



Usando o Amazon Comprehend Medical e LLMs para saúde e ciências biológicas

AWS Orientação prescritiva



AWS Orientação prescritiva: Usando o Amazon Comprehend Medical e LLMs para saúde e ciências biológicas

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

As marcas comerciais e imagens comerciais da Amazon não podem ser usadas no contexto de nenhum produto ou serviço que não seja da Amazon, nem de qualquer maneira que possa gerar confusão entre os clientes ou que deprecie ou desprestige a Amazon. Todas as outras marcas comerciais que não pertencem à Amazon pertencem a seus respectivos proprietários, que podem ou não ser afiliados, patrocinados pela Amazon ou ter conexão com ela.

Table of Contents

Introdução	1
Visão geral	1
Público-alvo	2
Objetivos	2
Abordagens técnicas	4
Usando o Amazon Comprehend Medical	4
Capacidades	5
Casos de uso	6
Combinando o Amazon Comprehend Medical com LLMs	7
Arquitetura	8
Casos de uso	10
Práticas recomendadas	11
Engenharia rápida	12
Usando LLMs	21
Casos de uso para um LLM	22
Personalização	22
Escolhendo um LLM	24
Como escolher uma abordagem	28
Considerações sobre a maturidade dos negócios	30
Avaliando LLMs	31
Dados de treinamento e teste	31
Métricas	32
Perguntas frequentes	34
Como faço para escolher entre o Amazon Comprehend Medical e um LLM?	34
Como posso fornecer os resultados do Amazon Comprehend Medical a um LLM?	34
Quais são algumas das melhores práticas ao usar o Amazon Comprehend Medical com LLMs	34
Devo usar um LLM médico pré-treinado ou ajustar um LLM geral para meu caso de uso na área de saúde?	35
Como faço para avaliar o desempenho de tarefas médicas LLMs de PNL?	35
Quais são as vantagens e desvantagens entre soluções LLM de alta complexidade e baixa complexidade?	35
Próximas etapas	36
AWS recursos	36

Outros recursos	37
Histórico do documento	38
Glossário	39
#	39
A	40
B	43
C	45
D	48
E	53
F	55
G	57
H	58
eu	59
L	62
M	63
O	67
P	70
Q	73
R	73
S	77
T	81
U	82
V	83
W	83
Z	84
.....	lxxxvi

Usando o Amazon Comprehend Medical e LLMs para saúde e ciências biológicas

Joe King, Rajesh Sitaraman e Ross Claytor, da Amazon Web Services

Dezembro de 2024 ([histórico do documento](#))

Visão geral

O volume cada vez maior de dados médicos e a necessidade de processamento eficiente e preciso impulsionaram a adoção do processamento de [linguagem natural \(PNL\)](#) com tecnologias de inteligência artificial e aprendizado de máquina (IA/ML). Modelos classificadores pré-treinados e modelos de [linguagem ampla \(LLMs\)](#) surgiram como ferramentas poderosas para várias tarefas médicas de PNL, incluindo respostas a perguntas clínicas, resumo de relatórios e geração de insights. No entanto, o domínio de saúde e ciências biológicas apresenta desafios únicos devido à complexidade da terminologia médica, do conhecimento específico do domínio e dos requisitos regulatórios. O uso eficaz de classificadores pré-treinados ou LLMs nesse domínio requer uma abordagem bem projetada que combine os pontos fortes desses modelos com recursos e técnicas específicos do domínio.

As práticas do setor em saúde e ciências biológicas tradicionalmente se baseiam em sistemas baseados em regras, codificação manual e processos de revisão por especialistas. Esses sistemas e processos são demorados e propensos a erros. A integração das tecnologias de IA e PNL, como o [Amazon Comprehend Medical e os modelos básicos do Amazon Bedrock](#), oferece soluções eficientes e escaláveis para o processamento de dados médicos e, ao mesmo tempo, melhora a precisão e a consistência.

Este guia explora o uso do Amazon Comprehend Medical LLMs e a automação inteligente no setor de saúde. Ele descreve as melhores práticas, desafios e abordagens práticas para simplificar a codificação médica, a extração de informações do paciente e os processos de resumo de registros. Ao usar os recursos do Amazon Comprehend Medical e LLMs, as organizações de saúde podem desbloquear novos níveis de eficiência operacional, reduzir custos e potencialmente melhorar o atendimento ao paciente.

O guia detalha as considerações exclusivas do domínio da saúde, como entender a terminologia médica, usar domínios específicos e abordar as limitações dos sistemas de LLMs IA/ML. Ele fornece

um caminho de decisão abrangente para gerentes de TI, arquitetos e líderes técnicos da área de saúde avaliarem a prontidão organizacional, avaliarem as opções de implementação e usarem as ferramentas adequadas Serviços da AWS para uma automação bem-sucedida.

Seguindo as diretrizes e as melhores práticas descritas neste guia, as organizações de saúde podem aproveitar o poder das tecnologias de IA/ML enquanto navegam pelas complexidades do domínio médico. Essa abordagem apóia a conformidade com as diretrizes éticas e regulatórias e promove o uso responsável dos sistemas de IA na área da saúde. Ele foi projetado para gerar insights precisos e privados.

Público-alvo

Este guia é destinado a partes interessadas em tecnologia, arquitetos, líderes técnicos e tomadores de decisão que desejam implementar soluções de processamento de linguagem natural baseadas em IA para análise e automação de dados médicos.

Objetivos

Organizações de saúde e ciências biológicas podem atingir várias metas de negócios usando o Amazon Comprehend Medical e LLMs. Esses resultados geralmente incluem aumento da eficiência operacional, redução de custos e melhoria do atendimento ao paciente. Esta seção descreve os principais objetivos comerciais e os benefícios associados à implementação das estratégias e melhores práticas descritas neste guia.

A seguir estão alguns dos objetivos que as organizações podem alcançar implementando as diretrizes e as melhores práticas deste guia:

- Reduza o tempo de desenvolvimento — O objetivo final deste guia é reduzir o tempo de desenvolvimento com o custo, diminuir a dívida técnica e mitigar possíveis falhas do projeto devido ao POC. Ao compreender os principais serviços de IA/ML, como o Amazon Comprehend Medical, e as vantagens e limitações do uso do LLM para tarefas de saúde, as empresas podem alcançar um tempo de lançamento mais rápido no mercado e aumentar sua velocidade no cumprimento dos objetivos comerciais.
- Extraia informações para automatizar as tarefas de codificação médica — Após as visitas aos pacientes, especialistas em codificação e provedores podem extrair informações de textos médicos, como notas subjetivas, objetivas, de avaliação e planejamento (SOAP). Isso pode reduzir os esforços de documentação manual e ajudar o profissional a se concentrar nas necessidades

do paciente. Ao combinar os recursos de reconhecimento de entidades do Amazon Comprehend Medical com LLMs, as organizações podem extrair informações médicas relevantes de registros de pacientes, notas clínicas e outras fontes de dados de saúde. Isso pode minimizar os erros humanos e promover práticas consistentes.

- Resuma os registros dos pacientes e a documentação clínica — O resumo automatizado do histórico do paciente, dos planos de tratamento e dos resultados médicos pode economizar um tempo valioso para os profissionais de saúde. LLMs pode ajudar a gerar documentação clínica abrangente e estruturada. Você pode obter mais contexto com o Amazon Comprehend Medical, usar um LLM de domínio médico ou ajustar um LLM com dados médicos. Essas abordagens podem ajudar a fornecer resumos precisos e garantir que a documentação esteja de acordo com os requisitos e padrões de conformidade.
- Support decisões clínicas e atendimento ao paciente — Usando a [vinculação ontológica](#) no Amazon Comprehend Medical e LLMs usando, os provedores podem responder perguntas médicas ou buscar recomendações sobre o atendimento ao paciente. Isso capacita os profissionais de saúde a tomar decisões informadas que melhoram os resultados dos pacientes e reduzem o risco de erros médicos.

Abordagens generativas de IA e PNL para saúde e ciências biológicas

O processamento de linguagem natural (PNL) é uma tecnologia de aprendizado de máquina que dá aos computadores a capacidade de interpretar, manipular e compreender a linguagem humana. As organizações de saúde e ciências biológicas têm grandes volumes de dados dos prontuários dos pacientes. Eles podem usar o software de PNL para processar automaticamente esses dados. Por exemplo, eles podem combinar a PNL com a IA generativa para simplificar a codificação médica, extrair informações do paciente e resumir os registros.

Dependendo da tarefa de PNL que você deseja realizar, arquiteturas diferentes podem ser mais adequadas para seu caso de uso. Este guia aborda as seguintes opções generativas de IA e PNL para aplicações de saúde e ciências biológicas em: AWS

- [Usando o Amazon Comprehend Medical](#)— Saiba como usar o Amazon Comprehend Medical de forma independente, sem integrá-lo a um modelo de linguagem grande (LLM).
- [Combinando o Amazon Comprehend Medical com LLMs](#)— Saiba como combinar o Amazon Comprehend Medical com um LLM em uma arquitetura Retrieval Augment Generation (RAG).
- [Usando LLMs](#)— Saiba como usar um LLM para aplicações de saúde e ciências biológicas, usando um LLM ajustado ou uma arquitetura RAG.

Usando o Amazon Comprehend Medical

[O Amazon Comprehend Medical](#) detecta e retorna informações úteis em textos clínicos não estruturados, como anotações médicas, resumos de alta, resultados de exames e notas de casos. AWS service (Serviço da AWS) Ele usa modelos de processamento de linguagem natural (PNL) para detectar entidades. Entidades são referências textuais a informações médicas, como condições médicas, medicamentos ou informações de saúde protegidas (PHI).

Important

O Amazon Comprehend Medical não é um substituto para aconselhamento, diagnóstico ou tratamento médico profissional. O Amazon Comprehend Medical fornece pontuações de confiança que indicam o nível de confiança na precisão das entidades detectadas. Identifique o limite de confiança certo para seu caso de uso e use limites de alta confiança

em situações que exigem alta precisão. Em certos casos de uso, os resultados devem ser revisados e verificados por revisores humanos devidamente treinados. Por exemplo, o Amazon Comprehend Medical só deve ser usado em cenários de atendimento ao paciente após uma revisão que assegure a precisão e uma opinião médica confiável por profissionais médicos treinados.

Você pode acessar o Amazon Comprehend Medical por meio do AWS Management Console, do AWS Command Line Interface (AWS CLI) ou do. AWS SDKs Eles AWS SDKs estão disponíveis para várias linguagens e plataformas de programação, como Java, Python, Ruby, .NET, iOS e Android. Você pode usar o SDKs para acessar programaticamente o Amazon Comprehend Medical a partir do seu aplicativo cliente.

Esta seção analisa os principais recursos do Amazon Comprehend Medical. Ele também discute as vantagens de usar esse serviço em comparação com um modelo de linguagem grande (LLM).

Capacidades do Amazon Comprehend Medical

O Amazon Comprehend Medical oferece APIs inferência em lote e quase em tempo real. Eles APIs podem ingerir texto médico e fornecer resultados para tarefas médicas de PNL usando o reconhecimento de entidades médicas e identificando relacionamentos entre entidades. Você pode realizar análises em arquivos únicos ou em lote em vários arquivos armazenados em um bucket do Amazon Simple Storage Service (Amazon S3). O Amazon Comprehend Medical oferece as seguintes operações de API de análise de texto para detecção síncrona de entidades:

- [Detectar entidades](#) — Detecta categorias médicas gerais, como anatomia, condição médica, categoria de PHI, procedimentos e expressões temporais.
- [Detectar PHI](#) — Detecta entidades específicas, como idade, data, nome e informações pessoais semelhantes.

O Amazon Comprehend Medical também inclui várias operações de API que você pode usar para realizar análises de texto em lote em documentos clínicos. Para saber mais sobre como usar essas operações de API, consulte [Lote de análise de texto APIs](#).

Use o Amazon Comprehend Medical para detectar entidades em textos clínicos e vincular essas entidades a conceitos em ontologias médicas padronizadas, incluindo as bases de conhecimento ICD-10-CM e SNOMED CT. RxNorm Você pode realizar análises em arquivos únicos ou em lote em

documentos grandes ou em vários arquivos armazenados em um bucket do Amazon S3. O Amazon Comprehend Medical oferece a seguinte ontologia que vincula operações de API:

- [Infer ICD1 0CM](#) — A operação Infer ICD1 0CM detecta possíveis condições médicas e as vincula aos códigos da versão 2019 da Classificação Internacional de Doenças, 10ª Revisão, Modificação Clínica (CID-10-CM). Para cada possível condição médica detectada, o Amazon Comprehend Medical lista os códigos e descrições ICD-10-CM correspondentes. As condições médicas listadas nos resultados incluem uma pontuação de confiança, que indica a confiança que o Amazon Comprehend Medical tem na precisão das entidades em relação aos conceitos correspondentes nos resultados.
- [InferRxNorm](#) — A InferRxNorm operação identifica os medicamentos que estão listados no prontuário do paciente como entidades. Ele vincula entidades a identificadores de conceito (RxCUI) do RxNorm banco de dados da National Library of Medicine. Cada RxCUI é exclusivo de diferentes dosagens e formas de dosagem. Os medicamentos listados nos resultados incluem uma pontuação de confiança, que indica a confiança que o Amazon Comprehend Medical tem na precisão das entidades que correspondem aos conceitos da base de conhecimento. RxNorm O Amazon Comprehend Medical lista os principais Rx CUIs que potencialmente coincidem com cada medicamento que ele detecta em ordem decrescente com base na pontuação de confiança.
- [InfersNomeDCT](#) — A operação InfersNomeDCT identifica possíveis conceitos médicos como entidades e os vincula a códigos da versão 2021-03 da Nomenclatura Sistematizada de Medicina, Termos Clínicos (SNOMED CT). O SNOMED CT fornece um vocabulário abrangente de conceitos médicos, incluindo condições médicas e anatomia, exames médicos, tratamentos e procedimentos. Para cada ID de conceito correspondente, o Amazon Comprehend Medical retorna os cinco principais conceitos médicos, cada um com uma pontuação de confiança e informações contextuais, como características e atributos. O conceito SNOMED CT IDs pode então ser usado para estruturar dados clínicos do paciente para codificação médica, relatórios ou análises clínicas quando usado com a polihierarquia do SNOMED CT.

Para obter mais informações, consulte [Análise de texto APIs](#) e [vinculação de ontologias APIs na documentação](#) do Amazon Comprehend Medical.

Casos de uso do Amazon Comprehend Medical

Como um serviço independente, o Amazon Comprehend Medical pode abordar o caso de uso da sua organização. O Amazon Comprehend Medical pode realizar tarefas como as seguintes:

- Ajuda com a codificação médica nos prontuários dos pacientes

- Detecte dados de informações de saúde protegidas (PHI)
- Validando medicamentos, incluindo atributos como dosagem, frequência e forma

Os resultados do Amazon Comprehend Medical são digeríveis para a maioria dos consultórios médicos. No entanto, talvez seja necessário considerar alternativas se tiver limitações como as seguintes:

- Definições de entidades diferentes — Por exemplo, sua definição FREQUENCY de entidade medicamentosa pode ser diferente. Para fins de frequência, o Amazon Comprehend Medical prevê conforme necessário, mas sua organização pode usar o termo pro re nata (PRN).
- Quantidade impressionante de resultados — Por exemplo, as anotações do paciente geralmente contêm vários sintomas e palavras-chave que são mapeados para vários códigos ICD-10-CM. No entanto, várias das palavras-chave não são aplicáveis ao diagnóstico. Nesse caso, o provedor deve avaliar várias entidades do ICD-10-CM e suas pontuações de confiança, o que requer tempo de processamento manual.
- Entidades personalizadas ou tarefas de PNL — Por exemplo, os provedores podem querer extrair evidências do PRN, como coletar conforme necessário para tratar a dor. Como isso não está disponível por meio do Amazon Comprehend Medical, uma AI/ML model is warranted. A different AI/ML solução diferente é necessária se a tarefa de PNL estiver fora do reconhecimento da entidade, como resumo, resposta a perguntas e análise de sentimentos.

Combinando o Amazon Comprehend Medical com grandes modelos de linguagem

Um [estudo de 2024 realizado pela NEJM AI](#) mostrou que usar um LLM, com solicitação zero, para tarefas de codificação médica geralmente leva a um desempenho ruim. Usar o Amazon Comprehend Medical com um LLM pode ajudar a mitigar esses problemas de desempenho. Os resultados do Amazon Comprehend Medical são um contexto útil para um LLM que está realizando tarefas de PNL. Por exemplo, fornecer contexto do Amazon Comprehend Medical para o grande modelo de linguagem pode ajudar você a:

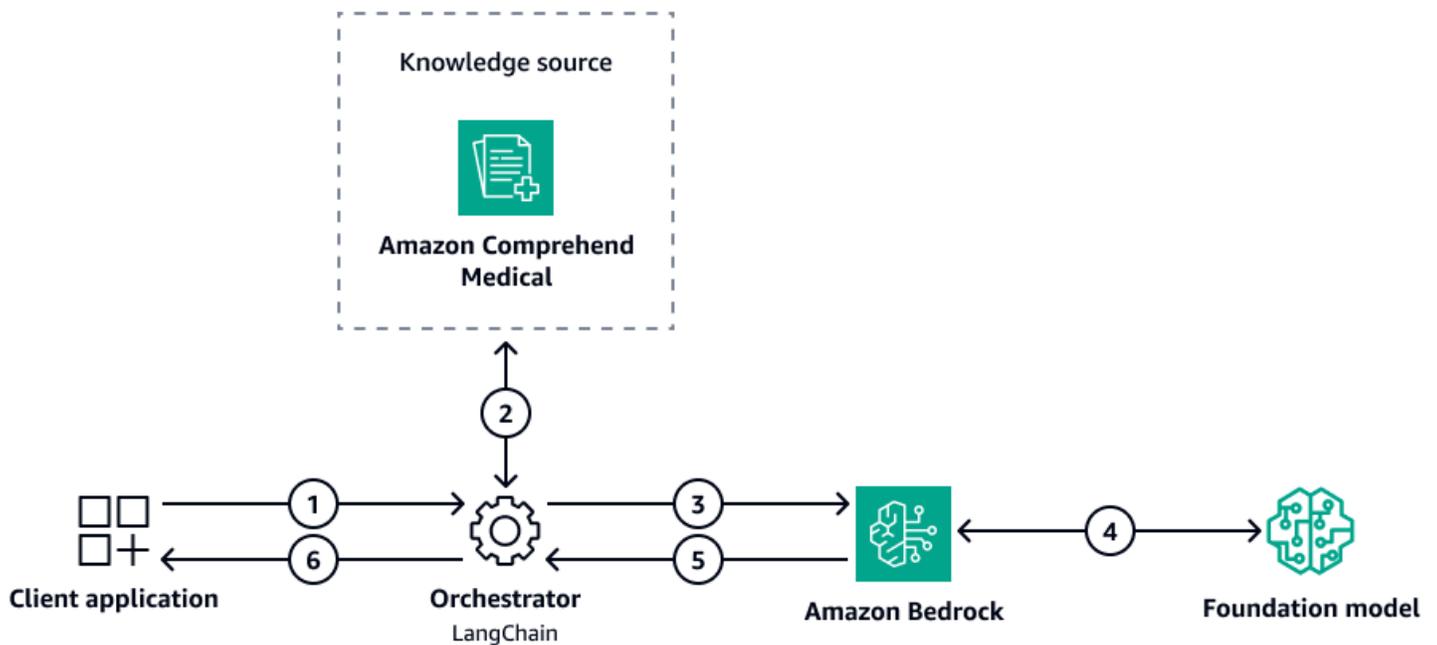
- Aumente a precisão das seleções de entidades usando os resultados iniciais do Amazon Comprehend Medical como contexto para o LLM
- Implemente reconhecimento personalizado de entidades, resumos, respostas a perguntas e casos de uso adicionais

Esta seção descreve como você pode combinar o Amazon Comprehend Medical com um LLM usando uma abordagem de geração aumentada de recuperação (RAG). 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. Para obter mais informações, consulte [O que é RAG](#).

Para ilustrar essa abordagem, esta seção usa o exemplo de codificação médica (diagnóstico) relacionada ao ICD-10-CM. Ele inclui uma arquitetura de amostra e modelos de engenharia rápidos para ajudar a acelerar sua inovação. Também inclui as melhores práticas para usar o Amazon Comprehend Medical em um fluxo de trabalho do RAG.

Arquitetura baseada em RAG com o Amazon Comprehend Medical

O diagrama a seguir ilustra uma abordagem RAG para identificar códigos de diagnóstico da CID-10-CM a partir de anotações de pacientes. Ele usa o Amazon Comprehend Medical como fonte de conhecimento. Em uma abordagem RAG, o método de recuperação geralmente recupera informações de um banco de dados vetorial contendo conhecimento aplicável. Em vez de um banco de dados vetoriais, essa arquitetura usa o Amazon Comprehend Medical para a tarefa de recuperação. O orquestrador envia as informações da nota do paciente para o Amazon Comprehend Medical e recupera as informações do código ICD-10-CM. O orquestrador envia esse contexto para o modelo de fundação downstream (LLM), por meio do Amazon Bedrock. O LLM gera uma resposta usando as informações do código ICD-10-CM, e essa resposta é enviada de volta ao aplicativo cliente.



O diagrama mostra o seguinte fluxo de trabalho do RAG:

1. O aplicativo cliente envia as anotações do paciente como uma consulta ao orquestrador. Um exemplo dessas anotações do paciente pode ser: "A paciente é uma paciente do Dr. X. A paciente se apresentou à sala de emergência na noite passada com aproximadamente 7 a 8 dias de história de dor abdominal, que tem sido persistente. Ela não teve febres ou calafrios definidos nem histórico de icterícia. O paciente nega qualquer perda significativa de peso recente."
2. O orquestrador usa o Amazon Comprehend Medical para recuperar códigos ICD-10-CM relevantes às informações médicas na consulta. Ele usa a API Infer ICD10CM para extrair e inferir os códigos ICD-10-CM das anotações do paciente.
3. O orquestrador cria um prompt que inclui o modelo de prompt, a consulta original e os códigos ICD-10-CM recuperados do Amazon Comprehend Medical. Ele envia esse contexto aprimorado para o Amazon Bedrock.
4. O Amazon Bedrock processa a entrada e usa um modelo básico para gerar uma resposta que inclui os códigos ICD-10-CM e suas evidências correspondentes da consulta. A resposta gerada inclui os códigos ICD-10-CM identificados e as evidências das anotações do paciente que apóiam cada código. Veja a seguir uma resposta de exemplo:

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

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

5. O Amazon Bedrock envia a resposta gerada ao orquestrador.
6. O orquestrador envia a resposta de volta ao aplicativo cliente, onde o usuário pode revisar a resposta.

Casos de uso para usar o Amazon Comprehend Medical em um fluxo de trabalho do RAG

O Amazon Comprehend Medical pode realizar tarefas específicas de PNL. Para obter mais informações, consulte [Casos de uso do Amazon Comprehend Medical](#).

Talvez você queira integrar o Amazon Comprehend Medical a um fluxo de trabalho do RAG para casos de uso avançados, como os seguintes:

- Gere resumos clínicos detalhados combinando entidades médicas extraídas com informações contextuais dos registros dos pacientes
- Automatize a codificação médica para casos complexos usando entidades extraídas com informações vinculadas à ontologia para atribuição de código
- Automatize a criação de notas clínicas estruturadas a partir de texto não estruturado usando entidades médicas extraídas
- Analise os efeitos colaterais dos medicamentos com base nos nomes e atributos dos medicamentos extraídos
- Desenvolva sistemas inteligentes de suporte clínico que combinem informações médicas extraídas com up-to-date pesquisas e diretrizes

Melhores práticas para usar o Amazon Comprehend Medical em um fluxo de trabalho do RAG

Ao integrar os resultados do Amazon Comprehend Medical em uma solicitação para um LLM, é essencial seguir as melhores práticas. Isso pode melhorar o desempenho e a precisão. A seguir estão as principais recomendações:

- Entenda as pontuações de confiança do Amazon Comprehend Medical — O Amazon Comprehend Medical fornece pontuações de confiança para cada entidade detectada e vinculação de ontologia. É fundamental entender o significado dessas pontuações e estabelecer limites apropriados para seu caso de uso específico. As pontuações de confiança ajudam a filtrar entidades de baixa confiança, reduzindo o ruído e melhorando a qualidade da entrada do LLM.
- Use pontuações de confiança na engenharia imediata — Ao criar solicitações para o LLM, considere incorporar as pontuações de confiança do Amazon Comprehend Medical como contexto adicional. Isso ajuda o LLM a priorizar ou avaliar entidades com base em seus níveis de confiança, melhorando potencialmente a qualidade da produção.
- Avalie os resultados do Amazon Comprehend Medical com dados reais — Os dados verdadeiros são informações que se sabe serem verdadeiras. Ele pode ser usado para validar se um aplicativo de IA/ML está produzindo resultados precisos. Antes de integrar os resultados do Amazon Comprehend Medical ao seu fluxo de trabalho de LLM, avalie o desempenho do serviço em uma amostra representativa de seus dados. Compare os resultados com anotações verdadeiras para identificar possíveis discrepâncias ou áreas de melhoria. Essa avaliação ajuda você a entender os pontos fortes e as limitações do Amazon Comprehend Medical para seu caso de uso.
- Selecione estrategicamente informações relevantes — O Amazon Comprehend Medical pode fornecer uma grande quantidade de informações, mas nem todas podem ser relevantes para sua tarefa. Selecione cuidadosamente as entidades, os atributos e os metadados que são mais relevantes para seu caso de uso. Fornecer muitas informações irrelevantes ao LLM pode introduzir ruído e potencialmente diminuir o desempenho.
- Alinhe as definições de entidades — Certifique-se de que as definições de entidades e atributos usados pelo Amazon Comprehend Medical estejam alinhadas com sua interpretação. Se houver discrepâncias, considere fornecer contexto ou esclarecimento adicional ao LLM para preencher a lacuna entre a produção do Amazon Comprehend Medical e seus requisitos. Se a entidade Amazon Comprehend Medical não atender às suas expectativas, você pode implementar a detecção personalizada de entidades incluindo instruções adicionais (e possíveis exemplos) no prompt.

- Forneça conhecimento específico do domínio — Embora o Amazon Comprehend Medical forneça informações médicas valiosas, ele pode não capturar todas as nuances do seu domínio específico. Considere complementar os resultados do Amazon Comprehend Medical com fontes adicionais de conhecimento específicas do domínio, como ontologias, terminologias ou conjuntos de dados selecionados por especialistas. Isso fornece um contexto mais abrangente para o LLM.
- Siga as diretrizes éticas e regulatórias — Ao lidar com dados médicos, é importante seguir os princípios éticos e as diretrizes regulatórias, como as relacionadas à privacidade de dados, segurança e uso responsável de sistemas de IA na área da saúde. Certifique-se de que sua implementação esteja em conformidade com as leis relevantes e as melhores práticas do setor.

Ao seguir essas melhores práticas, os profissionais de IA/ML podem usar com eficácia os pontos fortes do Amazon Comprehend Medical e LLMs Para tarefas médicas de PNL, essas melhores práticas ajudam a mitigar riscos potenciais e podem melhorar o desempenho.

Engenharia rápida para o contexto do Amazon Comprehend Medical

A [engenharia rápida](#) é o processo de projetar e refinar solicitações para orientar uma solução generativa de IA para gerar os resultados desejados. Você escolhe os formatos, frases, palavras e símbolos mais adequados que orientam a IA a interagir com seus usuários de forma mais significativa.

Dependendo da operação de API que você executa, o Amazon Comprehend Medical retorna as entidades detectadas, os códigos e descrições da ontologia e as pontuações de confiança. Esses resultados se tornam contextuais no prompt quando sua solução invoca o LLM de destino. Você deve criar a solicitação para apresentar o contexto dentro do modelo de solicitação.

Note

Os exemplos de instruções nesta seção seguem as orientações [antrópicas](#). Se você estiver usando um provedor de LLM diferente, siga as recomendações desse provedor.

Em geral, você insere o texto médico original e os resultados do Amazon Comprehend Medical no prompt. A seguir está uma estrutura de prompt comum:

```
<medical_text>  
medical text
```

```
</medical_text>

<comprehend_medical_text_results>
comprehend medical text results
</comprehend_medical_text_results>

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

Esta seção fornece estratégias para incluir os resultados do Amazon Comprehend Medical como contexto imediato para as seguintes tarefas comuns de PNL médica:

- [Filtrar resultados do Amazon Comprehend Medical](#)
- [Estenda as tarefas médicas de PNL com o Amazon Comprehend Medical](#)
- [Aplique grades de proteção com o Amazon Comprehend Medical](#)

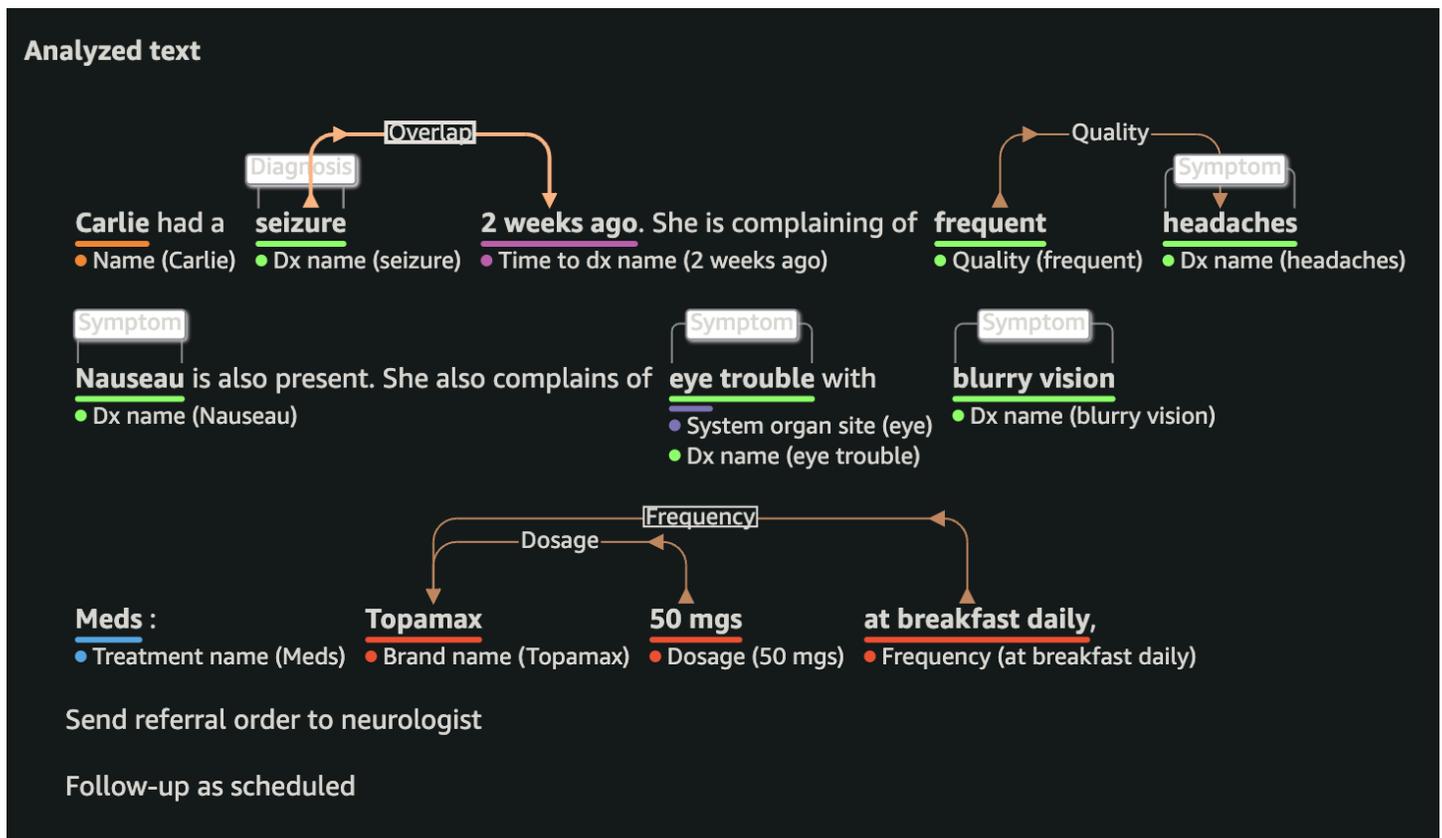
Filtrar resultados do Amazon Comprehend Medical

O Amazon Comprehend Medical normalmente fornece uma grande quantidade de informações. Talvez você queira reduzir o número de resultados que o profissional médico deve analisar. Nesse caso, você pode usar um LLM para filtrar esses resultados. As entidades do Amazon Comprehend Medical incluem uma pontuação de confiança que você pode usar como mecanismo de filtragem ao criar o prompt.

A seguir está um exemplo de nota do paciente:

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

Nesta nota do paciente, o Amazon Comprehend Medical detecta as seguintes entidades.



As entidades estão vinculadas aos seguintes códigos ICD-10-CM para convulsões e dores de cabeça.

Categoria	Código ICD-10-CM	Descrição do ICD-10-CM	Pontuação de confiança
Convulsão	R56.9	Convulsões não especificadas	0,8348
Convulsão	G40.909	Epilepsia, não especificada, não intratável, sem status epiléptico	0,5424
Convulsão	R56,00	Convulsões febris simples	0,4937
Convulsão	G40.09	Outras convulsões	0,4397

Categoria	Código ICD-10-CM	Descrição do ICD-10-CM	Pontuação de confiança
Convulsão	G40.409	Outras epilepsias generalizadas e síndromes epilépticas, não intratáveis, sem status epiléptico	0,4138
dores de cabeça	R51	dor de cabeça	0,4067
dores de cabeça	R51.9	Dor de cabeça, não especificada	0,3844
dores de cabeça	G4.52	Nova dor de cabeça persistente diária (NDPH)	0,3005
dores de cabeça	G44	Outra síndrome de dor de cabeça	0,2670
dores de cabeça	G4.8	Outras síndromes de cefaleia especificadas	0,2542

Você pode passar os códigos ICD-10-CM para o prompt para aumentar a precisão do LLM. Para reduzir o ruído, você pode filtrar os códigos ICD-10-CM usando a pontuação de confiança incluída nos resultados do Amazon Comprehend Medical. A seguir está um exemplo de solicitação que inclui somente códigos ICD-10-CM com uma pontuação de confiança superior a 0,4:

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

<comprehend_medical_results>
<icd-10>
```

```
<entity>
  <text>seizure</text>
  <code>
    <description>Unspecified convulsions</description>
    <code_value>R56.9</code_value>
    <score>0.8347607851028442</score>
  </code>
  <code>
    <description>Epilepsy, unspecified, not intractable, without status epilepticus</
description>
    <code_value>G40.909</code_value>
    <score>0.542376697063446</score>
  </code>
  <code>
    <description>Other seizures</description>
    <code_value>G40.89</code_value>
    <score>0.43966275453567505</score>
  </code>
  <code>
    <description>Other generalized epilepsy and epileptic syndromes, not intractable,
without status epilepticus</description>
    <code_value>G40.409</code_value>
    <score>0.41382506489753723</score>
  </code>
</entity>
<entity>
  <text>headaches</text>
  <code>
    <description>Headache</description>
    <code_value>R51</code_value>
    <score>0.4066613018512726</score>
  </code>
</entity>
<entity>
  <text>Nausea</text>
  <code>
    <description>Nausea</description>
    <code_value>R11.0</code_value>
    <score>0.6460834741592407</score>
  </code>
</entity>
<entity>
  <text>eye trouble</text>
```

```

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

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

```

Estenda as tarefas médicas de PNL com o Amazon Comprehend Medical

Ao processar textos médicos, o contexto do Amazon Comprehend Medical pode ajudar o LLM a selecionar melhores tokens. Neste exemplo, você deseja combinar os sintomas do diagnóstico com os medicamentos. Você também deseja encontrar textos relacionados a exames médicos, como termos relacionados a um exame de hemograma. Você pode usar o Amazon Comprehend Medical para detectar as entidades e os nomes dos medicamentos. Nesse caso, você usaria o [DetectEntitiesV2](#) e o Amazon [InferRxNorm](#) APIs Comprehend Medical.

A seguir está um exemplo de nota do paciente:

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

Para focar no código de diagnóstico, somente as entidades relacionadas `MEDICAL_CONDITION` ao tipo `DX_NAME` são usadas no prompt. Outros metadados são excluídos devido à irrelevância. Para entidades de medicamentos, o nome do medicamento junto com os atributos extraídos está incluído. Outros metadados de entidades medicamentosas do Amazon Comprehend Medical foram excluídos devido à irrelevância. Veja a seguir um exemplo de solicitação que usa resultados filtrados do Amazon Comprehend Medical. O prompt se concentra em `MEDICAL_CONDITION` entidades que têm o `DX_NAME` tipo. Esse prompt foi projetado para vincular com mais precisão os códigos de diagnóstico aos medicamentos e extrair com mais precisão os exames de pedidos médicos:

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

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

<rx_results>
<entity>
  <text>Topamax</text>
  <category>MEDICATION</category>
  <type>BRAND_NAME</type>
  <attributes>
    <attribute>
      <type>FREQUENCY</type>
      <text>at breakfast daily</text>
    </attribute>
    <attribute>
      <type>DOSAGE</type>
      <text>50 mgs</text>
    </attribute>
    <attribute>
      <type>ROUTE_OR_MODE</type>
      <text>by mouth</text>
    </attribute>
  </attributes>
</entity>
<entity>
  <text>Amoxicillan</text>
  <category>MEDICATION</category>
  <type>GENERIC_NAME</type>
  <attributes>
    <attribute>
      <type>ROUTE_OR_MODE</type>
      <text>by mouth</text>
    </attribute>
    <attribute>
      <type>DOSAGE</type>
      <text>25 mg</text>
    </attribute>
  </attributes>
</entity>
```

```
</attribute>
<attribute>
  <type>FREQUENCY</type>
  <text>twice a day</text>
</attribute>
</attributes>
</entity>
</rx_results>
```

```
<prompt>
```

Based on the patient note and the detected entities, can you please:

1. Link the diagnosis symptoms with the medications prescribed.

Provide your reasoning for the linkages.

2. Extract any entities related to medical order tests mentioned in the note.

```
</prompt>
```

Aplique grades de proteção com o Amazon Comprehend Medical

Você pode usar um LLM e o Amazon Comprehend Medical para criar grades de proteção antes que a resposta gerada seja usada. Você pode executar esse fluxo de trabalho em textos médicos não modificados ou pós-processados. Os casos de uso incluem abordar informações de saúde protegidas (PHI), detectar alucinações ou implementar políticas personalizadas para publicar resultados. Por exemplo, você pode usar o contexto do Amazon Comprehend Medical para identificar dados de PHI e, em seguida, usar o LLM para remover esses dados de PHI.

Veja a seguir um exemplo de informações de um prontuário de paciente que inclui PHI:

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

Veja a seguir um exemplo de solicitação que inclui os resultados do Amazon Comprehend Medical como contexto:

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

```
</original_text>

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

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

Usando grandes modelos de linguagem para casos de uso de saúde e ciências biológicas

Isso descreve como você pode usar modelos de linguagem grandes (LLMs) para aplicações de saúde e ciências biológicas. Alguns casos de uso exigem o uso de um grande modelo de linguagem para recursos generativos de IA. Há vantagens e limitações até mesmo para a maioria state-of-the-art LLMs, e as recomendações nesta seção foram elaboradas para ajudá-lo a atingir os resultados desejados.

Você pode usar o caminho de decisão para determinar a solução LLM apropriada para seu caso de uso, considerando fatores como conhecimento do domínio e dados de treinamento disponíveis. Além disso, esta seção discute os médicos pré-treinados populares LLMs e as melhores práticas para sua seleção e uso. Ele também discute as vantagens e desvantagens entre soluções complexas de alto desempenho e abordagens mais simples e de baixo custo.

Casos de uso para um LLM

O Amazon Comprehend Medical pode realizar tarefas específicas de PNL. Para obter mais informações, consulte [Casos de uso do Amazon Comprehend Medical](#).

Os recursos lógicos e generativos de IA de um LLM podem ser necessários para os casos de uso avançados de saúde e ciências biológicas, como os seguintes:

- Classificação de entidades médicas personalizadas ou categorias de texto
- Respondendo a perguntas clínicas
- Resumindo relatórios médicos
- Gerando e detectando insights a partir de informações médicas

Abordagens de personalização

É fundamental entender como LLMs são implementados. LLMs geralmente são treinados com bilhões de parâmetros, incluindo dados de treinamento de vários domínios. Esse treinamento permite que o LLM aborde as tarefas mais generalizadas. No entanto, muitas vezes surgem desafios quando o conhecimento específico do domínio é necessário. Exemplos de conhecimento de domínio em saúde e ciências biológicas são códigos clínicos, terminologia médica e informações de saúde necessárias para gerar respostas precisas. Portanto, usar o LLM como está (solicitação zero sem complementar o conhecimento do domínio) para esses casos de uso provavelmente resulta em resultados imprecisos. Há duas abordagens que você pode usar para superar esse desafio: Recuperação, geração aumentada e ajuste fino.

geração aumentada via recuperação

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. Um sistema RAG pode recuperar informações de ontologia médica (como classificações internacionais de doenças, arquivos nacionais de medicamentos e títulos de

assuntos médicos) de uma fonte de conhecimento. Isso fornece contexto adicional ao LLM para apoiar a tarefa médica de PNL.

Conforme discutido na [Combinando o Amazon Comprehend Medical com grandes modelos de linguagem](#) seção, você pode usar uma abordagem RAG para recuperar o contexto do Amazon Comprehend Medical. Outras fontes de conhecimento comuns incluem dados de domínio médico que são armazenados em um serviço de banco de dados, como Amazon OpenSearch Service, Amazon Kendra ou Amazon Aurora. Extrair informações dessas fontes de conhecimento pode afetar o desempenho da recuperação, especialmente com consultas semânticas que usam um banco de dados vetoriais.

Outra opção para armazenar e recuperar conhecimento específico do domínio é usar o [Amazon Q Business](#) em seu fluxo de trabalho do RAG. O Amazon Q Business pode indexar repositórios internos de documentos ou sites públicos (como [CMS.gov](#) para dados do ICD-10). O Amazon Q Business pode então extrair informações relevantes dessas fontes antes de passar sua consulta para o LLM.

Há várias maneiras de criar um fluxo de trabalho de RAG personalizado. Por exemplo, há muitas maneiras de recuperar dados de uma fonte de conhecimento. Para simplificar, recomendamos a abordagem comum de recuperação de usar um banco de dados vetoriais, como o Amazon OpenSearch Service, para armazenar conhecimento como incorporações. Isso exige que você use um modelo de incorporação, como um transformador de frases, para gerar incorporações para a consulta e para o conhecimento armazenado no banco de dados vetoriais.

Para obter mais informações sobre abordagens RAG totalmente gerenciadas e personalizadas, consulte [Opções e arquiteturas de geração aumentada de recuperação](#) em AWS

Ajuste

O ajuste fino de um modelo existente envolve usar um LLM, como um modelo Amazon Titan, Mistral ou Llama, e depois adaptar o modelo aos seus dados personalizados. Existem várias técnicas de ajuste fino, a maioria das quais envolve a modificação de apenas alguns parâmetros em vez de modificar todos os parâmetros do modelo. Isso é chamado de ajuste fino com eficiência de parâmetros (PEFT). Para obter mais informações, consulte [Hugging Face PEFT ativado](#). GitHub

A seguir estão dois casos de uso comuns em que você pode optar por ajustar um LLM para uma tarefa médica de PNL:

- Tarefa generativa — modelos baseados em decodificador realizam tarefas generativas de IA. Os profissionais de IA/ML usam dados reais básicos para ajustar um LLM existente. Por exemplo,

you can train the LLM using [MedQuAD](#), a public dataset of medical question answers. When you invoke a query for the LLM, you don't need a RAG approach to provide additional context to the LLM.

- **Incorporações** — modelos baseados em codificadores geram incorporações transformando texto em vetores numéricos. Esses modelos baseados em codificadores são normalmente chamados de modelos de incorporação. Um modelo de transformador de sentenças é um tipo específico de modelo de incorporação otimizado para sentenças. O objetivo é gerar incorporações a partir do texto de entrada. As incorporações são então usadas para análise semântica ou em tarefas de recuperação. Para ajustar o modelo de incorporação, você deve ter um conjunto de conhecimentos médicos, como documentos, que possa ser usado como dados de treinamento. Isso é feito com pares de texto baseados em semelhança ou sentimento para ajustar um modelo de transformador de frases. Para obter mais informações, consulte [Treinando e ajustando modelos de incorporação com Sentence Transformers v3 em Hugging Face](#).

You can use [Amazon SageMaker Ground Truth](#) to create a dataset of training data that is labeled and of high quality. You can use the output of the Ground Truth dataset to train your own models. You can also use the output as a dataset of training data for an Amazon SageMaker IA model. For more information about named entity recognition, text classification with a single label, and text classification with multiple labels, consult [Rotulagem de texto com Ground Truth](#) in the Amazon SageMaker AI documentation.

Escolhendo um LLM

[O Amazon Bedrock](#) is the recommended starting point to evaluate LLM performance. For more information, consult [Modelos de fundação compatíveis no Amazon Bedrock](#). You can use model evaluation workloads in Amazon Bedrock to compare the outputs of various models and, in turn, choose the model that is best suited to your use case. For more information, consult [Escolha o modelo de melhor desempenho usando as avaliações do Amazon Bedrock na documentação](#) of Amazon Bedrock.

Some LLMs have limited training on medical domain data. [Se seu caso de uso exigir o ajuste fino de um LLM ou de um LLM que o Amazon Bedrock não suporta, considere usar o Amazon AI. SageMaker](#) In SageMaker IA, you can use a LLM that is fine-tuned or choose an LLM that has been trained on medical domain data.

The following table lists popular LLMs that were trained on medical domain data.

LLM	Tarefas	Conhecimento	Arquitetura
BioBert	Recuperação de informações, classificação de texto e reconhecimento de entidade nomeada	Resumos de PubMed, artigos em texto completo e conhecimento geral do PubMedCentral domínio	Codificador
Clínica Albert	Recuperação de informações, classificação de texto e reconhecimento de entidade nomeada	Grande conjunto de dados multicêntrico, juntamente com mais de 3.000.000 de registros de pacientes de sistemas de prontuário eletrônico de saúde (EHR)	Codificador
GPT clínico	Sumarização, resposta a perguntas e geração de texto	Conjuntos de dados médicos extensos e diversos, incluindo registros médicos, conhecimento específico do domínio e consultas de diálogo em várias rodadas	Decodificador
GatorTron-GO	Sumarização, resposta a perguntas, geração de texto e recuperação de informações	Notas clínicas e literatura biomédica	Codificador
Med-bert	Recuperação de informações, classificação de texto e	Grande conjunto de dados de textos médicos, notas clínicas, trabalhos	Codificador

LLM	Tarefas	Conhecimento	Arquitetura
	reconhecimento de entidade nomeada	de pesquisa e documentos relacionados à saúde	
Palmeira vermelha	Resposta a perguntas para fins médicos	Conjuntos de dados de textos médicos e biomédicos	Decodificador
Medalha Paca	Tarefas de resposta a perguntas e diálogo médico	Uma variedade de textos médicos, abrangendo recursos como flashcards médicos, wikis e conjuntos de dados de diálogos	Decodificador
BioMedBert	Recuperação de informações, classificação de texto e reconhecimento de entidade nomeada	Exclusivamente resumos PubMed e artigos em texto completo de PubMedCentral	Codificador
BioMedLM	Sumarização, resposta a perguntas e geração de texto	Literatura biomédica a partir de fontes de PubMed conhecimento	Decodificador

A seguir estão as melhores práticas para o uso de médicos LLMs pré-treinados:

- Entenda os dados de treinamento e sua relevância para sua tarefa médica de PNL.
- Identifique a arquitetura LLM e sua finalidade. Os codificadores são apropriados para incorporações e tarefas de PNL. Os decodificadores são para tarefas de geração.
- Avalie os requisitos de infraestrutura, desempenho e custo para hospedar o LLM médico pré-treinado.

- Se for necessário um ajuste fino, garanta a veracidade ou o conhecimento precisos dos dados de treinamento. Certifique-se de mascarar ou redigir qualquer informação de identificação pessoal (PII) ou informação de saúde protegida (PHI).

As tarefas de PNL médica do mundo real podem ser diferentes das pré-treinadas LLMs em termos de conhecimento ou casos de uso pretendidos. Se um LLM específico de domínio não atender aos seus benchmarks de avaliação, você pode ajustar um LLM com seu próprio conjunto de dados ou treinar um novo modelo básico. Treinar um novo modelo de fundação é uma tarefa ambiciosa e, muitas vezes, cara. Para a maioria dos casos de uso, recomendamos o ajuste fino de um modelo existente.

Quando você usa ou ajusta um LLM médico pré-treinado, é importante abordar a infraestrutura, a segurança e as barreiras de proteção.

Infraestrutura

Em comparação com o uso do Amazon Bedrock para inferência sob demanda ou em lote, hospedar médicos pré-treinados LLMs (geralmente da Hugging Face) requer recursos significativos. Para hospedar médicos pré-treinados LLMs, é comum usar uma imagem de SageMaker IA da Amazon que é executada em uma instância do Amazon Elastic Compute Cloud EC2 (Amazon) com uma ou mais GPUs, como instâncias ml.g5 para computação acelerada ou instâncias ml.inf2 para. AWS Inferentia Isso ocorre porque LLMs consome uma grande quantidade de memória e espaço em disco.

Segurança e grades de proteção

Dependendo dos requisitos de conformidade da sua empresa, considere usar o Amazon Comprehend e o Amazon Comprehend Medical para mascarar ou redigir informações de identificação pessoal (PII) e informações de saúde protegidas (PHI) dos dados de treinamento. Isso ajuda a evitar que o LLM use dados confidenciais ao gerar respostas.

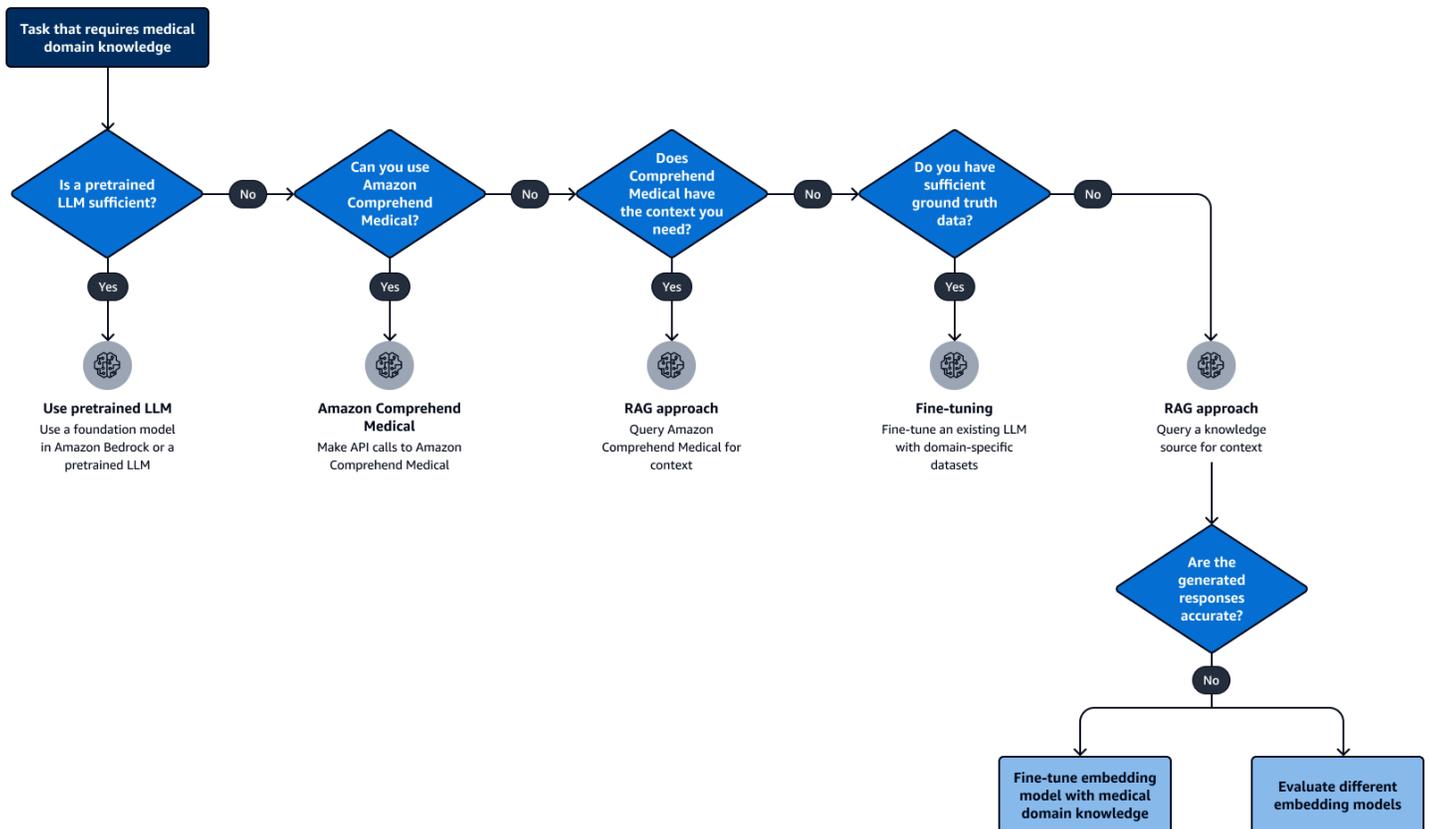
Recomendamos que você considere e avalie preconceitos, imparcialidade e alucinações em seus aplicativos generativos de IA. Se você estiver usando um LLM preexistente ou ajustando um, implemente grades de proteção para evitar respostas prejudiciais. As grades de proteção são proteções que você personaliza de acordo com seus requisitos generativos de aplicativos de IA e políticas de IA responsáveis. Por exemplo, você pode usar o [Amazon Bedrock Guardrails](#).

Escolhendo uma abordagem de PNL para saúde e ciências biológicas

A [Abordagens generativas de IA e PNL para saúde e ciências biológicas](#) seção descreve as seguintes abordagens para lidar com tarefas de processamento de linguagem natural (PNL) para aplicações de saúde e ciências biológicas:

- Usando o Amazon Comprehend Medical
- Combinando o Amazon Comprehend Medical com um LLM em um fluxo de trabalho de Retrieval Augment Generation (RAG)
- Usando um LLM aperfeiçoado
- Usando um fluxo de trabalho RAG

Ao avaliar as limitações conhecidas das LLMs tarefas do domínio médico e seu caso de uso, você pode escolher qual abordagem funcionará melhor para sua tarefa. A árvore decisória a seguir pode ajudá-lo a escolher uma abordagem LLM para sua tarefa médica de PNL:



O diagrama mostra o seguinte fluxo de trabalho:

1. Para casos de uso de saúde e ciências biológicas, identifique se a tarefa de PNL requer conhecimento de domínio específico. Conforme necessário, coordene com especialistas no assunto (SMEs).
2. Se você puder usar um LLM geral ou um modelo treinado em conjuntos de dados médicos, use um modelo básico disponível no Amazon Bedrock ou no LLM pré-treinado. Para obter mais informações, consulte [Escolhendo um LLM](#) neste guia.
3. Se os recursos de detecção de entidades e de vinculação de ontologias do Amazon Comprehend Medical atenderem ao seu caso de uso, use o Amazon Comprehend Medical. APIs Para obter mais informações, consulte [Usando o Amazon Comprehend Medical](#) neste guia.
4. Às vezes, o Amazon Comprehend Medical tem o contexto necessário, mas não dá suporte ao seu caso de uso. Por exemplo, você pode precisar de definições de entidade diferentes, receber um grande número de resultados, precisar de entidades personalizadas ou precisar de uma tarefa de PNL personalizada. Se for esse o caso, use uma abordagem RAG para consultar o contexto do Amazon Comprehend Medical. Para obter mais informações, consulte [Combinando o Amazon Comprehend Medical com grandes modelos de linguagem](#) neste guia.
5. Se você tiver uma quantidade suficiente de dados reais básicos, ajuste um LLM existente. Para obter mais informações, consulte [Abordagens de personalização](#) neste guia.
6. Se as outras abordagens não satisfizerem os objetivos médicos de suas tarefas de PNL, implemente uma solução RAG. Para obter mais informações, consulte [Abordagens de personalização](#) neste guia.
7. Depois de implementar a solução RAG, avalie se as respostas geradas são precisas. Para obter mais informações, consulte [Avaliação LLMs para aplicações de saúde e ciências biológicas](#) neste guia. [É comum começar com um modelo Amazon Titan Text Embeddings ou um modelo geral de transformador de frases, como o All-MiniLM-L6-v2.](#) No entanto, devido à falta de contexto de domínio, esses modelos podem não capturar a terminologia médica do texto. Se necessário, considere os seguintes ajustes:
 - a. Avalie outros modelos de incorporação
 - b. Ajuste o modelo de incorporação com conjuntos de dados específicos do domínio

Considerações sobre a maturidade dos negócios

A maturidade dos negócios é fundamental ao adaptar as soluções de LLM para aplicações de saúde e ciências biológicas. Essas organizações enfrentam vários níveis de complexidade durante a implementação LLMs, dependendo de seus critérios de aceitação. Frequentemente, organizações que não têm recursos de IA/ML investem no suporte de prestadores de serviços para criar soluções de LLM. Nessas situações, é importante entender as seguintes vantagens e desvantagens:

- Alto desempenho para alto custo e manutenção — Você pode precisar de uma solução complexa que envolva ajustes finos ou personalizados LLMs para atender a padrões de desempenho rigorosos. No entanto, isso acarreta custos e requisitos de manutenção mais altos. Talvez seja necessário contratar recursos especializados ou fazer parcerias com prestadores de serviços para manter essas soluções sofisticadas. Isso pode potencialmente retardar o desenvolvimento.
- Bom desempenho com baixo custo e manutenção — Como alternativa, você pode descobrir que serviços como o Amazon Bedrock ou o Amazon Comprehend Medical oferecem um desempenho aceitável. Embora essas LLMs ou abordagens possam fornecer resultados perfeitos, essas soluções geralmente podem fornecer resultados consistentes e de alta qualidade. Essas soluções têm menor custo e reduzem a carga de manutenção. Isso pode acelerar o desenvolvimento.

Se uma abordagem mais simples e de baixo custo fornecer consistentemente resultados de alta qualidade que atendam aos seus critérios de aceitação, considere se o aumento do desempenho compensa as compensações de custo, manutenção e tempo. No entanto, se a solução mais simples ficar significativamente aquém do desempenho desejado e se sua organização não tiver a capacidade de investimento em soluções complexas e seus requisitos de manutenção, considere adiar o desenvolvimento de IA/ML até que mais recursos ou soluções alternativas estejam disponíveis.

Além disso, para qualquer solução médica de PNL que dependa de um LLM, recomendamos que você realize monitoramento e avaliação contínuos. Avalie o feedback dos usuários ao longo do tempo e implemente avaliações periódicas para garantir que a solução continue atendendo aos seus objetivos comerciais.

Avaliação LLMs para aplicações de saúde e ciências biológicas

Esta seção fornece uma visão geral abrangente dos requisitos e considerações para avaliar grandes modelos de linguagem (LLMs) em casos de uso de saúde e ciências biológicas.

É importante usar dados reais básicos e feedback do SME para mitigar o viés e validar a precisão da resposta gerada pelo LLM. Esta seção descreve as melhores práticas para coletar e organizar dados de treinamento e teste. Também ajuda você a implementar barreiras e medir o viés e a imparcialidade dos dados. Ele também discute as tarefas médicas comuns de processamento de linguagem natural (PNL), como classificação de texto, reconhecimento de entidades nomeadas e geração de texto, e suas métricas de avaliação associadas.

Ele também apresenta fluxos de trabalho para realizar a avaliação do LLM durante a fase de experimentação do treinamento e a fase de pós-produção. O monitoramento do modelo e as operações de LLM são elementos importantes desse processo de avaliação.

Dados de treinamento e teste para tarefas médicas de PNL

As tarefas de PNL médica geralmente usam corporações médicas (como PubMed) ou informações do paciente (como notas de visitas de pacientes à clínica) para classificar, resumir e gerar insights. O pessoal médico, como médicos, administradores de serviços de saúde ou técnicos, varia em experiência e pontos de vista. Devido à subjetividade entre esses profissionais médicos, conjuntos menores de dados de treinamento e testes representam um risco de viés. Para mitigar esse risco, recomendamos as seguintes práticas recomendadas:

- Ao usar uma solução LLM pré-treinada, verifique se você tem uma quantidade adequada de dados de teste. Os dados do teste devem ser exatamente iguais ou muito parecidos com os dados médicos reais. Dependendo da tarefa, isso pode variar de 20 a mais de 100 registros.
- Ao ajustar um LLM, colete um número suficiente de registros rotulados (verdades fundamentais) de uma variedade SMEs do domínio médico alvo. Um ponto de partida geral são pelo menos 100 registros de alta qualidade, e recomendamos não mais do que 20 registros de cada PME. No entanto, dada a complexidade da tarefa e seus critérios de aceitação de precisão, mais registros podem ser necessários.
- Se necessário para seu caso de uso médico, implemente grades de proteção e meça o viés e a imparcialidade dos dados. Por exemplo, certifique-se de que o LLM evite diagnósticos errados

devido aos perfis raciais dos pacientes. Para obter mais informações, consulte a [Segurança e grades de proteção](#) seção deste guia.

Muitas empresas de pesquisa e desenvolvimento de IA, como a Anthropic, já implementaram grades de proteção em seus modelos básicos para evitar toxicidade. Você pode usar a detecção de toxicidade para verificar as solicitações de entrada e as respostas de saída de LLMs. Para obter mais informações, consulte [Detecção de toxicidade](#) na documentação do Amazon Comprehend.

Em qualquer tarefa generativa de IA, existe o risco de alucinação. Você pode mitigar esse risco executando tarefas de PNL, como classificação. Você também pode usar técnicas mais avançadas, como métricas de similaridade de texto. [BertScore](#) é uma métrica de similaridade de texto comumente adotada. Para obter mais informações sobre técnicas que você pode usar para mitigar a alucinação, consulte [Uma pesquisa abrangente sobre técnicas de mitigação de alucinações em modelos de linguagem ampla](#).

Métricas para tarefas médicas de PNL

Você pode criar métricas quantificáveis depois de estabelecer dados reais básicos e rótulos fornecidos pelas PME para treinamento e testes. Verificar a qualidade por meio de processos qualitativos, como testes de estresse e revisão dos resultados do LLM, é útil para um desenvolvimento rápido. No entanto, as métricas atuam como referências quantitativas que apoiam futuras operações de LLM e atuam como referências de desempenho para cada versão de produção.

Compreender a tarefa médica é fundamental. Normalmente, as métricas são mapeadas para uma das seguintes tarefas gerais de PNL:

- **Classificação de texto** — O LLM categoriza o texto em uma ou mais categorias predefinidas, com base na solicitação de entrada e no contexto fornecido. Um exemplo é classificar uma categoria de dor usando uma escala de dor. Exemplos de métricas de classificação de texto incluem:
 - [Precisão](#)
 - [Precisão](#), também conhecida como precisão macro
 - [Recall](#), também conhecido como recall de macro
 - [Pontuação F1](#), também conhecida como pontuação macro F1
 - [Perda de Hamming](#)

- Reconhecimento de entidade nomeada (NER) — Também conhecido como extração de texto, o reconhecimento de entidade nomeada é o processo de localizar e classificar entidades nomeadas que são mencionadas em texto não estruturado em categorias predefinidas. Um exemplo é extrair os nomes dos medicamentos dos prontuários dos pacientes. Exemplos de métricas do NER incluem:
 - [Precisão](#)
 - [Precisão](#)
 - [Recall](#)
 - [F1 score](#)
 - [Perda de Hamming](#)
- Geração — O LLM gera um novo texto processando a solicitação e o contexto fornecido. A geração inclui tarefas de resumo ou tarefas de resposta a perguntas. Exemplos de métricas de geração incluem:
 - [Substituta de Avaliação de Gisting Orientada a Recalls \(ROUGE\)](#)
 - [Métrica para avaliação de tradução com explícito ORdering \(METEOR\)](#)
 - [Subestudo de avaliação bilíngue \(BLEU\) \(para traduções\)](#)
 - [Distância da corda](#), também conhecida como similaridade de cosseno

Perguntas frequentes sobre casos de uso de saúde e ciências biológicas

A seguir estão as perguntas mais frequentes relacionadas ao uso do Amazon Comprehend Medical ou LLMs para tarefas médicas de PNL.

Como faço para escolher entre o Amazon Comprehend Medical e um LLM?

Se sua tarefa for detectar entidades médicas em seu texto médico, revise a [documentação do Amazon Comprehend Medical para entender quais entidades médicas](#) podem ser extraídas e se alguma das [ontologias](#) aborda seu caso de uso. Caso contrário, considere usar um LLM. Para obter mais informações, consulte [Casos de uso do Amazon Comprehend Medical](#) e [Casos de uso para um LLM](#) neste guia.

Como posso fornecer os resultados do Amazon Comprehend Medical a um LLM?

Você pode incorporar os resultados do Amazon Comprehend Medical como contexto em suas solicitações de LLM. Isso fornece conhecimento médico adicional e terminologia para o LLM. O contexto fornecido pode melhorar o desempenho do LLM em tarefas como reconhecimento de entidades, resumo ou resposta a perguntas. O guia fornece vários exemplos de como estruturar solicitações com os resultados do Amazon Comprehend Medical. Para obter mais informações, consulte [Combinando o Amazon Comprehend Medical com grandes modelos de linguagem](#) neste guia.

Quais são algumas das melhores práticas ao usar o Amazon Comprehend Medical com LLMs

Recomendamos usar as pontuações de confiança do Amazon Comprehend Medical para filtrar ou priorizar entidades dentro de suas solicitações. Também é importante avaliar seu desempenho em seus dados específicos e validar se as definições da entidade estão alinhadas com seus requisitos. A combinação do Amazon Comprehend Medical com fontes de conhecimento específicas do domínio

pode melhorar ainda mais o desempenho do LLM. Para obter mais informações, consulte [Melhores práticas para usar o Amazon Comprehend Medical em um fluxo de trabalho do RAG](#) neste guia.

Devo usar um LLM médico pré-treinado ou ajustar um LLM geral para meu caso de uso na área de saúde?

A decisão depende de seus requisitos específicos e da disponibilidade de dados de treinamento de alta qualidade. Um médico pré-treinado LLMs pode fornecer um bom ponto de partida. No entanto, talvez você ainda precise ajustá-los com os dados específicos do seu domínio. Se você tiver dados rotulados suficientes, ajustar um LLM geral pode ser uma opção viável. Para obter mais informações, consulte [Escolhendo um LLM](#) e [Escolhendo uma abordagem de PNL para saúde e ciências biológicas](#) neste guia.

Como faço para avaliar o desempenho de tarefas médicas LLMs de PNL?

Recomendamos o uso de métricas quantitativas, como exatidão, precisão, recordação e pontuação F1 para classificação de texto e tarefas de reconhecimento de entidades nomeadas. Você pode usar ROUGE e METEOR para tarefas de geração de texto. É importante ter dados confiáveis verdadeiros identificados por especialistas no assunto e implementar processos para monitorar o desempenho do modelo ao longo do tempo. Para obter mais informações, consulte [Avaliação LLMs para aplicações de saúde e ciências biológicas](#) neste guia.

Quais são as vantagens e desvantagens entre soluções LLM de alta complexidade e baixa complexidade?

Ajustar um LLM ou criar um LLM personalizado são soluções altamente complexas. Essas abordagens podem melhorar o desempenho, mas implicam custos e requisitos de manutenção mais altos. Soluções mais simples, como usar o Amazon Comprehend Medical pré-treinado LLMs ou o Amazon Comprehend Medical, podem oferecer desempenho aceitável com custos mais baixos e ciclos de desenvolvimento mais rápidos. No entanto, essas abordagens podem não atender aos rigorosos requisitos de precisão para alguns casos de uso. Para obter mais informações, consulte [Considerações sobre a maturidade dos negócios](#) neste guia.

Próximas etapas e recursos

Este guia ajuda você a Serviços da AWS automatizar tarefas médicas de PNL e IA generativa para aplicações do mundo real em ambientes de produção. Ele descreve como você pode usar o Amazon Comprehend Medical, apoiado no LLMs Amazon Bedrock, médico pré-treinado ou aperfeiçoado LLMs para atingir seus LLMs objetivos comerciais de saúde e ciências biológicas. Este guia descreve as vantagens e limitações das seguintes abordagens:

- Usando o Amazon Comprehend Medical de forma independente
- Fornecendo resultados do Amazon Comprehend Medical a um LLM
- Usando um LLM geral pré-treinado ou um LLM médico em uma abordagem de geração aumentada de recuperação (RAG)
- Ajustando um LLM geral ou LLM médico

Use a [árvore decisória](#) e as [considerações sobre maturidade dos negócios](#) neste guia para escolher entre essas abordagens com base no nível de maturidade de IA/ML da sua organização. Embora o Amazon Comprehend Medical e o Amazon Bedrock LLMs forneçam recursos poderosos, eles só são bem-sucedidos se você os implementar e avaliar adequadamente. Use as [informações e métricas de avaliação](#) descritas neste guia para validar o desempenho da sua solução.

Para as próximas etapas, recomendamos que os gerentes de TI, arquitetos e líderes técnicos da área de saúde trabalhem com profissionais de IA/ML para identificar suas tarefas médicas de PNL. Use este guia para escolher um caminho de desenvolvimento e, em seguida, use os recursos adequados Serviços da AWS para implementar com êxito uma solução automatizada em AWS.

AWS recursos

- Documentação do Amazon Comprehend Medical:
 - [Guia do desenvolvedor](#)
 - [API Reference](#)
- [Avaliação do modelo Amazon Bedrock](#)
- [Amazon SageMaker Ground Truth](#)
- [Detecção de toxicidade do Amazon Comprehend](#)
- [AWS Parceiros de competência em saúde](#)

Outros recursos

- [Tabela de classificação do Open Medical-LLM](#)
- [Uma pesquisa sobre grandes modelos de linguagem para assistência médica: de dados, tecnologia e aplicativos à responsabilidade e ética](#)
- [Modelos de linguagem grandes são codificadores médicos ruins — avaliação comparativa da consulta de códigos médicos](#)
- [Do iniciante ao especialista: modelando o conhecimento médico em geral LLMs](#)

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	—	16 de dezembro de 2024

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/rearquitar:** 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
- **Rever (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 descryptografia. É possível compartilhar a chave pública porque ela não é usada na descryptografia, 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 containerizados 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 recomendadas 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 CCoE](#) 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 ou Bitbucket 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/CD pode 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 (OU)s. 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

laC

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.

IloT

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 microserviço 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 microserviços 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: rehoste 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 todas as 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 microserviço com base em padrões de acesso a dados e outros requisitos. Se seus microserviços tiverem a mesma tecnologia de armazenamento de dados, eles poderão enfrentar desafios de implementação ou apresentar baixa performance. Os microserviços 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 microserviços](#).

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.