Fundamentos de IA e ML

Inteligência Artificial O que é Inteligência Artificial?

- A Inteligência Artificial (IA) é um campo da ciência da computação focado no desenvolvimento de sistemas que podem exibir comportamento inteligente, como raciocínio, aprendizagem e autonomia.
- A IA funciona combinando grandes quantidades de dados com algoritmos inteligentes. Estes algoritmos são treinados nos dados para aprender padrões e tomar decisões informadas. Esse processo permite que a IA execute tarefas que exigem inteligência semelhante à humana, como entender a linguagem e dirigir veículos.
- A AWS oferece uma variedade de serviços de IA pré-criados, bem como infraestrutura personalizável opções, projetadas para otimizar o desenvolvimento de IA e reduzir custos.

Como a IA processa informações e toma decisões?

Coleta de dados: sistemas de IA exigem grandes quantidades de dados para aprender. Esses dados podem ser qualquer coisa, de imagens e texto a valores numéricos.
Seleção de algoritmo: o algoritmo apropriado é escolhido com base na tarefa específica que a IA foi projetada para executar.
Estes são os principais conceitos e etapas no campo da Inteligência Artificial (IA) e Aprendizado de Máquina (AM): •

Aprendizado de Máquina: Envolve algoritmos de treinamento em grandes conjuntos de dados para identificar padrões e fazer previsões.

- **Aprendizado profundo:** um subconjunto do aprendizado de máquina que usa redes neurais artificiais para aprender de padrões complexos.
- **Processamento de Linguagem Natural (PLN):** A IA pode entender e responder à linguagem humana. •

Treinamento: O algoritmo é treinado nos dados coletados. Isso envolve alimentar os dados no algoritmo e permitindo que ele aprenda com ele.

- **Teste:** O algoritmo treinado é testado em novos dados para avaliar seu desempenho. • **Implantação:**

Depois que o algoritmo é testado e refinado, ele pode ser implantado em aplicações do mundo real.

Componentes-chave da arquitetura de aplicação de IA: A

arquitetura de inteligência artificial consiste em três camadas principais. Todas as camadas são

executadas em infraestrutura de TI que fornece os recursos de computação e memória necessários para a IA ser executada.

Camada 1: **Camada**

de dados A camada de dados é a pedra angular da IA. Ela envolve preparar e organizar dados para uso em aprendizado de máquina, processamento de linguagem natural e tecnologias de reconhecimento de imagem.

Camada 2: Camada de

modelo A camada de modelo foca nas capacidades de tomada de decisão do sistema de IA. As organizações frequentemente selecionam modelos de base pré-existentes ou modelos de linguagem grandes e os personalizam usando técnicas que incorporam dados relevantes.

Camada 3: Camada de

aplicação A camada de aplicação é a interface do sistema de IA voltada para o usuário. Ela permite que os usuários interajam com a IA, solicitem tarefas específicas, gerem informações ou tomem decisões baseadas em dados.

Aplicações da Inteligência Artificial: Chatbots e assistentes inteligentes-

- Os chatbots e assistentes inteligentes alimentados por IA são cada vez mais capazes de conduzir conversas que lembram interações humanas.

Processamento inteligente de documentos-

- O processamento inteligente de documentos (IDP) transforma documentos comerciais, como e-mails, imagens e PDFs em informações organizadas e estruturadas. O IDP alavanca tecnologias de IA, incluindo processamento de linguagem natural (NLP), aprendizado profundo e visão computacional, para extrair, categorizar e validar dados.

Monitoramento de desempenho do aplicativo-

- As ferramentas de APM orientadas por IA aproveitam dados históricos para antecipar problemas potenciais antes que eles ocorram. surgem. Além disso, essas ferramentas podem resolver problemas em tempo real, oferecendo aos desenvolvedores soluções práticas, garantindo operações suaves de aplicativos e resolvendo gargalos de desempenho.

Manutenção preditiva-

- A manutenção preditiva orientada por IA aproveita grandes quantidades de dados para detectar problemas potenciais que podem causar interrupções operacionais, de sistema ou de serviço. Ao antecipar problemas antes que eles surjam, a manutenção preditiva ajuda as empresas a minimizar o tempo de inatividade e evitar interrupções.

Pesquisa médica-

- A tecnologia de IA em pesquisa médica simplifica, automatiza tarefas repetitivas e lida com grandes conjuntos de dados. Ela pode ser empregada para aprimorar todo o processo de descoberta e desenvolvimento farmacêutico, transcrever registros médicos e acelerar o tempo de lançamento de novos produtos no mercado.

Business Analytics- •

O business analytics alavanca a IA para reunir, processar e examinar conjuntos de dados complexos. Com o analytics orientado por IA, você pode prever tendências futuras, identificar causas subjacentes dentro dos dados e agilizar tarefas demoradas.

Limitações da Inteligência Artificial na AWS:

Embora a AWS ofereça um conjunto robusto de serviços de IA e aprendizado de máquina, há limitações inerentes das quais os usuários devem estar cientes:

1. Qualidade e quantidade de dados:

- **Viés:** se os modelos de IA forem treinados com dados tendenciosos, isso pode levar a resultados injustos ou imprecisos.
- **Ruído:** dados incompletos ou ruidosos podem prejudicar a precisão e o desempenho do modelo.

2. Recursos Computacionais:

- **Custo:** Treinar e executar modelos de IA pode se tornar caro, especialmente para projetos em larga escala.
- **Infraestrutura:** Pode ser necessário acesso à infraestrutura de computação de alto desempenho para modelos complexos.

3. Considerações éticas:

- **Privacidade:** O tratamento de dados confidenciais levanta preocupações de privacidade, especialmente quando se usa IA para tarefas como reconhecimento facial ou processamento de linguagem natural.
- **Justiça:** os modelos de IA podem perpetuar preconceitos ou discriminações existentes se não forem concebidos e treinado cuidadosamente.

4. Supervisão humana:

- **Dependência:** as ferramentas relacionadas aos sistemas de IA ainda exigem supervisão humana para garantir que sejam usado de forma ética e eficaz.
- **Correção de erros:** os modelos de IA podem cometer erros, e a intervenção humana pode ser necessária para corrigir erros ou vieses.

Terminologias básicas de IA

O que é aprendizado de máquina?

Machine Learning (ML) é um subconjunto da inteligência artificial (IA) focado no desenvolvimento de algoritmos que melhoram automaticamente por meio de experiência e dados. Simplificando, o machine learning permite que os computadores aprendam com dados e tomem decisões ou previsões sem programação explícita.

O que é Deep Learning?

O aprendizado profundo, um subconjunto da inteligência artificial, permite que os computadores aprendam com dados como a cognição humana. Ao analisar padrões complexos em imagens, texto, áudio e outras formas de dados, os modelos de aprendizado profundo podem fornecer insights e previsões precisas. Esses modelos podem automatizar tarefas que tradicionalmente necessitam de inteligência humana, como descrição de imagem ou transcrição de áudio arquivos em texto.

O que é o Large Language Model [LLM]?

Large language models (LLMs) são modelos sofisticados de deep learning treinados em conjuntos de dados de texto extensos. Esses modelos utilizam uma arquitetura de transformador, uma estrutura de rede neural composta de um codificador e decodificador. O codificador e o decodificador alavancam mecanismos de autoatenção para extrair significado contextual de sequências de texto, compreendendo as relações entre palavras e frases.

O que é IA responsável?

Implicações mais amplas: os sistemas de IA têm efeitos substanciais sobre os indivíduos, comunidades, e o meio ambiente.

ÿ **Prioridades éticas:** as organizações devem enfatizar a justiça, a transparência e a ética considerações nas práticas de IA.

ÿ **Ato de equilíbrio:** as empresas devem equilibrar as práticas éticas de IA com a busca de

vantagem competitiva em um campo em rápida evolução.

O que são redes neurais?

Redes neurais artificiais servem como base para muitos sistemas de inteligência artificial. Inspiradas pela estrutura e função do cérebro humano, essas redes empregam unidades computacionais interconectadas, frequentemente chamadas de neurônios ou nós artificiais. Semelhantes aos neurônios biológicos, esses nós processam informações por meio de cálculos matemáticos. Ao trabalharem juntos em uma rede, esses nós podem resolver coletivamente problemas complexos e aprender com os dados.

O que é Processamento de Linguagem Natural (PLN)?

O processamento de linguagem natural (NLP) emprega redes neurais para extrair significado de dados textuais. Ele alavanca métodos computacionais projetados para compreender a linguagem humana, permitindo que máquinas processem palavras, gramática e estrutura de frases. Essa tecnologia facilita tarefas como resumo de documentos, interações de chatbot e análise de sentimentos.

O que é visão computacional?

A visão computacional, alimentada por aprendizado profundo, permite que os computadores interpretem e entendam informações visuais de imagens e vídeos. Essa tecnologia pode ser aplicada em vários domínios, como moderação de conteúdo para identificar imagens inapropriadas, reconhecimento facial e classificação de imagens. Também é crucial em veículos autônomos, onde ajuda a perceber o ambiente ao redor e a tomar decisões oportunas.

O que é reconhecimento de fala?

A tecnologia de reconhecimento de fala alavanca algoritmos de aprendizado profundo para decifrar a fala humana, discernir palavras individuais e compreender significado. Redes neurais são empregadas para transcrever a linguagem falada em texto escrito e para avaliar o tom emocional do falante. O reconhecimento de fala encontra
aplicações em vários domínios, como assistentes virtuais e sistemas de call center, onde é usado para interpretar comandos falados e executar ações correspondentes.

O que é IA generativa?

IA generativa é um tipo de inteligência artificial que pode produzir conteúdo original, incluindo imagens, vídeos, texto e áudio, com base em instruções textuais. Diferentemente da IA tradicional, que analisa principalmente dados, a IA generativa utiliza técnicas de aprendizado profundo e vastos conjuntos de dados para criar saídas criativas de
alta qualidade e semelhantes às humanas.

Embora essa tecnologia abra novas possibilidades para a expressão criativa, também há preocupações quanto a preconceito, conteúdo prejudicial e direitos de propriedade intelectual. Concluindo, a IA generativa marca um avanço significativo na capacidade da IA de gerar linguagem e conteúdo de qualidade humana.

Diferenças entre IA, ML, Deep Learning e Gen AI

Aspecto	Inteligência Artificial (IA)	Aprendizado de Máquina (ML)	Aprendizagem profunda	IA Generativa
Definição	A IA é um campo amplo de criando máquinas capazes de executar tarefas que normalmente exigem inteligência humana.	ML é um subconjunto da IA focado em sistemas que aprender com os dados para fazer decisões.	DL é um subconjunto de ML que usa redes neurais com múltiplas camadas (redes neurais profundas) para aprender com grandes quantidades de dados.	Gen AI é um tipo de IA focado em gerar novos conteúdos, como imagens, texto ou música, com base em padrões aprendidos.
Tipos	Sistemas baseados em regras, sistemas especialistas, árvores de decisão, ML e DL.	Supervisionado, não supervisionado, e reforço aprendizado.	Convolucional redes neurais (CNNs), redes neurais recorrentes (RNNs), transformadores.	Generativo Redes Adversárias (GANs), Autocodificadores Variacionais (VAEs), Grandes Modelos de Linguagem (LLMs).
Dados Dependência	Pode incluir regras baseadas sistemas sem aprendizagem componente.	Ele depende de rotulados ou dados não rotulados para aprendizagem padrões.	Requer grande quantidades de dados para treinar redes neurais profundas efetivamente.	São necessários grandes conjuntos de dados para treinamento para gerar resultados realistas.
Aplicações	Robótica, processamento de linguagem natural, sistemas especialistas, etc.	Análise preditiva, sistemas de recomendação, detecção de fraudes.	Reconhecimento de imagem, PNL, autônomo veículos.	Geração de imagens e vídeos, geração de textos (ex.: chatbots), composição musical.
Exemplos	Siri, robôs autônomos, computadores que jogam xadrez.	Filtros de spam, recomendações personalizadas, previsão do tempo.	Sistemas de reconhecimento facial, carros autônomos e linguagem tradução serviços.	DALL-E para geração de imagens, ChatGPT para geração de texto e DeepDream para criação de imagens.

Compreendendo o modelo de fundação

O que são modelos de fundação?

Os modelos de fundação (FMs) são redes neurais de aprendizado profundo em larga escala treinadas em conjuntos de dados extensos. Eles revolucionaram a abordagem que os cientistas de dados adotam para o aprendizado de máquina (ML) ao fornecer um ponto de partida que acelera o desenvolvimento de novos aplicativos de IA. Esses modelos são projetados para executar uma ampla gama de tarefas gerais, como compreensão de linguagem, geração de texto e imagem e processamento de linguagem natural (NLP).

Características dos modelos de fundação

Adaptabilidade:

Os modelos de fundação são excepcionalmente versáteis e capazes de executar uma variedade de tarefas com alta precisão com base em prompts de entrada. Isso os torna significativamente diferentes dos modelos de ML tradicionais, que normalmente são projetados para tarefas específicas, como análise de sentimentos, classificação de imagens ou previsão de tendências.

Natureza de uso geral:

Devido ao seu grande tamanho e treinamento amplo, os modelos de fundação podem servir como modelos base para aplicações mais especializadas. Ao longo dos anos, esses modelos cresceram em complexidade e tamanho, com modelos como BERT e GPT-4 mostrando essa evolução.

Aplicações dos modelos de base

Processamento de linguagem:

Os modelos de base se destacam em tarefas de linguagem natural, incluindo responder perguntas, escrever scripts e traduzir idiomas.

Compreensão visual:

Esses modelos são altamente eficazes em visão computacional, identificação de imagens, geração de imagens a partir de texto e edição de fotos e vídeos.

Geração de código:

Os modelos de fundação podem escrever e depurar código em várias linguagens de programação com base em instruções de linguagem natural.

Engajamento centrado no ser humano:

Eles dão suporte a processos de tomada de decisão, como diagnósticos clínicos e análises, aprendendo continuamente com informações humanas durante a inferência.

Exemplos de Modelos de Fundação

BERT (2018):

Um modelo bidirecional treinado em um vasto conjunto de dados, capaz de analisar texto e prever sentenças. Ele estabeleceu as bases para modelos futuros como o GPT.

GPT (Generative Pre-trained Transformer):

Lançado pela OpenAI, os modelos GPT evoluíram do GPT-1 com 117 milhões de parâmetros para GPT-4, que ostenta 170 trilhões de parâmetros. Esses modelos são capazes de tarefas que vão desde geração de texto até resposta a perguntas.

Amazon Titan:

Um modelo de base da Amazon oferece modelos generativos e de incorporação para tarefas como resumo de texto, extração de informações e personalização.

Desafios com modelos de fundação

Alta demanda de recursos: O desenvolvimento de modelos de fundação requer infraestrutura substancial, o que o torna caro e demorado.

Complexidade de integração:

Para uso prático, esses modelos devem ser integrados em sistemas de software, o que envolve desenvolvimento adicional para engenharia rápida e ajuste fino.

Problemas de compreensão e confiabilidade:

Embora os modelos de base possam gerar respostas coerentes, eles podem ter dificuldades para entender o contexto e podem produzir respostas não confiáveis ou tendenciosas.

Suporte da AWS para modelos de fundação

Amazon Bedrock:

Este serviço simplifica o desenvolvimento e o dimensionamento de aplicativos de IA generativa, oferecendo acesso a modelos básicos por meio de uma API, permitindo que os usuários escolham o modelo mais adequado às suas necessidades.

Amazon SageMaker JumpStart:

Um hub para modelos e soluções de ML, o SageMaker JumpStart fornece acesso a uma ampla variedade de modelos básicos, incluindo os mais populares, como Llama 2 e Falcon, dando suporte ao desenvolvimento de diversos aplicativos de IA.

Referência:

https://aws.amazon.com/what-is/foundation-models/

Modelos de IA: Tipos

1. Modelos de Visão Computacional

- **Amazon Rekognition:** Este serviço fornece modelos pré-treinados para análise de imagem e vídeo, incluindo recursos como detecção de objetos, reconhecimento facial e cena detecção.

2. Modelos de Processamento de Linguagem Natural (PLN)

- **Amazon Comprehend:** Usado para analisar texto, o Amazon Comprehend pode executar análise de sentimentos, reconhecimento de entidades e detecção de linguagem.
- **Amazon Translate:** fornece serviços de tradução em tempo real entre diferentes idiomas. • **Amazon Lex:** fornece interfaces de conversação, permitindo a criação de chatbots que podem interagir por voz e texto.
- **Amazon Polly:** converte texto escrito em fala realista em vários idiomas.

3. Modelos de reconhecimento de fala

- **Amazon Transcribe:** este serviço converte fala em texto, o que o torna útil para transcrições, legendas e muito mais.

4. Modelos de processamento de documentos

• **Amazon Textract:** extrai texto, tabelas e outros dados de documentos digitalizados, facilitando o processamento e a análise de informações em papel.

5. Modelos de Recomendação e Previsão

- **Amazon Personalize:** Oferece recomendações personalizadas analisando o comportamento do usuário e preferências. •

Amazon Forecast: utiliza dados de séries temporais para prever tendências futuras, como vendas previsões ou necessidades de inventário.

6. Modelos de busca e recuperação de informação

- **Amazon Kendra:** Um serviço de pesquisa empresarial que usa aprendizado de máquina para fornecer resultados de pesquisa relevantes em documentos e fontes de dados.

7. Modelos de aprendizado de máquina personalizados

- **Amazon SageMaker:** Uma plataforma abrangente para construir, treinar e implementar modelos de machine learning personalizados. Ela suporta uma ampla gama de algoritmos e frameworks.

8. Modelos de IA generativos

- **Amazon Bedrock:** Um serviço que fornece acesso a modelos fundamentais para IA generativa, permitindo que os usuários criem aplicativos personalizados, como geração de texto ou criação de imagens. • **SageMaker**

JumpStart: Oferece modelos e soluções pré-treinados para tarefas de IA generativa, que pode ser ajustado para necessidades específicas.

9. Modelos de IA de ponta

• **Inferência de ML do AWS IoT Greengrass:** permite a inferência de aprendizado de máquina em dispositivos de ponta, permitindo que modelos sejam implantados em ambientes onde o processamento em tempo real é crítico.

10. Modelos de IA híbrida

- **Amazon Neptune ML:** Integra aprendizado de máquina com bancos de dados gráficos, permitindo aplicações avançadas de análise de dados e gráficos de conhecimento.

Aprendizado de máquina

O que é aprendizado de máquina?

- **Conceito central:** o aprendizado de máquina gira em torno da criação de algoritmos que facilitam tomada de decisão e previsões. Esses algoritmos melhoram seu desempenho ao longo do tempo processando mais dados.
- **Programação tradicional vs. ML:** Diferentemente da programação tradicional, onde um computador segue instruções predefinidas, o machine learning envolve fornecer um conjunto de exemplos (dados) e uma tarefa. O computador então descobre como realizar a tarefa com base nesses exemplos.

- **Exemplo:** Para ensinar um computador a reconhecer imagens de gatos, não lhe damos instruções específicas instruções. Em vez disso, fornecemos milhares de imagens de gatos e deixamos o algoritmo de aprendizado de máquina identificar padrões e características comuns. Com o tempo, o algoritmo melhora e pode reconhecer gatos em novas imagens que ele não viu antes.

Tipos de aprendizado de máquina

O aprendizado de máquina pode ser amplamente classificado em três tipos:

1. **Aprendizagem supervisionada:** o algoritmo é treinado em dados rotulados, permitindo que ele faça previsões baseadas em pares de entrada-saída.
2. **Aprendizagem não supervisionada:** o algoritmo descobre padrões e relacionamentos dentro dados não rotulados.
3. **Aprendizagem por reforço:** o algoritmo aprende por tentativa e erro, recebendo feedback com base em suas ações.

Aplicações da Aprendizagem de Máquina

O aprendizado de máquina impulsiona muitos dos avanços tecnológicos atuais:

- **Assistentes de voz:** assistentes pessoais como Siri e Alexa contam com ML para entender e responder às dúvidas dos usuários.
- **Sistemas de recomendação:** plataformas como Netflix e Amazon usam ML para sugerir conteúdo e produtos com base no comportamento do usuário.
- **Carros autônomos:** veículos autônomos usam ML para navegar e fazer decisões.
- **Análise preditiva:** as empresas usam ML para prever tendências e fazer análises baseadas em dados decisões.

ML Pipeline: Componentes com serviços da AWS

Um pipeline de aprendizado de máquina (ML) na AWS se refere a um fluxo de trabalho estruturado que automatiza os vários estágios envolvidos no desenvolvimento, treinamento e implantação de modelos de aprendizado de máquina.

1. **Coleta de dados**

   - **Amazon S3 (Simple Storage Service):** Usado para armazenar grandes conjuntos de dados. A AWS fornece e armazenamento escalável para dados estruturados e não estruturados.
   - **AWS Glue:** Um serviço de integração de dados que ajuda a descobrir, preparar e combinar dados em várias fontes para análise.
   - **Amazon RDS (Relational Database Service):** para armazenar e gerenciar dados relacionais que podem ser usados para treinar modelos de ML.

2. **Análise Exploratória de Dados (EDA)**

   - **Amazon SageMaker Studio:** fornece um ambiente integrado onde cientistas de dados

     pode executar EDA usando notebooks Jupyter. Ele suporta bibliotecas de visualização como Matplotlib,

     Seaborn e Pandas para análise estatística e exploração de dados. • **Amazon Athena:** Um serviço de consulta interativo que permite analisar dados no Amazon
     S3 usando SQL. Útil para análise rápida sem a necessidade de mover dados.

3. **Pré-processamento de**

   dados • **AWS Glue e AWS Data Wrangler:** Essas ferramentas ajudam na limpeza, normalização e transformar dados brutos em um formato adequado para modelagem. Isso pode envolver manipulação de valores ausentes, normalização e dimensionamento de dados.
   - **Processamento do Amazon SageMaker:** permite a execução de trabalhos de pré-processamento que podem ser dimensionados para

     lidar com grandes conjuntos de dados.

4. **Engenharia de recursos** •

   Amazon SageMaker Feature Store: Um repositório totalmente gerenciado para armazenar, recuperar e compartilhar recursos entre diferentes modelos e equipes. Ele ajuda a automatizar o processo de extração e gerenciamento de recursos.
   - **Amazon SageMaker Data Wrangler:** simplifica o processo de transformação de recursos, permitindo que os usuários criem novos recursos combinando os existentes.

5. **Treinamento de modelo**

   - **Amazon SageMaker:** Suporta treinamento de modelos personalizados usando uma ampla variedade de algoritmos integrados ou seu próprio código. Ele também oferece treinamento distribuído, permitindo que você dimensione trabalhos de treinamento em várias instâncias.
   - **AWS Deep Learning AMIs:** fornece ambientes pré-configurados com deep learning populares estruturas de aprendizagem como TensorFlow, PyTorch e Apache MXNet.

6. **Ajuste de hiperparâmetros**

- **Amazon SageMaker Automatic Model Tuning:** Também conhecido como hiperparâmetro otimização (HPO), este serviço ajusta automaticamente os hiperparâmetros do modelo para melhorar o desempenho, usando técnicas como otimização bayesiana.

7. **Avaliação do modelo**

- **Amazon SageMaker Debugger:** Oferece insights sobre o processo de treinamento monitorando e criando perfis de jobs de treinamento. Ajuda a identificar problemas como overfitting e underfitting analisando métricas de treinamento.
- **Amazon SageMaker Model Monitor:** usado após a implantação para rastrear o modelo desempenho e detectar desvios de dados ao longo do tempo, garantindo que o modelo permaneça preciso.

8. **Implantação do modelo**

- **Amazon SageMaker Endpoint:** permite que você implante seus modelos treinados em tempo real, tornando-os acessíveis via API para inferência.
- **Amazon Elastic Kubernetes Service (EKS):** oferece suporte à implantação de modelos em um ambiente gerenciado pelo Kubernetes para aplicativos maiores e mais complexos.

9. **Monitoramento**

- **Amazon CloudWatch:** monitora modelos implantados em tempo real, coletando e rastreando métricas, registrando e disparando alertas para problemas de desempenho ou infraestrutura do modelo.
- **Amazon SageMaker Model Monitor:** monitora continuamente os modelos implantados para desvio de conceito, problemas de qualidade de dados e outras anomalias que podem afetar a precisão do modelo ao longo do tempo.

Fundamentos das Operações de ML (MLOps)

MLOps na AWS é um conjunto de práticas que combinam Machine Learning (ML) e DevOps para otimizar o desenvolvimento, a implantação e o gerenciamento de modelos de ML no ambiente de nuvem da Amazon Web Services (AWS).

1. **Experimentação**
   - **Prototipagem rápida:** serviços da AWS como o Amazon SageMaker permitem que cientistas de dados criem, testem e iterem rapidamente em modelos de aprendizado de máquina usando notebooks Jupyter e algoritmos pré-criados.
   - **Rastreamento de experimentos:** o SageMaker Experiments ajuda a rastrear e comparar diferentes execuções de modelos, capturando parâmetros, configurações e resultados para melhor reprodutibilidade e colaboração.
2. **Processos repetíveis**
   - **Automação de pipeline:** o SageMaker Pipelines automatiza todo o fluxo de trabalho de aprendizado de máquina, desde a preparação de dados até a implantação do modelo, garantindo que cada etapa seja repetível e consistente.
   - **Infraestrutura como código (IaC):** usando AWS CloudFormation ou Terraform, você pode definir e implantar infraestrutura de maneira consistente e repetível, garantindo que os ambientes sejam idênticos em diferentes estágios.
3. **Sistemas escaláveis**
   - **Recursos elásticos:** a AWS fornece recursos de computação escaláveis, como instâncias EC2 e Instâncias gerenciadas pelo SageMaker que podem ser dimensionadas automaticamente para cima ou para baixo com base na carga de trabalho, garantindo o uso eficiente dos recursos.
   - **Treinamento distribuído:** o SageMaker oferece suporte ao treinamento distribuído, permitindo treinamento em larga escala modelos sejam treinados mais rapidamente em várias GPUs ou instâncias.
4. **Gerenciando Dívida Técnica**
   - **Controle de versão: o controle** de versão de modelos, conjuntos de dados e código garante que você possa rastrear alterações, reproduzir resultados e evitar problemas causados por dados desatualizados ou inconsistentes. componentes.
   - **Registro de modelo:** o SageMaker Model Registry ajuda a gerenciar diferentes versões de modelos, armazenando metadados e promovendo modelos através de vários estágios de desenvolvimento e produção.
5. **Atingindo a prontidão para a produção**
   - **Integração contínua/implantação contínua (CI/CD):** implementar pipelines de CI/CD com o AWS CodePipeline ou Jenkins integra alterações de código, testes e implantações perfeitamente, garantindo que os modelos estejam sempre prontos para produção.

- **Segurança e conformidade:** a AWS fornece ferramentas como o AWS Identity and Access Management (IAM) e o AWS Key Management Service (KMS) para proteger dados, modelos e pipelines, garantindo a conformidade com os padrões do setor.

6. Monitoramento de modelo

- **Monitoramento de desempenho:** o SageMaker Model Monitor monitora automaticamente os modelos implantados modelos para precisão e desvio de desempenho, alertando as equipes sobre quaisquer problemas que possam exigir atenção. •

Registro e análise: o AWS CloudWatch e o AWS X-Ray podem ser usados para registrar previsões de modelos, rastrear métricas de desempenho e diagnosticar problemas em tempo real.

7. Retreinamento do modelo

- **Retreinamento automatizado:** SageMaker Pipelines e Step Functions podem automatizar o

processo de retreinamento quando o desempenho de um modelo cai ou novos dados ficam disponíveis. • **Detecção**

de desvio de dados: ferramentas de monitoramento como o SageMaker Model Monitor podem detectar quando a distribuição de dados de entrada muda, acionando um pipeline de retreinamento de modelo para garantir que o modelo

permanece preciso.

8. Escalabilidade e flexibilidade

- **Implantação escalável:** os endpoints do SageMaker podem ser dimensionados automaticamente para lidar com

aumentando o tráfego, garantindo que o modelo possa fornecer previsões de forma eficiente, independentemente da carga.

- **Endpoints multimodelo:** permite a implantação de vários modelos em um único endpoint,

otimizando a utilização de recursos e reduzindo custos.

9. Colaboração e Governança

- **Ferramentas de colaboração:** o SageMaker Studio fornece uma interface unificada onde cientistas e engenheiros de dados podem colaborar, compartilhar experimentos e trabalhar em modelos juntos.
- **Governança e Auditoria:** A AWS fornece ferramentas para manter a governança, como

SageMaker Clarify para detecção de viés e SageMaker Model Monitor para garantir a conformidade do modelo com as regras de negócios.

10. Gerenciamento de dívida técnica •

Gerenciamento de artefatos: usar serviços como o S3 para armazenar conjuntos de dados, modelos e logs ajuda a gerenciar e organizar artefatos de forma eficiente, reduzindo a dívida técnica associada a recursos desorganizados.

- **Reutilização de código:** Utilizando código modular e práticas padronizadas entre equipes minimiza o trabalho redundante e acelera projetos futuros.

Amazon SageMaker

O que é o Amazon SageMaker?

- O Amazon SageMaker é uma plataforma abrangente que capacita os usuários a desenvolver, treinar,

  e implementar modelos de aprendizado de máquina de forma eficiente. Este serviço totalmente gerenciado oferece uma ampla gama de ferramentas, incluindo notebooks, depuradores, profilers, pipelines e MLOps capacidades, para otimizar todo o ciclo de vida do ML.

- O Amazon SageMaker oferece uma variedade de ferramentas pré-criadas, incluindo algoritmos, pré-treinados modelos e modelos de soluções para agilizar o desenvolvimento e a implantação de modelos de aprendizado de máquina para ajudar cientistas de dados e profissionais.

Seleção de Algoritmo:

Tipo de problema	Algoritmo apropriado
Classificação Binária	Regressão Logística, XGBoost, etc.
Classificação multiclasse	XGBoost, Aprendiz Linear, etc.
Regressão	Aprendiz Linear, XGBoost, etc.
Detecção de objetos	R-CNN, SSD, etc. mais rápidos.
Detecção de anomalias	Floresta de corte aleatório, etc.
Agrupamento	K-Means, DBSCAN, etc.
Modelagem de Tópicos	Alocação latente de Dirichlet (LDA)
Sistemas de Recomendação	Máquinas de fatoração, etc.

Características:

Preparar dados -

- **SageMaker Feature Store:-** Amazon SageMaker Feature Store é uma plataforma centralizada

  projetado para armazenar, compartilhar e gerenciar recursos usados em modelos de aprendizado de máquina. Os recursos são as entradas de dados nas quais os modelos se baseiam durante o treinamento e a inferência.
- **SageMaker Data Wrangler:-** O Amazon SageMaker Data Wrangler seleciona, entende,

  e transforma dados para prepará-los para aprendizado de máquina (ML) em minutos. reduz dados

  tempo de preparação para dados tabulares, de imagem e de texto de semanas para minutos. Ele permite uma rápida avaliação da precisão do modelo de ML e ajuda a identificar problemas potenciais antes Implantação.

- ML geoespacial com Amazon SageMaker: - O Amazon SageMaker capacita cientistas de dados e engenheiros de ML a criar, treinar e implantar modelos de ML usando dados geoespaciais, como imagens de satélite, mapas
  e dados de localização.

Construir -

- **SageMaker Notebooks:-** Amazon SageMaker Notebooks oferece um Jupyter totalmente gerenciado

  ambiente, permitindo que cientistas de dados e engenheiros de ML explorem, analisem e desenvolvam modelos de aprendizado de máquina com eficiência.
- **SageMaker Jumpstart:-** O Amazon SageMaker JumpStart é um hub de aprendizado de máquina (ML) que pode ajudar você a avaliar, comparar e selecionar rapidamente modelos do Foundation com base em métricas de qualidade e responsabilidade predefinidas para executar tarefas como resumo de artigos e geração de imagens.

- **SageMaker Studio Lab:-** Amazon SageMaker Studio Lab é um serviço gratuito baseado em

  JupyterLab de código aberto que permite aos clientes usar recursos de computação da AWS para criar e executar seus notebooks Jupyter.

Train -

- **Treinamento do modelo SageMaker: -** O treinamento do modelo Amazon SageMaker simplifica o

  processo de treinamento e ajuste de modelos de aprendizado de máquina, reduzindo significativamente o tempo e os custos, ao mesmo tempo que elimina a necessidade de gerenciamento de
infraestrutura. • **Experimentos do SageMaker: -** O Amazon SageMaker oferece um recurso MLflow gerenciado que simplifica o aprendizado de máquina e a experimentação de IA generativa. Cientistas de dados podem usar facilmente o MLflow dentro do SageMaker para treinamento, registro e implantação de modelos.
  Os administradores podem estabelecer rapidamente ambientes MLflow seguros e escaláveis na AWS. •

SageMaker HyperPod:- O Amazon SageMaker HyperPod simplifica o processo de construção e otimização da infraestrutura de ML para treinamento de modelos de base, reduzindo significativamente o tempo de treinamento em 40%. Ao distribuir automaticamente as cargas de trabalho de treinamento em milhares de aceleradores, o HyperPod permite o processamento paralelo e acelera o desempenho do modelo.

Implantar

- Implantação do modelo SageMaker: - O Amazon SageMaker simplifica a implantação de modelos de aprendizado

de máquina, incluindo modelos básicos, oferecendo ótima relação custo-benefício para solicitações de inferência em vários aplicativos.

- **SageMaker Pipelines:-** Amazon SageMaker Pipelines é um fluxo de trabalho sem servidor

  serviço de orquestração que automatiza fluxos de trabalho de aprendizado de máquina (ML) e modelos de grandes linguagens (LLM).

ML de ponta a ponta -

- **SageMaker MLOps:-** O Amazon SageMaker oferece ferramentas especializadas para gerenciar operações de aprendizado de máquina (MLOps), simplificando e padronizando processos em todo o ciclo de vida do aprendizado de máquina. • **SageMaker**

Canvas:- O Amazon SageMaker Canvas oferece uma interface visual que simplifica

o processo de machine learning. Ele permite que você prepare dados, crie e implante modelos de ML de forma eficiente. •

SageMaker Studio:- O Amazon SageMaker Studio fornece um conjunto abrangente de ferramentas para todas as etapas de desenvolvimento de machine learning, incluindo preparação de dados, construção de modelos, treinamento, implantação e gerenciamento.

Recurso	Tela do SageMaker	Estúdio SageMaker
Alvo Público	Cientistas de dados e engenheiros de ML com experiência limitada em codificação, habilidades avançadas de codificação	Cientistas de dados e engenheiros de ML com
Interface	Interface visual sem código	Ambiente de desenvolvimento integrado (IDE)
Modelo Prédio	Seleção automatizada de modelos e treinamento	Seleção manual de modelos e treinamento usando vários algoritmos
MLOps	Recursos básicos do MLOps (monitoramento, controle de versão)	Recursos avançados de MLOps (criação de pipeline, rastreamento de experimentos)
Casos de uso	Prototipagem rápida, análise exploratória de dados, modelos simples de ML	Modelos complexos de ML, pipelines personalizados, projetos de pesquisa

SageMaker Ground Truth: - O Amazon SageMaker Ground Truth fornece uma plataforma robusta para incorporar conhecimento humano ao processo de aprendizado de máquina, aprimorando o desempenho do modelo por meio de feedback contínuo.

Governança de ML -

- **Governança de ML com SageMaker:-** O Amazon SageMaker oferece recursos de governança especializados para garantir práticas de aprendizado de máquina responsáveis. O Amazon SageMaker Role Manager permite que
os administradores estabeleçam rapidamente as permissões necessárias. • **SageMaker Clarify:-**

O SageMaker Clarify simplifica o processo de identificação de potenciais

vieses no seu conjunto de dados. Especifique os recursos de entrada com os quais você está preocupado, como gênero ou idade, e o SageMaker Clarify conduzirá uma análise completa para descobrir quaisquer vieses potenciais presentes nesses recursos.

Fundamentos da IA Generativa

IA Generativa

O que é IA generativa?

- A IA generativa é uma forma de inteligência artificial capaz de gerar conteúdo original, como texto, recursos visuais e áudio.
- Você pode treiná-lo para entender e gerar texto em linguagens humanas e de programação.

  Ele também pode aprender assuntos complexos como arte, química, biologia, etc. Ao aproveitar dados de treinamento anteriores, ele pode aplicar seu conhecimento a tarefas novas e desconhecidas.
- Usando a AWS, você pode desenvolver e expandir rapidamente aplicativos de IA generativa adaptados aos seus dados específicos, casos de uso e necessidades do cliente. Beneficie-se de segurança e privacidade de
  nível empresarial, acesso a modelos fundamentais de ponta e uma abordagem centrada em dados.

Como funciona a IA Generativa?

A IA generativa utiliza modelos de aprendizado de máquina, que são amplamente treinados em grandes conjuntos de dados, para produzir novos conteúdos.
Modelos de fundação:- Modelos de fundação são modelos de machine learning treinados em conjuntos de dados massivos e diversos sem rótulos de tarefas específicos. Eles são capazes de executar uma ampla variedade de tarefas gerais.
Grandes modelos de linguagem:- LLMs são uma classe de FMs. Esses modelos são capazes de executar uma variedade de funções linguísticas, incluindo sumarização, geração de texto, classificação, diálogo aberto e extração de
informações.

Benefícios da IA Generativa:

- Acelera a pesquisa

- Melhora a experiência do cliente • Otimiza os processos de negócios • Aumenta a produtividade dos funcionários

Modelos de IA generativos:

Modelos de difusão: - Os modelos de difusão criam novos dados introduzindo gradualmente ruído em amostras de dados existentes e, em seguida, removendo-o cuidadosamente. Esse processo envolve adicionar alterações aleatórias controladas aos dados originais em várias iterações. O modelo garante que os dados gerados permaneçam coerentes
e realistas controlando cuidadosamente o nível de ruído.

Uma vez que os dados tenham sido suficientemente ruidosos, o modelo de difusão inverte o processo. Ele remove gradualmente o ruído, passo a passo, até produzir uma nova amostra de dados que se assemelha muito ao original. Esse processo de redução de ruído reverso permite que o modelo gere novos dados que são consistentes com a distribuição de dados subjacente.

Redes Adversariais Generativas (GAN): - GANs são amplamente utilizadas na geração de imagens realistas, transformação de estilos visuais e tarefas de aumento de dados.
Redes Adversariais Generativas (GANs) empregam um processo de treinamento competitivo entre duas redes neurais. A rede geradora produz dados falsos introduzindo ruído aleatório, enquanto a rede discriminadora tenta diferenciar entre esses dados falsos e dados reais.
À medida que o treinamento progride, o gerador refina sua capacidade de criar dados cada vez mais realistas, enquanto o discriminador se torna mais hábil em identificar dados falsos. Esse processo iterativo continua até que o gerador gere dados que sejam indistinguíveis de dados autênticos, mesmo para o discriminador.

Autocodificadores Variacionais:- VAEs criam uma representação comprimida de dados, frequentemente chamada de espaço latente. Esse espaço latente é uma construção matemática que captura a essência dos dados. Imagine-o como um código único, resumindo as principais características dos dados.
Por exemplo, ao estudar rostos, o espaço latente contém números que representam o formato dos olhos, o formato do nariz, as maçãs do rosto e as orelhas.
VAEs usam duas redes neurais. A rede codificadora transforma dados de entrada em uma média e variância para cada dimensão do espaço latente. Uma amostra aleatória é extraída de uma distribuição gaussiana usando esses parâmetros, resultando em uma representação do espaço latente. Essa representação compactada serve como uma versão simplificada da entrada. A rede decodificadora pega essa amostra latente e tenta reconstruir os dados originais. A qualidade da reconstrução é avaliada usando métricas matemáticas que comparam a saída reconstruída com a entrada original.

Modelos baseados em transformadores: - O modelo de IA generativa baseado em transformadores se baseia nos conceitos de codificador e decodificador de VAEs.
Modelos baseados em transformadores usam um mecanismo de autoatenção. A autoatenção ajuda a habilitar modelos generativos para priorizar palavras significativas durante o processamento. Modelos baseados em transformadores alavancam múltiplas camadas de codificadores múltiplas chamadas cabeças de atenção para identificar diversas interconexões entre palavras. Cada cabeça foca em seções distintas da sequência de entrada, facilitando uma análise abrangente dos dados.

Ferramentas para construir aplicações de IA generativas:

- Amazon Bedrock
- Amazon SageMaker
- Treinamento AWS
- Inferência da AWS
- Instâncias P5 do Amazon EC2
- UltraClusters do Amazon EC2

Aplicações generativas alimentadas por IA:

- Amazon Q
- PartyRock - Parque infantil Amazon Bedrock • AWS HealthScribe

Casos de uso de IA generativa:

- Chatbots e Assistentes Virtuais
- Análise Conversacional • Assistente de Funcionário • Geração de Código
- Personalização
- Modelos de soluções de produtividade e criatividade

Limitações da IA generativa na AWS:

Segurança:- O uso de dados proprietários para treinar modelos de IA generativos levanta preocupações sobre privacidade e segurança de dados. É crucial implementar medidas que protejam informações sensíveis e impeçam acesso não autorizado.
Criatividade:- Embora a IA generativa possa produzir conteúdo criativo, suas saídas são frequentemente limitadas pelos dados em que foi treinada. A criatividade humana, que envolve compreensão mais profunda,

ressonância emocional e pensamento original, continua sendo um desafio para a IA replicar completamente.
Custo: - Treinar e executar modelos de IA generativa exigem recursos computacionais significativos.

Soluções baseadas em nuvem oferecem uma abordagem mais acessível e econômica em comparação à criação de modelos do zero.

Explicabilidade:- Modelos de IA generativos são frequentemente considerados "caixas pretas", dificultando a compreensão de seus processos de tomada de decisão. Melhorar sua interpretabilidade e transparência é essencial para construir confiança e promover uma adoção mais ampla.

Matriz de escopo de segurança GenAI

Uma Generative AI Security Scoping Matrix oferece uma abordagem estruturada para organizações avaliarem e implementarem medidas de segurança em todo o ciclo de vida de aplicativos de IA. Ao categorizar preocupações de segurança, ela fornece uma estrutura direcionada para proteger sistemas de

IA. • A AWS fornece vários serviços para proteger cargas de trabalho de IA generativa. Os serviços da AWS variam significativamente em sua infraestrutura subjacente, software, mecanismos de acesso e manipulação de dados. Para simplificar o gerenciamento de segurança, organizamos esses serviços em categorias lógicas chamadas 'escopos'.

Escopo (determine seu escopo):

• Para começar, você precisará determinar em qual escopo seu caso de uso se

encaixa. • Os escopos são numerados de 1 a 5, representando da menor à maior propriedade que sua organização tem sobre o modelo de IA e seus dados associados.

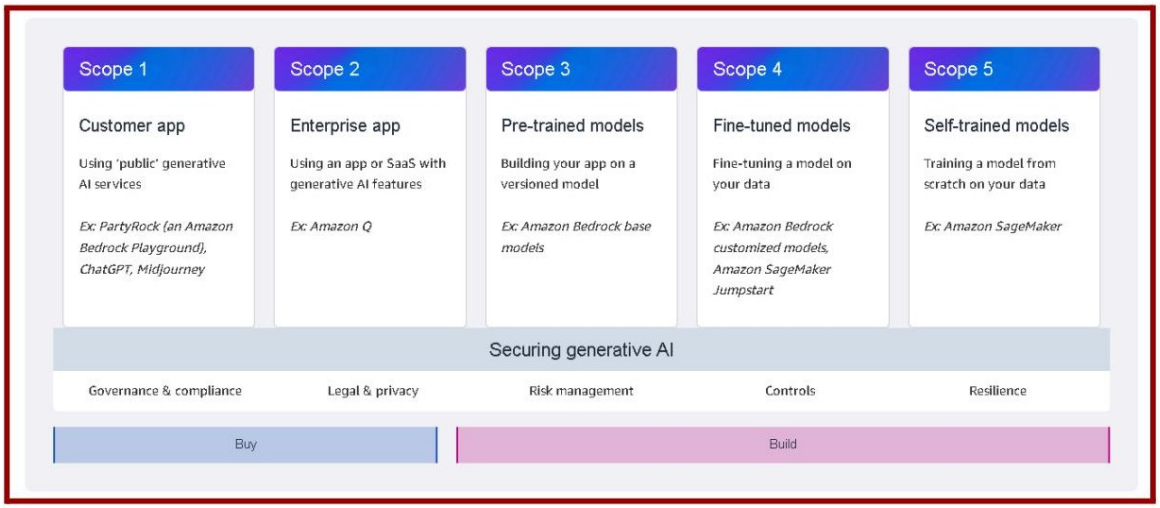


Figura 1: *Matriz de escopo de segurança de IA generativa*

Comprando IA generativa:

- **Escopo 1:** Aplicativo do consumidor – Sua empresa consome uma IA generativa pública de terceiros serviço, que é gratuito ou pago. Neste escopo, você não tem propriedade ou acesso aos dados ou modelo de treinamento subjacentes. Você só pode interagir com o serviço por meio de suas APIs ou aplicativos fornecidos, aderindo aos termos de uso do provedor.

Exemplo: Um trabalhador usa um chatbot de IA generativa para fazer um brainstorming de uma campanha de marketing

conceitos.

- **Escopo 2:** Aplicativo empresarial – Sua empresa usa um aplicativo empresarial de terceiros que possui recursos de IA generativa, e um relacionamento comercial é estabelecido entre sua organização e o fornecedor.

Exemplo: você usa um aplicativo de agendamento empresarial de terceiros que possui um recurso de IA generativa incorporado para ajudar a redigir pautas de reuniões.

- **Escopo 3:** Modelos pré-treinados – Sua empresa utiliza um modelo de fundação de IA generativa de terceiros existente para alimentar seu aplicativo. Este modelo é acessado e integrado às suas operações por meio de uma interface de programação de aplicativo.

  Exemplo: Um chatbot de suporte ao cliente foi desenvolvido utilizando o modelo de base Anthropic Claude, acessado por meio da API Amazon Bedrock. • **Escopo 4:** Modelos ajustados

– Sua empresa refina um modelo de base de IA generativa de terceiros existente, ajustando-o com dados específicos para sua empresa, gerando um modelo novo e aprimorado, especializado para sua carga de trabalho.

Exemplo: Ao aproveitar um modelo básico por meio de uma API, você pode criar um aplicativo de marketing que adapta materiais promocionais especificamente para seus produtos e serviços.

- **Escopo 5:** Modelos autotreinados – Sua empresa cria e treina um modelo de IA generativo do zero usando dados que você possui ou adquire. Você possui todos os aspectos do modelo.

  Exemplo: sua empresa deseja criar um modelo treinado exclusivamente em dados profundos e específicos do setor para licenciar empresas nesse setor, criando um modelo completamente novo

Mestrado em Direito.

Ao identificar as aplicações específicas da IA generativa, as equipes de segurança podem priorizar seus esforços e avaliar os riscos potenciais dentro de cada domínio de segurança.

Vamos examinar como o escopo influencia os requisitos de segurança dentro de cada disciplina de segurança. •

Governança

e conformidade – Implementação de políticas, procedimentos e

mecanismos de relatórios podem permitir que as empresas operem com eficiência e, ao mesmo tempo, mitiguem riscos.

- **Legal e privacidade –** Os requisitos legais, regulamentares e de privacidade específicos para uso ou criando soluções de IA generativas.
- **Gestão de riscos –** Avaliação de riscos associados à IA generativa e proposição de contramedidas.
- **Controles –** Implementação de medidas de segurança para reduzir riscos. •

Resiliência – Projetar sistemas de IA generativos confiáveis que atendam consistentemente às necessidades de negócios.

SLAs.

Amazon SageMaker JumpStart

O Amazon SageMaker JumpStart é um hub de machine learning que pode acelerar seu desenvolvimento de ML. Usando o SageMaker JumpStart, você pode selecionar, avaliar e comparar FMs rapidamente com base em métricas de qualidade e responsabilidade predefinidas para executar tarefas como geração de imagens e resumo de artigos.

Características:

Modelos de Fundação:- Descubra uma variedade de modelos de fundação de provedores líderes como AI21 Labs, Databricks, Hugging Face, Meta, Mistral AI, Stability AI e Alexa. Esses modelos podem ser usados para realizar uma ampla gama de tarefas, incluindo resumir artigos e gerar texto, imagens ou vídeos.

Algoritmos integrados:- Você pode utilizar modelos de solução integrados por meio do SageMaker Python SDK. Esses algoritmos abordam tarefas comuns de ML, incluindo classificação de dados de imagem, texto e tabular, bem como análise de sentimento.
Soluções pré-criadas: - O SageMaker JumpStart oferece soluções pré-criadas de ponta a ponta para aplicações comuns de aprendizado de máquina, como previsão de demanda, avaliação de risco de crédito, detecção de fraudes e visão computacional.

Benefícios do SageMaker JumpStart: Modelos
de base disponíveis publicamente Algoritmos de ML integrados Soluções

personalizáveis

Suporte à colaboração **Casos**

de uso e vantagens: **1. Integração do modelo de fundação**

- Implante modelos como LLaMA 2 e Stable Diffusion no modo VPC, mesmo sem internet. • Acesse modelos pré-treinados para fácil implantação e ajuste.

2. Grandes Modelos de Linguagem (LLMs)

- Simplifica a implantação e o ajuste de LLMs, incluindo modelos de parâmetros de 40 milhões para tarefas de PNL.

3. Classificação de texto

- Modelos pré-construídos para classificação de texto com opções de personalização.

4. Geração de imagem •

Implante o Stable Diffusion XL para geração de imagens de alta qualidade.

5. Soluções sem código

- Implantação rápida e sem código para soluções de IA rápidas, acessíveis a não especialistas.

6. Recursos de aprendizagem

- Tutoriais em vídeo e guias para fácil implantação e ajuste de modelos.

Amazon Bedrock

- O Amazon Bedrock é um serviço gerenciado sem servidor que fornece vários modelos de base (FMs) de alto desempenho das principais empresas de IA, como AI21 Labs, Anthropic, Cohere, Meta, Mistral AI, Stability AI e Amazon.
- Esses modelos são acessíveis por meio de uma API unificada para criar aplicativos de IA generativos com foco em segurança, privacidade e práticas responsáveis de IA.
- Com o Amazon Bedrock, você pode testar e comparar rapidamente diferentes modelos de fundação para encontre o melhor ajuste para seu caso de uso.
- Esses modelos podem ser adaptados aos seus dados exclusivos usando técnicas como ajuste fino e Recuperação de Geração Aumentada (RAG). • Além

disso, você pode criar agentes que podem executar tarefas usando os sistemas da sua empresa e informações.

Como o Amazon Bedrock ajuda a criar aplicativos de IA generativa?

- **Escolha do modelo -** Escolha entre uma variedade de FMs líderes: a API única da Amazon Bedrock permite que você alterne facilmente entre diferentes modelos de base e suas atualizações. • **Personalização -** Adapte

modelos de forma privada com seus dados: a personalização do modelo permite que você

entregue experiências de usuário diferenciadas e personalizadas. Ajuste modelos de base com seus dados para criar experiências únicas e personalizadas.

- **RAG -** Entregar respostas de FM mais relevantes: Para fornecer aos FMs dados relevantes da empresa, as organizações usam RAG. Essa técnica alimenta dados em prompts para melhorar as respostas.
- **Agentes -** Execute tarefas complexas em todos os sistemas da empresa: os agentes do Amazon Bedrock automatizam tarefas complexas usando os sistemas e dados da sua empresa. Os agentes analisam solicitações, executam APIs relevantes e fornecem respostas seguras e privadas.

A Amazon Bedrock oferece modelos em 3 estados:

- **Ativo:** O fornecedor do modelo está desenvolvendo ativamente esta versão e ela continuará a ser atualizado com correções de bugs e pequenas melhorias.
- **Legado:** Uma versão é marcada como legado quando uma versão mais avançada oferece desempenho superior resultados. A Amazon Bedrock determina uma data de EOL para versões desatualizadas.
- **EOL:** Esta versão está desatualizada e inoperante. Solicitações feitas a ela falharão.

Casos de uso:

- **Geração de texto -** Produza conteúdo exclusivo para seu blog, mídias sociais e páginas da web. • **Assistentes virtuais -**

Crie assistentes que entendam as dúvidas dos usuários, dividam tarefas automaticamente, interajam conversacionalmente para reunir os detalhes necessários e executem ações para concluir a tarefa solicitada.

- **Pesquisa de texto e imagem -** Identifique e compile informações relevantes para responder a perguntas e fornecer recomendações baseadas em um grande conjunto de dados textuais e visuais.

- **Resumo de texto -** Obtenha resumos concisos de documentos extensos, como artigos, relatórios, trabalhos de pesquisa, documentação técnica e até livros, para extrair informações essenciais de forma eficaz.

- **Geração de imagens -** Gere imagens realistas e visualmente envolventes para campanhas publicitárias, sites, apresentações e outros aplicativos.

Agentes Amazon Bedrock: os agentes Amazon Bedrock permitem que você desenvolva e configure agentes autônomos para seu aplicativo.

Características:

- Os agentes do Amazon Bedrock acessam com segurança os dados da sua empresa e aprimoram as solicitações dos usuários

  com informações relevantes e fornecer respostas precisas.

- Os agentes do Amazon Bedrock orquestram e analisam as tarefas, dividindo-as em ordem lógica apropriada usando as capacidades de raciocínio do FM.
- Os agentes Amazon Bedrock permitem a geração e execução dinâmica de código em um

  ambiente seguro. Isso automatiza consultas analíticas complexas que antes eram difíceis de abordar usando apenas o raciocínio do modelo.
- Os agentes do Amazon Bedrock permitem que você incorpore lógica de negócios em seu ambiente escolhido

  serviço de backend. Além disso, a funcionalidade de retorno de controle permite que você execute ações demoradas em segundo plano (de forma assíncrona) enquanto continua o fluxo de orquestração.

- Os agentes do Amazon Bedrock possuem a capacidade de manter a memória durante as interações, permitindo que eles se lembrem de conversas históricas e melhorem a precisão de tarefas com várias etapas.

Guarda-corpos de Amazon Bedrock:

O Amazon Bedrock Guardrails ajuda você a estabelecer proteções para aplicativos de IA generativa alinhados aos seus casos de uso específicos e políticas de IA responsáveis.
Os Amazon Bedrock Guardrails oferecem proteções personalizáveis adicionais além das proteções integradas dos modelos de fundação.
O Amazon Bedrock Guardrails protege seus aplicativos de IA generativa avaliando tanto os prompts do usuário quanto as respostas do modelo. • Bloqueando mais
conteúdo prejudicial. • Filtrando respostas

imprecisas para tarefas de RAG e sumarização. • Os clientes podem personalizar e implementar salvaguardas de segurança, privacidade e veracidade.
dentro de uma solução unificada.

Características:

- Os Amazon Bedrock Guardrails podem ser combinados com os Amazon Bedrock Agents e

Bases de conhecimento para criar aplicativos de IA generativos que sigam suas políticas de IA responsável.

- Os clientes podem estabelecer vários guardrails, cada um adaptado com uma combinação diferente de controles, e aplicar esses guardrails a vários aplicativos e casos de uso.
- O Amazon Bedrock Guardrails oferece filtros de conteúdo personalizáveis para rastrear informações prejudiciais

  conteúdo, incluindo discurso de ódio, insultos, material sexualmente sugestivo, violência, má conduta (incluindo atividade criminosa) e ataques imediatos (injeção imediata e jailbreak).

- O Amazon Bedrock Guardrails emprega **verificações de aterramento contextual** para identificar e filtrar alucinações quando as respostas se desviam das informações fornecidas, como serem factualmente incorretas ou
  introduzir novos dados.

PartyRock - Amazon Bedrock Playground: PartyRock é

uma ferramenta poderosa projetada para permitir que você explore e experimente os vários modelos de fundação disponíveis na plataforma Amazon Bedrock. **Ela** é projetada especificamente para entretenimento e criatividade.

Características:

- **Parque de bate-papo -** O parque de bate-papo permite que você interaja com a conversa

  modelos disponíveis no Amazon Bedrock. Quando você insere um prompt no modelo • **Playground de texto -** O playground de texto oferece uma plataforma para explorar os modelos de texto do Amazon Bedrock. Ao
  inserir um prompt de texto, você pode ver a saída gerada do modelo. • **Playground de imagem -** O playground de imagem permite que você explore os recursos de
  Modelos de imagem do Amazon Bedrock. Ao inserir uma descrição de texto, você pode ver como o modelo transforma suas palavras em uma representação visual.

Link de referência:

https://docs.aws.amazon.com/bedrock/latest/studio-ug/guardrails.html

Amazon Q

O Amazon Q é um assistente generativo com tecnologia de IA que ajuda a acelerar o desenvolvimento de software e usa dados internos das empresas.

O Amazon Q Business é um assistente generativo com tecnologia de IA que pode fornecer respostas, resumos, geração de conteúdo e conclusão segura de tarefas com base nos dados da sua empresa.

O Amazon Q Developer oferece suporte a desenvolvedores e profissionais de TI em diversas tarefas, incluindo codificação, testes, atualizações de aplicativos, diagnóstico de erros, avaliações de segurança e otimização de recursos da AWS.

Características:

O Amazon Q oferece recursos avançados de planejamento e raciocínio para transformar e implementar novos recursos de código conforme solicitado pelos desenvolvedores.

O Amazon Q é capaz de entender e respeitar identidades de governança, funções e permissões atuais, fornecendo interações personalizadas.

O Amazon Q integra-se com:

- **Amazon QuickSight**
- **Conexão Amazon**
- **Cadeia de suprimentos da AWS**

Aplicações de Modelos de Fundação

Engenharia rápida

- **Os prompts** são um conjunto específico de entradas fornecidas pelo usuário que direcionam os LLMs Amazon Bedrock para produzir resultados relevantes.
- **A engenharia de prompts** é o processo de projetar prompts de texto para obter o resultado desejado respostas de um Large Language Model (LLM).

- Os engenheiros de prompt usam métodos de tentativa e erro para gerar textos de entrada que orientam um a IA generativa do aplicativo funcione conforme o esperado.

Benefícios da Prompt Engineering:

• Maior controle do desenvolvedor • Melhor experiência do usuário • Maior flexibilidade Técnicas de engenharia rápida:

- **Solicitação de cadeia de pensamento:-** A solicitação de cadeia de pensamento envolve dividir perguntas complexas em etapas menores e lógicas, semelhantes a um processo de pensamento. Esta técnica aumenta a capacidade do modelo de raciocinar e resolver problemas de forma eficaz. •

Solicitação de árvore de pensamento:- A técnica de árvore de pensamento estende a solicitação de cadeia de pensamento gerando várias próximas etapas potenciais e avaliando cada uma usando uma abordagem de busca em árvore.

- **Solicitação maiêutica: -** A solicitação maiêutica é semelhante à solicitação da árvore de pensamento. O modelo é solicitado a responder a uma pergunta com uma explicação, seguida de solicitações para explicar partes da explicação. Explicações inconsistentes são eliminadas, melhorando o desempenho no raciocínio complexo de senso comum. • **Solicitação baseada em complexidade: -** Este método de engenharia de solicitações utiliza múltiplos rollouts de cadeia de pensamento, selecionando aqueles com as sequências de pensamento mais longas e as conclusões mais frequentes. •

Solicitação de conhecimento gerado: - Esta técnica envolve gerar fatos relevantes para dar suporte à solicitação e, em seguida, concluí-la. Essa estratégia geralmente resulta em conclusões de maior qualidade, pois o modelo é guiado por informações relevantes. • **Solicitação do menor para o maior: -** Este método envolve solicitar ao modelo que identifique e abordar subproblemas sequencialmente. Isso permite que subproblemas subsequentes se beneficiem de soluções para os anteriores. •

Solicitação de auto-refinamento: - Nesta abordagem, o modelo é solicitado a resolver um problema, avaliar sua própria solução e, então, melhorá-la considerando o problema original, sua solução e sua avaliação. Este ciclo se repete até que haja uma razão predeterminada para parar.

- **Solicitação de estímulo direcional: -** Este método de engenharia de solicitação emprega uma dica ou sugestão, como palavras-chave especificadas, para direcionar o modelo de linguagem em direção ao resultado pretendido.

Casos de uso:

Especialização no assunto:

- **Exemplo:-** Imagine um médico usando um modelo de linguagem para gerar diagnósticos potenciais para um paciente complexo. Ao inserir sintomas e detalhes do paciente, a IA, guiada por prompts cuidadosamente elaborados, pode listar possíveis doenças e restringir as opções com base em informações adicionais.

Pensamento crítico:

- **Exemplo: -** Na tomada de decisão, um modelo pode ser solicitado a avaliar várias opções, pesar seus prós e contras e sugerir o curso de ação mais adequado.

Criatividade:

- **Exemplo: -** Escritores podem usar modelos projetados por prompt para fazer brainstorming de ideias para histórias, gerando personagens, cenários e pontos da trama. Designers gráficos podem empregar esses modelos para gerar paletas de cores que transmitam emoções específicas e, então, criar designs usando essas paletas.

Recuperação de Geração Aumentada (RAG)

A Geração Aumentada de Recuperação (RAG) é um processo para aprimorar as capacidades de grandes modelos

de linguagem referenciando uma base de conhecimento confiável além de seus dados de treinamento originais, resultando em respostas mais precisas e informativas. • Grandes Modelos de Linguagem
(LLMs) são treinados em conjuntos de dados massivos com bilhões de

parâmetros para gerar saída original para tarefas como responder perguntas, traduzir idiomas e completar frases. A Retrieval Augmented Generation (RAG) aprimora as capacidades dos LLMs integrando-os com bases de conhecimento específicas de domínio ou organizacionais.

• Essa abordagem evita a necessidade de retreinamento do modelo e fornece uma maneira econômica de garantir que os resultados do LLM permaneçam precisos, relevantes e valiosos em diversos contextos.

Benefícios da geração aumentada de recuperação: •

Implementação econômica • **Informações atuais**

• **Maior confiança do usuário**

• **Mais controle do desenvolvedor**

Como funciona o Retrieval Augmented Generation?

O LLM incorpora o novo conhecimento e seus dados de treinamento para criar melhores respostas.

• **Criar dados externos-**

• Os novos dados fora do conjunto de dados de treinamento original do LLM são chamados de *dados externos.*

Esses dados podem vir de várias fontes, como APIs, bancos de dados ou repositórios de

documentos, e podem existir em diferentes formatos, como arquivos, registros de banco de dados ou texto longo.

• Outra técnica de IA, a incorporação de modelos de linguagem, transforma dados textuais em representações numéricas armazenadas em bancos de dados vetoriais, criando uma biblioteca de conhecimento que
os modelos de IA generativos podem entender.

• **Recuperar informações relevantes-**

• O próximo passo é realizar uma pesquisa de relevância. As consultas do usuário também são convertidas em vetores para recuperar informações relevantes. A relevância foi calculada e estabelecida usando cálculos e representações de vetores matemáticos.

• **Aumentar o prompt LLM-** • O modelo

RAG pode gerar respostas mais precisas e informativas para o usuário

consultas integrando esses dados recuperados aos prompts fornecidos pelo modelo de linguagem grande.

Essa técnica, conhecida como engenharia de prompt, facilita a comunicação efetiva entre o usuário e a IA.

• **Atualizar dados externos-**

- Para garantir que as informações recuperadas estejam atualizadas, é essencial atualizar os documentos e atualizar seus embeddings correspondentes de forma assíncrona. Isso pode ser alcançado por meio de processos automatizados que acontecem em tempo real ou agendando atualizações regulares.

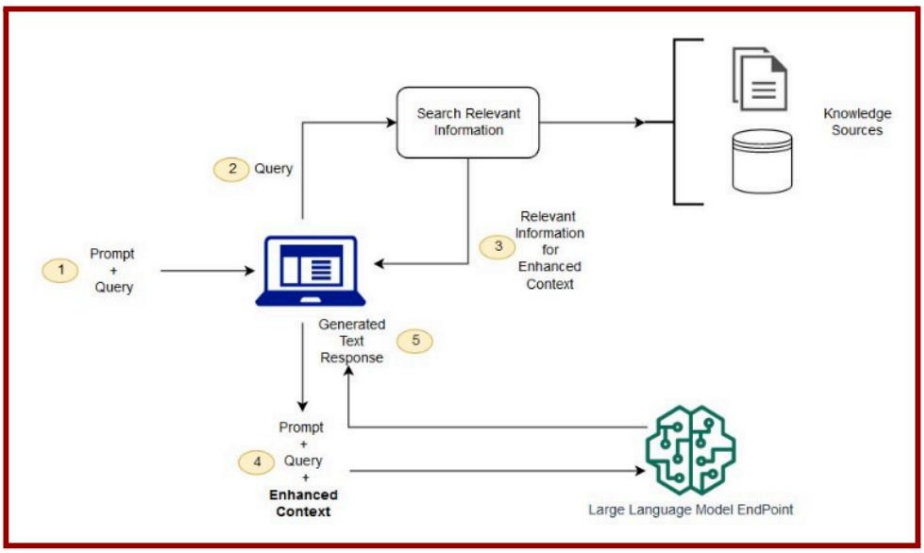


Figura: Geração Aumentada de Recuperação

RLHF - Aprendizagem por reforço a partir do feedback humano

• O aprendizado por reforço a partir do feedback humano (RLHF) é uma técnica de aprendizado de máquina que aprimora o desempenho do modelo de ML ao incorporar o feedback humano. • As técnicas de aprendizado por reforço (RL) treinam o software para tomar decisões que maximizam

recompensas, tornando seus resultados mais precisos.

Como funciona o RLHF?

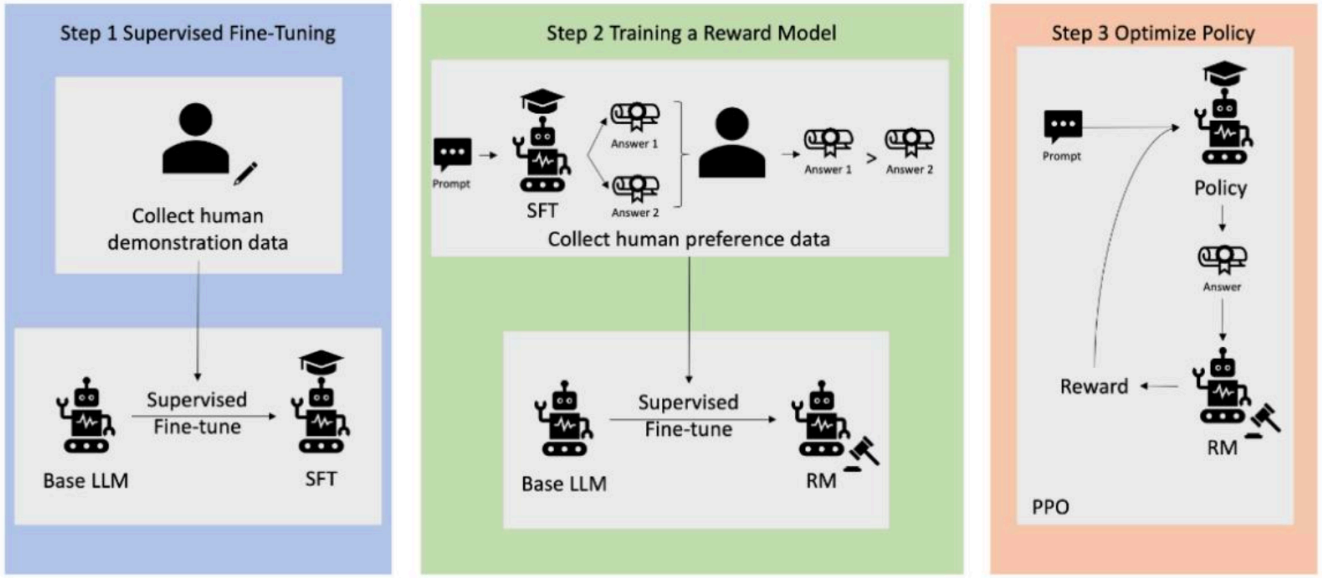
O RLHF envolve um processo de quatro etapas para preparar um modelo pronto para implantação.

- **Coleta de dados: -** Uma coleção de prompts escritos por humanos e suas respectivas respostas é estabelecida para servir como material de treinamento para o modelo de linguagem.
- **Ajuste fino supervisionado de um modelo de linguagem:-** Você pode usar um modelo comercial pré-treinado como modelo base para o processo RLHF. Você pode ajustar o modelo para a base de conhecimento interna da empresa usando a técnica de geração aumentada de recuperação (RAG). Após o ajuste fino, avalie a saída do modelo em relação a prompts predeterminados, comparando suas respostas a exemplos gerados por humanos coletados no estágio inicial.

- **Construindo um modelo de recompensa separado: -** A partir de um conjunto de múltiplas respostas do modelo respondendo a um único prompt, os avaliadores humanos podem expressar suas preferências para cada opção.

  Ao analisar essas classificações, construímos um modelo de recompensa capaz de prever automaticamente como um humano pontuaria qualquer resposta dada. •

Otimize o modelo de linguagem com o modelo baseado em recompensa:- O modelo de linguagem usa o modelo de recompensa para refinar iterativamente sua estratégia de geração de resposta. Ao avaliar internamente várias respostas potenciais, o modelo seleciona aquela que ele prevê que renderá a maior recompensa de acordo com os critérios do modelo de recompensa. Essa abordagem adaptativa garante que as saídas do modelo se alinhem mais de perto com as preferências humanas.



Aplicações do RLHF:

- RLHF pode ser usado na geração de imagens de IA: por exemplo, avaliando o grau de realismo, tecnicidade ou humor da obra de arte
- O RLHF pode gerar música que se alinha com emoções específicas ou criar trilhas sonoras. • O RLHF pode melhorar o tom de um assistente de voz, tornando-o mais acessível, curioso e confiável.

Como os serviços da AWS podem ser utilizados para atender aos requisitos do RLHF?

O Amazon SageMaker Ground Truth simplifica a tarefa de rotular e anotar dados para Aprendizado por Reforço a partir de Feedback Humano (RLHF), garantindo uma entrada humana precisa e confiável.

Estudo orientado para a recordação para avaliação de teoria [ROUGE]:

- ROUGE, ou Recall-Oriented Understudy for Gisting Evaluation, é um conjunto de métricas comumente usado em processamento de linguagem natural (NLP) para avaliar a qualidade do texto gerado por máquina. • Essas métricas

se concentram principalmente na comparação do texto gerado e do texto de referência da verdade básica (escrito por humanos) de um conjunto de dados de validação. •
As medidas ROUGE são projetadas para avaliar vários aspectos da similaridade do texto, como a precisão e a recordação de n-gramas (sequências contíguas de palavras) em dados gerados pelo sistema.
e textos de referência.

- O objetivo é determinar a eficácia com que o modelo pode gerar texto semelhante a o conteúdo original.

Diretrizes para IA Responsável

IA responsável

IA responsável se refere ao desenvolvimento de sistemas de IA que sejam justos, transparentes, responsáveis, seguros e imparciais.

Componentes da IA responsável:

ÿ **Justiça -** Considerando os efeitos potenciais em grupos diversos. ÿ **Explicabilidade**

- Entendendo e avaliando saídas do sistema. ÿ **Privacidade e segurança -**

Obtendo, utilizando e protegendo dados e modelos. ÿ **Segurança -** Salvaguardando contra falhas do sistema e uso malicioso. ÿ **Controlabilidade -** Estabelecendo mecanismos para monitorar e controlar ações de IA. ÿ **Veracidade e robustez -** Alcançando saídas precisas do sistema, tanto em condições normais quanto adversas.

ÿ **Governança -** Promover IA responsável integrando as melhores práticas em todo o fornecimento de IA corrente.

ÿ **Transparência -** Garantir que as partes interessadas tenham o conhecimento necessário para se envolver efetivamente com o sistema de IA.

Serviços e ferramentas para construir IA Responsável: avaliações do modelo de base (FM):-

• Avaliação de modelo no Amazon Bedrock • Amazon SageMaker Esclarecer

Implementando salvaguardas em IA generativa:- • Guardrails no Amazon Bedrock

Detectando viés:-

Os vieses são desequilíbrios nos dados ou disparidades no desempenho de um modelo em diferentes grupos. •

Amazon SageMaker Esclarecer

Explicando as previsões do modelo:-

• **Amazon SageMaker Esclarecer Monitoramento e revisão humana** • Amazon SageMaker Model Monitor • IA aumentada da Amazon

Melhorar a governação

• Governança de ML do Amazon SageMaker

Amazon SageMaker Esclarecer

O aprendizado de máquina oferece oportunidades para identificar e medir vieses em todo o ciclo de vida do ML.

O Amazon SageMaker Clarify ajuda a detectar viés em dados e modelos antes, durante e depois do treinamento: 1.

Viés pré-

treinamento: detecta viés nos dados brutos antes do início do treinamento do modelo.

2. **Viés pós-treinamento:** mede o viés nas saídas do modelo após o treinamento.

3. **Monitoramento de viés:** monitore continuamente o viés nas previsões do modelo após a implantação.

Benefícios do SageMaker Clarify:

- Avalie modelos de fundação (FMs) em minutos: - Automatize a avaliação de fundações

  modelos para suas aplicações de IA generativa com base em critérios como precisão, resiliência e viés para manter princípios de IA responsáveis.
- Crie confiança nos modelos de ML: - Avalie o desempenho do seu FM durante a personalização usando métodos automatizados e baseados em humanos.
- Métricas e relatórios acessíveis e baseados em ciência: - Gere métricas, relatórios e exemplos práticos fáceis de usar para dar suporte à personalização de FM e ao fluxo de trabalho de MLOps.

Estratégia do SageMaker Clarify para lidar com preconceitos

- **Métricas de viés:** o SageMaker Clarify oferece métricas independentes de modelo para medir viés e justiça baseada em diferentes conceitos de justiça.
- **Automação:** O SageMaker Clarify automatiza a detecção e o monitoramento de vieses em todo o o ciclo de vida do ML.
- **Monitoramento de dados:** o SageMaker Clarify rastreia o viés nas previsões do modelo após a implantação, garantindo supervisão contínua do comportamento do modelo.
- **Ferramentas para detecção de viés**

- **Blocos de notas de exemplo do SageMaker Clarify:** o SageMaker Clarify fornece um bloco de notas para detecção

  de viés e explicabilidade, ajudando os usuários a executar trabalhos de detecção de viés e interpretar atribuições de recursos.

O notebook de amostra para detecção de viés pode ser executado no **Amazon SageMaker Studio** usando **Python 3 (Data Science).** Ele percorre o processo de detecção de viés e explica as previsões do modelo.
O SageMaker Clarify oferece ferramentas abrangentes para medir e monitorar vieses, ajudando a garantir que os modelos de aprendizado de máquina sejam justos e imparciais em todos os estágios de desenvolvimento e implantação.

Monitor de modelo do Amazon SageMaker

O Amazon SageMaker Model Monitor monitora a qualidade dos modelos de aprendizado de máquina do Amazon SageMaker em produção.

O Model Monitor permite que você implemente monitoramento contínuo usando várias abordagens;

• Monitoramento contínuo com um ponto de extremidade em tempo

real. • Monitoramento contínuo com um trabalho de transformação em lote executado

regularmente. • Monitoramento dentro do cronograma para trabalhos de transformação em lote assíncronos.

- O Model Monitor permite que você estabeleça alertas que notificam a qualidade do modelo

  desvios. Ao identificar prontamente esses desvios, você pode tomar ações corretivas, como retreinar modelos, auditar sistemas upstream ou corrigir problemas de qualidade.

- O Model Monitor oferece opções de monitoramento pré-construídas e personalizáveis. Você pode usar os recursos pré-construídos para configuração rápida ou escrever seu código para análise personalizada.

O Model Monitor oferece vários tipos de monitoramento:

• **Monitorar a qualidade dos dados -** Monitorar o desvio na qualidade

dos dados. • **Monitorar a qualidade do modelo -** Monitorar o desvio nas métricas de qualidade do modelo,

como precisão. • **Monitorar o desvio de viés para modelos em produção -** Monitorar o viés nas previsões do seu modelo. • **Monitorar o desvio de atribuição de recurso para modelos em produção -** Monitorar o desvio no recurso

atribuição.

O Amazon SageMaker Model Monitor monitora continuamente o desempenho dos seus modelos de machine learning em produção. Ele alerta você automaticamente se quaisquer alterações significativas ou problemas de qualidade forem detectados. O Model Monitor usa regras para identificar desvios do comportamento esperado e notificá-lo imediatamente.

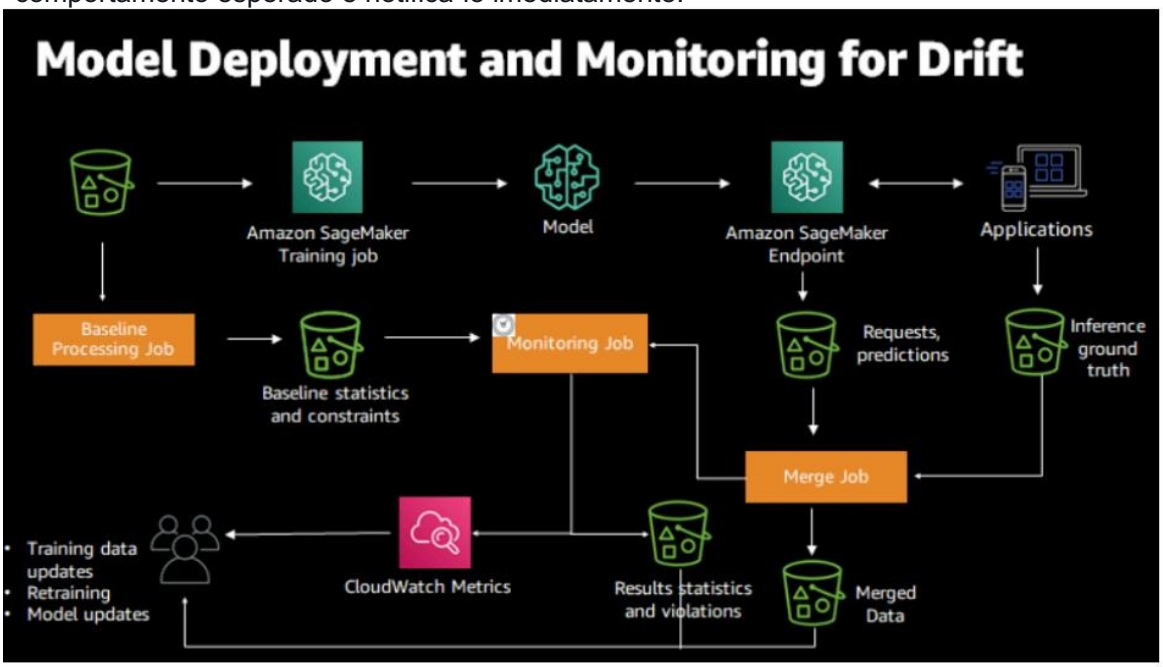


Figura: Monitor de modelo do Amazon SageMaker

Cartões de modelo do Amazon SageMaker:

- A governança do modelo é uma estrutura que fornece visibilidade sistemática da máquina

  desenvolvimento, validação e uso de modelos de aprendizado (ML). O Amazon SageMaker fornece ferramentas de governança de ML desenvolvidas especificamente para gerenciar acesso de controle, rastreamento de atividades e relatórios em todo o ciclo de vida de ML.
- Os cartões de modelo do Amazon SageMaker são essencialmente um modelo padronizado para documentar, recuperar e compartilhar informações essenciais do modelo, do desenvolvimento à implantação.

Características:

- Fornecer orientação sobre o uso apropriado do modelo. • Apoiar os

processos de auditoria, oferecendo informações detalhadas sobre o treinamento do modelo e métricas de desempenho. •
Comunique o valor comercial específico que o modelo deve entregar.

Serviços de IA gerenciados pela AWS

Amazonas Polly

O que é Amazon Polly?

Conversão de texto em fala: transforma texto escrito em palavras faladas.

Serviço baseado em nuvem: opera na nuvem, eliminando a necessidade de infraestrutura local.

Opções de voz: oferece várias opções de voz, incluindo conversão de texto em fala neural (NTTS) para uma fala com som natural.

Características

- **Sem taxas de configuração:** pague apenas pelo texto convertido em fala, sem custos iniciais de configuração.
- **Suporte multilíngue:** acesse vários idiomas e vozes de conversão de texto em fala neural (NTTS) para criar aplicativos habilitados para fala.
- **Cache de fala e repetição:** utilize recursos de cache e repetição para Amazon Polly's fala gerada, disponível em formatos como MP3.

Amazon Comprehend

O que é o Amazon Comprehend?

Processamento de Linguagem Natural (PLN): usa PNL para extrair insights do conteúdo do documento.

Extração de insights: identifica entidades, frases-chave, linguagem, sentimentos e outros elementos dentro de documentos.

Desenvolvimento de produtos: aproveite a compreensão da estrutura do documento para desenvolver novos produtos.

Características:

- **Extraia insights de diversas fontes de texto:** analise textos de documentos, tickets de suporte, avaliações de produtos, e-mails e mídias sociais para descobrir insights valiosos. • **Otimize o processamento de documentos:** melhore os fluxos de trabalho extraindo informações importantes,

  incluindo texto, frases, tópicos e sentimentos de documentos como reivindicações de seguros.

- **Classificação de documentos personalizada:** diferencie seu negócio treinando modelos para classificar documentos e reconhecer termos específicos, sem precisar de habilidades avançadas de aprendizado de máquina.
- **Proteja informações confidenciais:** proteja e gerencie o acesso a dados confidenciais identificando e redigindo informações de identificação pessoal (PII) em documentos.

Casos de uso:

- Analisar dados comerciais e de call center •

  Indexar e pesquisar avaliações de produtos • Gerenciar resumos jurídicos • Lidar com processamento de documentos financeiros

Amazon Rekognition

O que é Amazon Rekognition?

Serviço baseado em nuvem: utiliza tecnologia avançada de visão computacional para análise de imagens e vídeos.

Não é necessária experiência em aprendizado de máquina: acessível por meio de uma API intuitiva.

Integração com o Amazon S3: analisa rapidamente imagens e vídeos armazenados no Amazon S3.

Principais

recursos: Inclui detecção de objetos e texto, identificação de conteúdo inseguro e análise facial.

Análise facial •

Verificação do usuário: identifica e verifica identidades individuais.

• **Catalogação:** organiza e gerencia dados faciais para vários aplicativos. • **Segurança pública:** aumenta a segurança por meio de vigilância e monitoramento. • Detecte, analise e compare rostos em transmissões ao vivo e vídeos gravados.

Análise de imagem:

• **Detecção de objetos, cenas e conceitos:** detecte e classifique vários objetos, cenas, conceitos e celebridades presentes nas imagens.
• **Detecção de texto:** identifique texto impresso e manuscrito em imagens, apoiando vários idiomas.

Análise de vídeo: •

Detecção de objetos, cenas e conceitos: categorize objetos, cenas, conceitos e celebridades aparecendo em vídeos.
• **Detecção de texto:** reconheça texto impresso e manuscrito em vídeos em diferentes idiomas. • **Rastreamento de pessoas:**

monitore indivíduos identificados em vídeos conforme eles se movem pelos quadros.

Casos de uso:

• Simplifique a recuperação de conteúdo com a análise automática do Amazon Rekognition, permitindo fácil capacidade de pesquisa de imagens e vídeos.
• Aumente a segurança com a detecção de vivacidade facial do Rekognition, evitando falsificação de identidade além das senhas tradicionais.
• Localize rapidamente indivíduos em seu conteúdo visual usando o reconhecimento facial eficiente do Rekognition recurso de pesquisa.
• Garanta a segurança do conteúdo com a capacidade do Rekognition de detectar conteúdo explícito, inapropriado e conteúdo violento, facilitando a filtragem proativa.
• Beneficie-se da elegibilidade HIPAA, tornando o Amazon Rekognition adequado para lidar com informações de saúde protegidas em aplicações de assistência médica.

Amazon Lex

O que é o Amazon Lex?

Crie chatbots: crie interfaces de conversação usando processamento de linguagem natural.

Aproveita a tecnologia Alexa: utiliza a mesma tecnologia que capacita a Alexa para compreensão avançada de idiomas.

Integração perfeita: integra-se facilmente com outros serviços da AWS para aprimorar a funcionalidade do chatbot e a experiência do usuário.

Características:

- Integre sem esforço a IA que compreende a intenção, retém o contexto e automatiza as tarefas básicas tarefas em vários idiomas.
- Projete e implante IA conversacional omnicanal com um único clique, sem a necessidade de gerenciar hardware ou infraestrutura. • Conecte-se

perfeitamente com outros serviços da AWS para acessar dados, executar lógica de negócios, monitorar desempenho e muito mais.

- Pague somente por solicitações de fala e texto, sem custos iniciais ou taxas mínimas.

Casos de uso:

- **Habilitar agentes virtuais e assistentes de voz:** Ofereça aos usuários opções de autoatendimento por meio de agentes de contact center virtuais e resposta de voz interativa (IVR), permitindo que eles executem tarefas de forma autônoma, como agendar compromissos ou alterar senhas.

- **Automatize respostas a perguntas frequentes:** desenvolva soluções de conversação que respondam a perguntas comuns, aprimorando os fluxos de conversação do Connect & Lex com pesquisa em linguagem natural para perguntas frequentes, fornecida pelo Amazon Kendra.
- **Melhore a produtividade com bots de aplicativos:** simplifique as tarefas do usuário em aplicativos usando chatbots eficientes, integrando-se perfeitamente ao software empresarial por meio do AWS Lambda e mantendo o controle de acesso preciso via IAM.
- **Extraia insights de transcrições:** crie chatbots usando transcrições do contact center para maximizar as

informações capturadas, reduzindo o tempo de design e agilizando a implantação do bot de semanas para horas.

Transcrição da Amazon

O que é o Amazon Transcribe?

Conversão de fala em texto: transforma fala em áudio em texto escrito.

Tecnologia de aprendizado profundo: utiliza reconhecimento automático de fala (ASR) para transcrição precisa.
Aplicações versáteis: Ideal para gerar transcrições de diversas fontes de áudio, como reuniões, entrevistas e vídeos.

Características

- **Otimizado para casos de uso específicos:** ideal para chamadas de atendimento ao cliente, transmissões ao vivo e

  legendagem de mídia.
- **Transcrição médica:** converte discurso médico em texto para documentação clínica com alta precisão.
- **Estrutura de custos:** as cobranças são baseadas nos segundos de fala convertidos por mês.

Casos de uso

- **Atendimento ao cliente:** melhore as interações com o cliente transcrevendo chamadas de serviço para análise e melhoria.
- **Transmissões ao vivo:** gere legendas em tempo real para eventos e transmissões ao vivo. •

Documentação médica: simplifique a documentação clínica transcrevendo o discurso médico com precisão.

Amazon Translate

O que é o Amazon Translate?

Tradução automática neural: usa redes neurais para traduções de texto precisas e naturais.

Pares de idiomas: traduz texto entre inglês e vários outros idiomas.

Conversão de origem-destino: converte texto de um idioma de origem para um idioma de destino com base em pares de idiomas selecionados.

Benefícios do Amazon Translate:

- **Traduções de alta qualidade -** Fornece traduções precisas e em evolução em vários aplicações.
- **Traduções em lote e em tempo real -** Integre traduções em lote e em tempo real em seu aplicativos perfeitamente usando uma única chamada de API.
- **Personalização -** Personalize sua saída traduzida por ML para definir nomes de marcas, modelos nomes e outros termos exclusivos.

Casos de uso:

- **Traduza conteúdo gerado pelo usuário:** traduza automaticamente conteúdo gerado pelo usuário, incluindo postagens de mídia social, perfis e comentários, instantaneamente em tempo real. • **Analise conversas on-line em diferentes idiomas:** use um aplicativo de processamento de linguagem natural para analisar texto em vários idiomas e obter insights sobre a opinião pública sobre sua marca, produto ou serviço. • **Crie comunicações multilíngues entre usuários:** implemente recursos de tradução de idiomas em tempo real em sistemas de bate-papo, e-mail, helpdesk e emissão de tíquetes para permitir que agentes que falam inglês se comuniquem efetivamente com clientes em todo o mundo.

Amazon Mechanical Turk (MTurk)

O que é o Amazon Mechanical Turk?

- **Mercado de Crowdsourcing:** O MTurk conecta indivíduos e empresas com um mercado global, força de trabalho virtual para diversas tarefas.
- **Tipos de tarefas:** inclui validação de dados simples, pesquisa, participação em pesquisas, moderação de conteúdo e muito mais.
- **Como funciona:** Os solicitantes postam tarefas (HITs) que os trabalhadores concluem on-line. O sistema

  garante que os trabalhadores sejam pagos apenas pelo trabalho satisfatório, e você pode usar testes de qualificação para selecionar trabalhadores qualificados.

Vantagens •

Eficiência aprimorada: automatize tarefas manuais repetitivas para agilizar os fluxos de trabalho. MTurk ajuda a concluir tarefas rapidamente, liberando recursos internos para um trabalho mais estratégico.

- **Escalabilidade flexível:** aumente ou diminua facilmente a força de trabalho sem as complexidades de gerenciar uma equipe interna temporária. O MTurk fornece acesso a uma equipe global 24x7
  força de trabalho.

- **Redução de custos:** Reduza os custos de mão de obra e despesas gerais com um modelo de pagamento por tarefa. O MTurk

  ajuda a gerenciar despesas de forma eficaz, ao mesmo tempo em que alcança resultados que podem ser desafiadores com uma equipe dedicada.

Por que usar o MTurk? •

Experiência humana: conclui tarefas melhor que computadores, como moderação de conteúdo e desduplicação de dados.

- **Crowdsourcing eficiente:** divide projetos complexos em microtarefas gerenciáveis para trabalhadores distribuídos, melhorando a escalabilidade e reduzindo o esforço manual.

Casos de uso do MTurk em coleta e anotação de dados de aprendizado de máquina

- **Coleta de dados eficiente:** o MTurk simplifica a coleta e a rotulagem de grandes conjuntos de dados necessários para treinar modelos de ML. Ele acelera o processo de anotação de dados, como marcação de imagens ou categorização de texto.

Melhoria do modelo

- **Iteração Contínua:** Use o MTurk para ajustes e aprimoramentos contínuos em modelos de ML. A entrada humana ajuda a refinar modelos fornecendo feedback e fazendo as correções necessárias.

Humano no circuito (HITL)

- **Incorporando feedback humano:** o MTurk oferece suporte a fluxos de trabalho HITL onde os humanos

  o feedback é essencial para validação e retreinamento do modelo. Por exemplo, anotar imagens com caixas delimitadoras ajuda a criar conjuntos de dados precisos para tarefas de visão computacional, especialmente quando soluções automatizadas falham.

IA aumentada da Amazon [Amazon A2I]

• Amazon Augmented AI (Amazon A2I) é um serviço que traz revisão humana de ML

previsões para todos os desenvolvedores, removendo o trabalho pesado associado à construção de sistemas de revisão humana ou ao gerenciamento de um grande número de revisores humanos. • Por exemplo, extrair informações de formulários de solicitação de hipoteca digitalizados pode exigir revisão humana devido a digitalizações de baixa qualidade ou caligrafia ruim. • A construção de sistemas de revisão humana pode ser demorada e cara porque

envolve a implementação de fluxos de trabalho complexos, a criação de software personalizado para gerenciar tarefas e resultados de revisão e o gerenciamento de grandes grupos de revisores.

O Amazon A2I oferece processos de revisão humana integrados para tarefas típicas de ML, como moderação de conteúdo e extração de texto de documentos.

• Moderação de conteúdo

• Extração de formulários

• Classificação de imagens

Casos de uso:

ÿ **Use o Amazon A2I com o Amazon Textract: -** Use o Amazon Textract para selecionar e envie documentos do seu conjunto de dados para humanos para revisão.

ÿ **Use o Amazon A2I com o Amazon Rekognition:-** Use o Amazon Textract para selecionar aleatoriamente

e envie imagens do seu conjunto de dados para humanos para revisão.

ÿ **Use o Amazon A2I para revisar inferências de ML em tempo real: -** Use o Amazon A2I para revisar

inferências de baixa confiança em tempo real feitas por um modelo implantado em um SageMaker.

Refine continuamente seu modelo incorporando feedback dos dados de saída do A2I. ÿ **Use o Amazon A2I com o Amazon Comprehend: -** Tenha humanos para revisar inferências sobre dados de texto usando o Amazon Comprehend. ÿ

Use o Amazon A2I com o Amazon Transcribe: -

Tenha humanos para revisar transcrições de

arquivos de vídeo ou áudio usando o Amazon Transcribe.

AWS DeepRacer

O AWS DeepRacer é um veículo autônomo em escala 1/18 habilitado para aprendizado por reforço (RL), projetado para fins educacionais com serviços de suporte no ecossistema de aprendizado de máquina da AWS.

- O AWS DeepRacer oferece uma plataforma para desenvolver e testar modelos de aprendizado de reforço profundo. Você pode treinar esses modelos em um ambiente simulado e, em seguida, implantá-los em um carro físico do AWS DeepRacer para corrida autônoma.

Veículos AWS DeepRacer:

1. O dispositivo original do AWS DeepRacer é um modelo físico de carro em escala 1/18 com uma câmera montada e um módulo de computação integrado. O módulo de computação executa inferência para dirigir-se ao longo de uma pista. O módulo de computação e o chassi do veículo são alimentados por baterias dedicadas conhecidas como bateria de computação e bateria de acionamento, respectivamente.
2. O carro de corrida virtual se torna o dispositivo original AWS DeepRacer, o dispositivo Evo ou

   várias recompensas digitais que podem ser ganhas ao participar das corridas do AWS DeepRacer League Virtual Circuit.
3. O dispositivo AWS DeepRacer Evo é o dispositivo original com um kit de sensor opcional. O kit inclui uma câmera adicional e LIDAR (detecção e alcance de luz), que permitem que o carro detecte objetos atrás e laterais a si mesmo.

Segurança, conformidade e governança para soluções de IA

Amazonas Macie

O que é Amazon Macie? • Amazon Macie é uma solução de segurança de dados que emprega algoritmos de aprendizado de máquina e técnicas de reconhecimento de padrões para identificar e proteger dados confidenciais. • Ao aproveitar os recursos de aprendizado de máquina e correspondência de padrões, o Amazon Macie não apenas detecta dados confidenciais, mas também oferece insights sobre potenciais ameaças à segurança de dados. • Além disso, ele facilita medidas automatizadas para mitigar esses riscos, melhorando a segurança geral.

Características:

- Implementar processos automatizados para detectar dados confidenciais em larga escala.
- Use o Amazon Macie com o Amazon Textract, Amazon Rekognition e Amazon SageMaker para descobrir e proteger dados confidenciais armazenados no Amazon S3.

Link privado da AWS

O que é AWS PrivateLink?

- O AWS PrivateLink é um serviço de rede usado para conectar-se a serviços da AWS hospedados por outras contas da AWS (chamados de serviços de endpoint) ou ao AWS Marketplace. • Sempre que um endpoint de interface VPC (endpoint de interface) é criado para serviço no VPC, uma Elastic Network Interface (ENI) na sub-rede necessária com um endereço IP privado também é criada, servindo como um ponto de entrada para o tráfego destinado ao serviço.

Pontos de extremidade da interface • Serve como um ponto de entrada para o tráfego destinado a um serviço AWS ou um ponto de extremidade VPC serviço. Pontos de extremidade do gateway

- É um gateway na tabela de rotas que roteia o tráfego apenas para o Amazon S3 e o DynamoDB.

Características:

- Proporciona segurança ao não permitir a internet pública e reduzir a exposição a ameaças, como força bruta e ataques DDoS.
- Com a ajuda do AWS PrivateLink, o endpoint da interface VPC conecta seu VPC diretamente à API do SageMaker ou ao SageMaker Runtime sem usar um gateway de internet, dispositivo NAT, conexão VPN. • Defina um endpoint de interface VPC para o Amazon Rekognition conectar seu VPC a Reconhecimento da Amazon.

Capacidades das ferramentas de análise de custos da AWS

- **Explorador de custos da AWS:**

  - Fornece uma interface visual para analisar e rastrear o uso e os custos da AWS. • Permite que os usuários visualizem dados históricos, prevejam custos futuros e dividam os gastos por serviço ou conta. • Ajuda a identificar oportunidades de economia de custos com recomendações sobre Instâncias reservadas e planos de economia. •

Faturamento e gerenciamento de custos da AWS:

- Um hub central para gerenciar faturamento, pagamentos e orçamentos da AWS. • Oferece insights detalhados sobre suas cobranças mensais, permitindo que você configure o faturamento alertas.
- Fornece ferramentas para gerenciar e otimizar orçamentos da AWS, incluindo a definição de limites de custo e uso.

- **Consultor confiável da AWS:**

  - Fornece recomendações em tempo real para otimizar os custos da AWS.
  - Detecta recursos subutilizados que podem ser redimensionados ou desligados para um tamanho menor despesas.
  - Oferece sugestões em outras áreas, como desempenho, segurança e falhas tolerância.

Técnicas para Rastreamento e Alocação de Custos

- **Marcação de recursos:**

  - Anexa metadados personalizados (tags) a recursos da AWS, como instâncias, bancos de dados ou buckets do S3 para rastrear custos por departamento, projeto ou ambiente. • Melhora a precisão da alocação de custos, tornando mais simples identificar quais equipes ou projetos estão usando mais recursos.

- **Tags de alocação de custos:**

  - Tags personalizadas e geradas pela AWS que permitem organizar e alocar custos em áreas de negócios específicas.
  - Essas tags podem ser usadas para filtrar e agrupar dados de custo no AWS Cost Explorer, AWS Faturamento e relatórios.

- **Contas vinculadas:**

  - Permite que várias contas da AWS sejam agrupadas em uma única entidade de cobrança, permitindo faturamento unificado e fácil rastreamento de custos em vários departamentos ou equipes.
  - Facilita a alocação de custos entre contas, ajudando a rastrear e analisar gastos entre diferentes unidades ou equipes.

Funções, políticas e grupos do IAM

1. **Funções do IAM:**

   - **Definição:** as funções do IAM são identidades da AWS com permissões específicas que podem ser assumidas por serviços ou usuários da AWS para executar determinadas ações.
   - **Caso de uso de aprendizado de máquina:** no AWS Machine Learning, as funções permitem que serviços como Amazon SageMaker, AWS Glue e outros acessem recursos como buckets do S3 ou tabelas do DynamoDB.

   - **Exemplo:** Um trabalho de treinamento do Amazon SageMaker requer acesso a um bucket do S3 para recuperar dados de treinamento. Você cria uma função do IAM com as permissões necessárias e a atribui ao trabalho de treinamento do SageMaker.

2. **Políticas do IAM:**

   - **Definição:** as políticas do IAM são documentos no formato JSON que descrevem as permissões para

     Funções ou usuários do IAM. Eles determinam quais ações são permitidas ou restritas em AWS específicos recursos.

   - **Aplicação de Aprendizado de Máquina:** Essas políticas regulam o acesso ao aprendizado de máquina

     recursos. Por exemplo, eles podem permitir que o Amazon SageMaker recupere dados de ou salvar resultados em buckets S3.

   - **Exemplo:** Uma política pode conceder ao SageMaker a capacidade de executar s3:GetObject e

     s3:PutObject comandos em um bucket S3 designado, permitindo que ele acesse e armazene conjuntos de dados de treinamento e modelo **3. Grupos**

IAM: • **Definição:**

grupos IAM são coleções de usuários IAM que compartilham as mesmas permissões. Por

Ao atribuir políticas a um grupo, todos os membros herdam as permissões do grupo. • **Caso de uso de aprendizado de máquina:** os grupos podem ser usados para gerenciar permissões para equipes trabalhando em projetos de aprendizado de máquina, garantindo controles de acesso consistentes entre os membros da equipe.

- **Exemplo:** você pode criar um grupo para cientistas de dados com permissões para acessar o SageMaker, o S3 e outros serviços relacionados a ML, simplificando o gerenciamento de permissões para vários usuários.

4. **AWS Identity and Access Management (IAM):** • **Definição:** IAM é

um serviço da AWS que facilita o gerenciamento seguro do acesso à AWS. serviços e recursos.

- **Aplicação de Machine Learning:** O IAM permite a criação e o gerenciamento de funções,

  políticas e grupos para controlar quem e quais serviços podem interagir com seus recursos e dados de aprendizado de máquina.

- **Exemplo:** Configurar o IAM para o Amazon SageMaker envolve a configuração de funções e políticas

  que determinam quais usuários ou serviços podem iniciar trabalhos de treinamento, implantar modelos e acessar dados.

5. **Políticas de Bucket:**

- **Definição:** As políticas de bucket são políticas de controle de acesso aplicadas diretamente ao Amazon S3 buckets para definir quem pode acessar o bucket e seus objetos.
- **Caso de uso de aprendizado de máquina:** para tarefas de ML, as políticas de bucket controlam o acesso aos dados armazenados no S3 que são usados para treinamento, validação e teste de modelos de aprendizado de

máquina. • **Exemplo:** uma política de bucket pode permitir apenas funções específicas do IAM associadas ao SageMaker para ler dados de um bucket S3.

6. **Gerenciador de funções do SageMaker:**

- **Definição:** O SageMaker Role Manager é um recurso do Amazon SageMaker que facilita a criação e o gerenciamento de funções do IAM usadas pelos serviços do SageMaker.
- **Caso de uso de aprendizado de máquina:** simplifica o processo de atribuição das informações necessárias permissões para trabalhos e endpoints do SageMaker, garantindo acesso seguro aos recursos.
- **Exemplo:** Ao criar um trabalho de treinamento do SageMaker, o Role Manager ajuda você

  associar uma função do IAM ao trabalho, fornecendo as permissões necessárias para acessar dados do S3 e gravar saídas.

Esses componentes trabalham juntos para garantir acesso seguro e controlado aos recursos de aprendizado de máquina da AWS, ajudando a gerenciar permissões de forma eficaz e, ao mesmo tempo, mantendo a integridade e a segurança dos seus fluxos de trabalho de ML.

Segurança e conformidade para Amazon SageMaker

Ambientes Isolados

- **Configuração de VPC privada:** implante componentes do SageMaker, como Studio, notebooks, trabalhos de treinamento e instâncias de hospedagem em uma Virtual Private Cloud (VPC) sem conectividade com a Internet. Esse arranjo mantém os recursos do SageMaker protegidos do acesso externo à Internet, reforçando a segurança.
- **Restrição de acesso à Internet:** durante a configuração do SageMaker Studio ou notebooks, escolha a configuração de acesso à rede somente VPC para impedir o acesso direto à Internet. Essa configuração restringe a conectividade à Internet e garante que todo o tráfego de dados permaneça dentro da AWS. rede.

Endpoints e políticas de VPC

- **Endpoints de interface:** use endpoints de interface VPC para se conectar a serviços da AWS como S3 e APIs do SageMaker sem expor dados à internet pública. Isso garante que a comunicação permaneça segura dentro da infraestrutura da AWS.
- **Políticas de endpoint:** defina políticas de endpoint VPC para controlar quem pode acessar recursos e quais ações eles podem executar. Por exemplo, restringir o acesso ao bucket S3 a determinados domínios ou usuários do SageMaker Studio.

Controle de acesso e segurança

- **Acesso restrito a VPC:** implemente políticas de IAM para restringir o acesso ao SageMaker recursos, garantindo que somente usuários dentro do VPC possam se conectar ao SageMaker Studio ou notebooks. As políticas podem especificar endereços IP permitidos ou endpoints VPC.
- **Detecção e prevenção de intrusão:** utilize o AWS Gateway Load Balancer (GWLB) para integrar dispositivos de segurança de terceiros, como firewalls e sistemas de detecção de intrusão, em sua rede AWS. Isso ajuda a monitorar e gerenciar o tráfego de rede de forma eficaz.

Medidas de segurança adicionais

- **Gateway NAT:** para serviços ou recursos fora da AWS que não suportam endpoints VPC, configure um gateway NAT. Configure grupos de segurança para gerenciar conexões de saída.
- **Firewall de rede AWS:** use o Firewall de rede AWS para filtrar entradas e saídas tráfego da web. Ele suporta filtragem da web para tráfego não criptografado e permite o bloqueio de sites específicos para tráfego criptografado por meio de Server Name Indication (SNI).

Essas medidas garantem que os ambientes e dados do SageMaker permaneçam seguros, em conformidade e isolados de acesso não autorizado.

Referência:

https://docs.aws.amazon.com/whitepapers/latest/ml-best-practices-public-sector-organizations/security- and-compliance.htm