, trabalho necessário para criar e operar um aplicativo, mas que não fornece valor direto ao usuário final nem oferece vantagem competitiva. Exemplos de tarefas indiferenciadas incluem aquisição, manutenção e planejamento de capacidade.

ambientes superiores

Veja o ambiente.

V

aspiração

Uma operação de manutenção de banco de dados que envolve limpeza após atualizações incrementais para recuperar armazenamento e melhorar a performance.

controle de versões

Processos e ferramentas que rastreiam mudanças, como alterações no código-fonte em um repositório.

emparelhamento da VPC

Uma conexão entre duas VPCs que permite rotear o tráfego usando endereços IP privados. Para ter mais informações, consulte O que é emparelhamento de VPC? na documentação da Amazon VPC.

Vulnerabilidade

Uma falha de software ou hardware que compromete a segurança do sistema.

W

cache quente

Um cache de buffer que contém dados atuais e relevantes que são acessados com frequência. A instância do banco de dados pode ler do cache do buffer, o que é mais rápido do que ler da memória principal ou do disco.

dados mornos

Dados acessados raramente. Ao consultar esse tipo de dados, consultas moderadamente lentas geralmente são aceitáveis.

função de janela

Uma função SQL que executa um cálculo em um grupo de linhas que se relacionam de alguma forma com o registro atual. As funções de janela são úteis para processar tarefas, como calcular uma média móvel ou acessar o valor das linhas com base na posição relativa da linha atual.

workload

Uma coleção de códigos e recursos que geram valor empresarial, como uma aplicação voltada para o cliente ou um processo de back-end.

workstreams

Grupos funcionais em um projeto de migração que são responsáveis por um conjunto específico de tarefas. Cada workstream é independente, mas oferece suporte aos outros workstreams do projeto. Por exemplo, o workstream de portfólio é responsável por priorizar aplicações, planejar ondas e coletar metadados de migração. O workstream de portfólio entrega esses ativos ao workstream de migração, que então migra os servidores e as aplicações.

MINHOCA

Veja escrever uma vez, ler muitas.

WQF

Consulte Estrutura de qualificação AWS da carga de trabalho.

escreva uma vez, leia muitas (WORM)

Um modelo de armazenamento que grava dados uma única vez e evita que os dados sejam excluídos ou modificados. Os usuários autorizados podem ler os dados quantas vezes forem necessárias, mas não podem alterá-los. Essa infraestrutura de armazenamento de dados é considerada imutável.

Z

exploração de dia zero

Um ataque, geralmente malware, que tira proveito de uma vulnerabilidade de dia zero.

vulnerabilidade de dia zero

Uma falha ou vulnerabilidade não mitigada em um sistema de produção. Os agentes de ameaças podem usar esse tipo de vulnerabilidade para atacar o sistema. Os desenvolvedores frequentemente ficam cientes da vulnerabilidade como resultado do ataque.

aviso zero-shot

Fornecer a um LLM instruções para realizar uma tarefa, mas sem exemplos (fotos) que possam ajudar a orientá-la. O LLM deve usar seu conhecimento pré-treinado para lidar com a tarefa. A

eficácia da solicitação zero depende da complexidade da tarefa e da qualidade da solicitação. Veja também a solicitação de algumas fotos.

aplicação zumbi

Uma aplicação que tem um uso médio de CPU e memória inferior a 5%. Em um projeto de migração, é comum retirar essas aplicações.

As traduções são geradas por tradução automática. Em caso de conflito entre o conteúdo da tradução e da versão original em inglês, a versão em inglês prevalecerá.