



Avaliação generativa da carga de trabalho de IA

# AWS Recomendações



# AWS Recomendações: Avaliação generativa da carga de trabalho de IA

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

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

# Table of Contents

|                                                                       |    |
|-----------------------------------------------------------------------|----|
| Introdução .....                                                      | 1  |
| Objetivo deste guia .....                                             | 2  |
| Público-alvo e benefícios .....                                       | 2  |
| Escopo .....                                                          | 2  |
| Resultados de negócios desejados .....                                | 4  |
| Considerações e pré-requisitos de avaliação .....                     | 7  |
| Comece com casos de uso claros .....                                  | 7  |
| Garanta o alinhamento dos negócios .....                              | 8  |
| Implemente governança e supervisão .....                              | 8  |
| Dados de endereço e pré-requisitos técnicos .....                     | 8  |
| Considere os requisitos de recursos computacionais .....              | 8  |
| Abordar as implicações de privacidade e segurança .....               | 9  |
| Envolve as partes interessadas cedo .....                             | 9  |
| Faça iterações e aprenda .....                                        | 9  |
| Questionário generativo de avaliação da carga de trabalho de IA ..... | 10 |
| Prontidão .....                                                       | 11 |
| Casos de uso .....                                                    | 13 |
| Arquitetura .....                                                     | 16 |
| Armazenamento .....                                                   | 17 |
| Regulamentos e conformidade .....                                     | 19 |
| Integração .....                                                      | 20 |
| Teste .....                                                           | 22 |
| Implantação e automação .....                                         | 23 |
| Estratégia de dados .....                                             | 26 |
| Traduzindo insights de avaliação em resultados acionáveis .....       | 29 |
| Próximas etapas .....                                                 | 31 |
| Perguntas frequentes .....                                            | 32 |
| Qual é o objetivo principal? .....                                    | 32 |
| Quem deve usar essa avaliação? .....                                  | 32 |
| Quais são os principais componentes? .....                            | 32 |
| Como isso ajuda a definir a arquitetura? .....                        | 32 |
| Quais são os benefícios? .....                                        | 32 |
| Como podemos implementar isso com sucesso? .....                      | 33 |
| Quais são os desafios? .....                                          | 33 |

|                                                                   |        |
|-------------------------------------------------------------------|--------|
| Quais são os requisitos regulatórios e de conformidade? .....     | 33     |
| Qual é o papel das partes interessadas? .....                     | 33     |
| Como podemos medir o sucesso? .....                               | 34     |
| Como a abordagem difere com base no tamanho da organização? ..... | 34     |
| Recursos .....                                                    | 35     |
| Histórico do documento .....                                      | 36     |
| .....                                                             | xxxvii |

# Avaliação generativa da carga de trabalho de IA

Tabby Ward e Deepak Dixit, da Amazon Web Services (AWS)

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

A avaliação generativa da carga de trabalho de IA é um método estratégico que visa avaliar e melhorar a preparação de uma organização para criar ou atualizar suas cargas de trabalho generativas de IA. Essa avaliação é importante porque a incorporação da IA generativa às operações comerciais pode mudar muito a forma como as coisas funcionam e fornecer novas eficiências e capacidades. No entanto, para adotar a IA generativa com sucesso, é essencial entender completamente os sistemas atuais e ter um plano claro para o futuro.

As cargas de trabalho generativas de IA se referem a tarefas computacionais que envolvem o uso de modelos de inteligência artificial que podem criar novos conteúdos, como texto, imagens, código ou outros tipos de dados. Essas cargas de trabalho normalmente exigem poder computacional substancial, hardware especializado, como GPUs, e grandes conjuntos de dados para treinamento e inferência. A integração de cargas de trabalho generativas de IA às operações apresenta vários desafios:

- Requisitos de infraestrutura: provisionamento dos recursos computacionais significativos e do hardware especializado que os modelos generativos de IA exigem.
- Gerenciamento de dados: Garantindo a qualidade, a privacidade e a conformidade dos dados ao lidar com grandes conjuntos de dados.
- Lacuna de habilidades: falta de experiência em tecnologias de IA e implantação de modelos.
- Considerações éticas: lidar com preconceitos, justiça e transparência no conteúdo gerado por IA.
- Complexidade de integração: incorporação perfeita da IA generativa aos fluxos de trabalho existentes e aos sistemas legados.
- Gerenciamento de custos: equilibrar os benefícios potenciais com os altos custos de implementação e operação.

Superar esses desafios requer um planejamento cuidadoso, investimento em infraestrutura e talento e uma abordagem estratégica para a implementação.

## Objetivo deste guia

A IA generativa está rapidamente se tornando um componente essencial em muitos setores. Ela oferece oportunidades transformadoras, mas também apresenta desafios em termos de integração, conformidade e escalabilidade. Muitas organizações lutam para aproveitar totalmente a IA devido a bases tecnológicas fracas, resistência a mudanças e problemas de qualidade de dados. A avaliação generativa da carga de trabalho de IA aborda esses desafios identificando os requisitos de modernização, definindo o escopo da implementação e desafiando sistemas e pensamentos legados. Também ajuda a determinar produtos mínimos viáveis (MVPs) e ajuda a desenvolver uma arquitetura de solução alvo, garantindo uma abordagem estruturada e estratégica para a adoção da IA.

Este guia serve como uma abordagem estruturada para ajudar as organizações a lidar com as complexidades da adoção de tecnologias generativas de IA. Em vez de definir claramente os requisitos desde o início, o guia ajuda a:

- Identificar possíveis casos de uso da IA generativa em sua organização.
- Avaliando a prontidão da sua organização para a adoção generativa da IA.
- Definir e refinar metas de casos de uso e metas ampliadas.
- Determinar o escopo e os requisitos para a implementação generativa da IA.
- Desenvolvendo uma arquitetura de solução alvo.

## Público-alvo e benefícios

Essa avaliação foi projetada especificamente para arquitetos de soluções, arquitetos corporativos e arquitetos de aplicativos que desejam avaliar os aspectos técnicos da modernização generativa da carga de trabalho de IA. Também é valioso para gerentes de programas e pessoas que desejam avaliar a prontidão geral, a alocação de recursos e os requisitos de capacitação de sua equipe. As melhores práticas do setor enfatizam a importância de uma avaliação abrangente para garantir a prontidão para a adoção da IA. Isso inclui avaliar a arquitetura, o armazenamento, a conformidade, a integração, os testes, a implantação e a automação.

## Escopo

Os tópicos a seguir estão no escopo do método generativo de avaliação da carga de trabalho da IA:

- Tecnologias e modelos atuais de IA generativa (por exemplo, modelos de linguagem grande, modelos de geração de imagens)
- Aplicativos restritos de IA que usam técnicas generativas
- Integração da IA generativa com sistemas e fluxos de trabalho existentes
- Estratégias de dados para treinar e ajustar modelos generativos de IA
- Considerações éticas e práticas responsáveis de IA para aplicativos atuais de IA generativa
- Estratégias de teste e implantação para IA generativa em ambientes de produção
- Considerações sobre segurança e privacidade para implementações generativas de IA
- Otimização de desempenho e escalabilidade de cargas de trabalho generativas de IA
- Casos de uso e aplicações da IA generativa em vários setores
- Avaliação de resultados generativos de IA e processos de garantia de qualidade

Os tópicos a seguir estão fora do escopo:

- Cenários de inteligência artificial geral (AGI) e superinteligência artificial (ASI)
- Futuros avanços especulativos em IA além dos modelos generativos atuais
- Aplicativos de computação quântica em IA
- Computação neuromórfica e interfaces cérebro-computador
- Consciência e autoconsciência em sistemas de IA
- Impactos sociais de longo prazo da IA avançada além das atuais aplicações generativas de IA
- Estruturas regulatórias para tecnologias hipotéticas de IA do futuro
- Debates filosóficos sobre a natureza da inteligência e da consciência nas máquinas
- Casos extremos ou casos de uso altamente especulativos de IA
- Especificações técnicas detalhadas de modelos ou arquiteturas de IA proprietários

# Resultados de negócios desejados

A avaliação generativa da carga de trabalho de IA visa fornecer vários resultados direcionados que são cruciais para modernizar com sucesso as cargas de trabalho generativas de IA. Esses resultados garantem que as organizações estejam bem preparadas para integrar as tecnologias de IA de forma eficaz e eficiente.

Para cada resultado desejado, a avaliação generativa da carga de trabalho da IA se concentra em:

- **Interdependências:** identifique e esclareça quaisquer interdependências entre o resultado e outros aspectos do processo de modernização. Isso inclui entender como um resultado pode influenciar ou ser influenciado por outros, para garantir uma abordagem holística da modernização.
- **Alinhamento das partes interessadas:** delineie estratégias para alinhar várias partes interessadas a cada resultado. Isso envolve comunicar o valor e o impacto de cada resultado a diferentes níveis e departamentos organizacionais, para promover a adesão e o suporte.
- **Priorização:** nos casos em que vários casos de uso ou resultados são identificados, forneça uma estrutura para priorizá-los com base em fatores como impacto nos negócios, requisitos de recursos e alinhamento estratégico.
- **Melhoria contínua:** Para cada resultado, estabeleça mecanismos para avaliação e refinamento contínuos. Isso garante que os esforços de modernização permaneçam adaptáveis e responsivos às mudanças nos cenários tecnológicos e nas necessidades comerciais.

Aqui está uma discussão detalhada de cada resultado desejado:

## Arquitetura de destino

- **Definição:** A avaliação ajuda a definir uma arquitetura-alvo clara e escalável para cargas de trabalho generativas de IA.
- **Componentes:** isso inclui selecionar serviços de nuvem apropriados, projetar pipelines de dados e garantir a interoperabilidade do sistema.
- **Benefícios:** uma arquitetura bem definida oferece suporte à escalabilidade, confiabilidade e otimização do desempenho, além de fornecer uma base sólida para a modernização.

## Prontidão para o cliente

- **Avaliação:** avalie o estado atual da infraestrutura, dos processos e da cultura da organização para determinar a prontidão para a adoção generativa da modernização da IA.
- **Crítérios:** Isso envolve avaliar as capacidades técnicas, a qualidade dos dados e a disposição organizacional de abraçar a mudança.
- **Resultado:** A identificação de lacunas e áreas de melhoria garante que a organização esteja preparada para uma transição suave para soluções e tecnologias modernas.

### Metas de caso de uso e metas ampliadas

- As metas de casos de uso estabelecem objetivos claros para a implementação da solução alvo, com foco em problemas ou oportunidades comerciais específicos.

Uma meta de caso de uso no contexto da modernização generativa da IA se refere a um objetivo específico e mensurável que uma organização pretende alcançar implementando soluções generativas de IA. Essas metas geralmente estão alinhadas com objetivos comerciais mais amplos e se concentram em abordar desafios ou oportunidades específicos dentro da organização.

Exemplos de metas de casos de uso podem incluir:

- Reduzindo o tempo de resposta do atendimento ao cliente em 50% usando chatbots generativos alimentados por IA.
- Melhorando a eficiência da revisão de código em 30% por meio da análise generativa de código assistida por IA.
- Aumentar a precisão da detecção de fraudes em 25% usando o reconhecimento generativo de padrões de IA.
- As metas de expansão definem metas ambiciosas que ultrapassam os limites do que a modernização generativa da IA pode alcançar dentro da organização.
- **Impacto:** definir metas alcançáveis e ambiciosas ajuda a alinhar as iniciativas generativas de modernização da IA com os objetivos estratégicos de negócios e incentiva a inovação.

### Estimativa de esforço

- **Objetivo:** A estimativa precisa do esforço ajuda no planejamento de recursos e garante que os projetos sejam entregues no prazo e dentro do orçamento.
- **Escopo:** estime os recursos, o tempo e o orçamento necessários para implementar o plano generativo de modernização da IA.
- **Fatores:** considere a complexidade técnica, os desafios de integração e os riscos potenciais.

## Necessidades de capacitação

- **Treinamento e desenvolvimento:** identifique as habilidades e os conhecimentos necessários para a adoção bem-sucedida da modernização generativa da IA.
- **Recursos:** Determine a necessidade de programas de treinamento, workshops e outras atividades de capacitação.
- **Resultado:** garantir que a equipe esteja equipada com as habilidades necessárias aumenta a eficácia das iniciativas generativas de modernização da IA e apoia o sucesso a longo prazo.

## Plano de implementação

- **Roteiro:** desenvolva um plano detalhado que descreva as etapas necessárias para alcançar a modernização generativa da IA.
- **Marcos:** defina os principais marcos e resultados para acompanhar o progresso.
- **Benefícios:** um plano de implementação claro fornece orientação e responsabilidade, além de facilitar uma abordagem estruturada para a modernização generativa da IA.

# Considerações e pré-requisitos de avaliação

## Comece com casos de uso claros

Identifique problemas ou oportunidades comerciais específicos que a IA generativa pode resolver. Concentre-se em casos de uso que se alinhem às metas estratégicas de negócios e ofereçam benefícios mensuráveis. Priorize casos de uso que tenham como alvo desafios comuns na organização para garantir que a arquitetura da solução possa servir como padrão para vários cenários.

Iniciar o processo de avaliação com uma compreensão geral das possíveis aplicações generativas de IA é benéfico, mas não obrigatório. O [questionário](#) incluído neste guia acomoda vários níveis de preparação, desde organizações que têm casos de uso bem definidos até aquelas que têm apenas ideias gerais. O processo de avaliação serve para:

- Refine e esclareça essas ideias de casos de uso iniciais.
- Identifique novos casos de uso em potencial.
- Desenvolva metas específicas e mensuráveis para cada caso de uso.
- Avalie a viabilidade e o impacto potencial de cada caso de uso.

Vamos considerar um exemplo hipotético: uma empresa de serviços financeiros decide explorar a modernização generativa da IA. Eles começam com uma ideia ampla de melhorar o atendimento ao cliente e os processos de detecção de fraudes.

- Avaliação inicial: o questionário os ajuda a avaliar seus sistemas atuais, a qualidade dos dados e a prontidão organizacional para a adoção generativa da IA.
- Refinamento do caso de uso: por meio do processo de avaliação, eles refinam suas ideias iniciais em dois casos de uso específicos:
  - Implementação de um chatbot generativo com inteligência artificial para consultas de clientes
  - Usando IA generativa para detecção de fraudes em transações em tempo real
- Definição de metas: para cada caso de uso, eles definem metas específicas:
  - Reduza o tempo de resposta do atendimento ao cliente em 40 por cento em 6 meses
  - Melhore a precisão da detecção de fraudes em 20% e reduza os falsos positivos em 15%
- Metas de expansão: eles também estabeleceram essas metas ambiciosas:

- Alcance 80% de satisfação do cliente com respostas assistidas por IA
- Desenvolva um modelo preditivo de detecção de fraudes que identifique novos padrões de fraude
- Definição de MVP: o questionário os ajuda a determinar um MVP para cada caso de uso, com foco em recursos essenciais que agregam valor imediato.
- Arquitetura de destino: por fim, eles desenvolvem uma arquitetura de destino que suporta um ou ambos os casos de uso e garante escalabilidade e integração com os sistemas existentes.

## Garanta o alinhamento dos negócios

Alinhe as iniciativas generativas de IA com a estratégia e os objetivos gerais de negócios. Para cada caso de uso, desenvolva uma proposta de valor clara que demonstre como a IA generativa contribui para o crescimento, a eficiência ou a inovação dos negócios. Estabeleça métricas para medir o impacto das implementações generativas de IA nos principais indicadores de desempenho (KPIs).

## Implemente governança e supervisão

Crie um comitê diretor multifuncional para supervisionar as iniciativas generativas de IA. Desenvolva políticas e diretrizes para o uso responsável da IA, abordando considerações éticas e possíveis preconceitos. Estabeleça um processo de revisão para projetos generativos de IA para garantir a conformidade com os padrões organizacionais e os requisitos regulatórios.

## Dados de endereço e pré-requisitos técnicos

Avalie e melhore a qualidade dos dados e implemente práticas de governança de dados para garantir entradas confiáveis para modelos generativos de IA. Desenvolva uma estratégia de dados que aborde a coleta, o armazenamento e o gerenciamento de dados específicos às necessidades generativas de IA. Avalie e aprimore a infraestrutura de dados para suportar o volume e a velocidade de dados necessários para cargas de trabalho generativas de IA.

## Considere os requisitos de recursos computacionais

Avalie a infraestrutura de TI atual e identifique lacunas na capacidade computacional para cargas de trabalho generativas de IA. Planeje recursos computacionais escaláveis, considerando opções como serviços em nuvem ou clusters de computação de alto desempenho locais. Otimize a alocação

de recursos para equilibrar desempenho e economia para cargas de trabalho de treinamento e inferência.

## Abordar as implicações de privacidade e segurança

Implemente medidas de segurança robustas para proteger dados confidenciais usados em treinamentos e operações de IA generativa. Garanta a conformidade com os regulamentos de proteção de dados, como o Regulamento Geral de Proteção de Dados (GDPR) ou a Lei de Privacidade do Consumidor da Califórnia (CCPA) ao lidar com informações pessoais. Desenvolva protocolos para implantação e monitoramento seguros de modelos para evitar o acesso não autorizado ou o uso indevido de recursos generativos de IA.

## Envolva as partes interessadas cedo

Envolva as principais partes interessadas desde o início para obter a adesão e o apoio da liderança. Comunique claramente os benefícios e o impacto potencial das iniciativas de modernização, especificamente para cargas de trabalho generativas de IA. Forneça treinamento e recursos para ajudar as partes interessadas a entender as tecnologias generativas de IA e suas implicações.

## Faça iterações e aprenda

Adote uma abordagem incremental que permita refinar as soluções-alvo. Use ciclos de feedback para melhorar continuamente a arquitetura e os processos da carga de trabalho. Avalie regularmente o desempenho e o impacto das implementações generativas de IA e ajuste as estratégias conforme necessário com base nos resultados do mundo real e nas necessidades comerciais em evolução.

# Questionário generativo de avaliação da carga de trabalho de IA

As seções a seguir fornecem perguntas que você pode usar para avaliar diferentes aspectos da modernização generativa da carga de trabalho de IA para sua organização. Esse questionário abrangente avalia a prontidão da sua organização para adotar e implementar cargas de trabalho generativas de IA com perguntas sobre as principais áreas, incluindo casos de uso, arquitetura, armazenamento, conformidade, integração, testes, implantação e estratégia de dados. Ao abordar aspectos críticos da implementação generativa da IA, da infraestrutura técnica às considerações regulatórias, esse questionário ajuda a identificar pontos fortes, lacunas e oportunidades em sua jornada de modernização da IA.

Seções:

- [Prontidão](#)
- [Casos de uso](#)
- [Arquitetura](#)
- [Armazenamento](#)
- [Regulamentos e conformidade](#)
- [Integração](#)
- [Teste](#)
- [Implantação e automação](#)
- [Estratégia de dados](#)

Você também pode baixar o questionário no formato Microsoft Excel e usá-lo para registrar suas informações.



[Baixe o questionário](#)

# Prontidão

| Pergunta                                                                                                          | Exemplo de resposta                                                                                                                                                                                                                                                 |
|-------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Você tem AWS contas que podem ser usadas para essas cargas de trabalho?                                           | Sim ou não.                                                                                                                                                                                                                                                         |
| Você tem um contrato corporativo existente com AWS?                                                               | Sim ou não.                                                                                                                                                                                                                                                         |
| Quão escalável é sua infraestrutura de nuvem atual para lidar com cargas de trabalho generativas de IA?           | Nossa infraestrutura de nuvem é altamente escalável, com recursos de escalonamento automático para recursos computacionais e sistemas de armazenamento distribuído, projetados para lidar com cargas de trabalho generativas de IA em grande escala com eficiência. |
| Você tem recursos de pipeline de dados para pré-processamento e engenharia de recursos em grande escala?          | Nossos pipelines de dados usam estruturas de processamento distribuído, como o Apache Spark, para pré-processamento de dados em grande escala e engenharia de recursos, com suporte para processamento de dados em lote e streaming.                                |
| Você tem capacidade de provisionamento e gerenciamento de contas?                                                 | Sim ou não.                                                                                                                                                                                                                                                         |
| Como você descreveria a alfabetização e a prontidão da sua organização para adotar tecnologias generativas de IA? | Nossa organização investiu pesadamente em programas educacionais de IA, e a maioria da equipe técnica concluiu o treinamento básico de IA/ML. A organização tem uma cultura de inovação que abraça novas tecnologias, incluindo IA generativa.                      |
| Que experiência em IA/ML existe em sua organização e como ela é distribuída?                                      | Temos um Centro de Excelência em IA dedicado com cientistas de dados e engenheiros de ML experientes. Capacitamos especiali                                                                                                                                         |

| Pergunta                                                                                                               | Exemplo de resposta                                                                                                                                                                                                                                                                                                              |
|------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                                                        | stas do domínio em diferentes unidades de negócios para nos tornarmos alfabetizados em IA e identificarmos casos de uso generativos de IA.                                                                                                                                                                                       |
| Você tem um caso de negócios de alto nível que articula os objetivos, benefícios e custos do programa de nuvem?        | Sim ou não.                                                                                                                                                                                                                                                                                                                      |
| Qual é o seu cronograma para levar a solução à produção?                                                               | Semanas, meses e assim por diante.                                                                                                                                                                                                                                                                                               |
| Um compromisso de financiamento foi assumido por suas principais partes interessadas (por exemplo, CFO, CIT/CTO, COO)? | Sim ou não.                                                                                                                                                                                                                                                                                                                      |
| Como você garante a conformidade com os regulamentos de proteção de dados em suas iniciativas generativas de IA?       | Temos uma equipe de conformidade dedicada que trabalha em estreita colaboração com nossas equipes de IA. Realizamos avaliações regulares de impacto na privacidade, implementamos princípios de proteção de dados por design e mantemos registros detalhados de processamento de dados para todos os projetos generativos de IA. |
| Quão maduros estão seus sistemas existentes que se integram às novas tecnologias generativas de IA?                    | Nossa arquitetura de TI é baseada em microsserviços e permite APIs a integração flexível de novas tecnologias generativas de IA. Esses sistemas são padronizados em formatos e protocolos de dados comuns para garantir a interoperabilidade.                                                                                    |

| Pergunta                                                                                                                   | Exemplo de resposta                                                                                                                                                                                                                                                           |
|----------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Que experiência você tem na operacionalização de modelos de ML e como isso pode se aplicar aos sistemas generativos de IA? | Estabelecemos MLOps práticas, incluindo pipelines automatizados de implantação de modelos, sistemas de monitoramento e estruturas de testes A/B. Essas práticas estão sendo adaptadas para lidar com os requisitos exclusivos dos modelos generativos de IA em grande escala. |

## Casos de uso

| Pergunta                                                                                                      | Exemplo de resposta                                                                                                                                                                                                                                                                                                                                                 |
|---------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Qual é a meta principal ou o critério de sucesso do caso de uso?                                              | Para melhorar o tempo de resposta do suporte ao cliente, aumentar as conversões de vendas e aprimorar as recomendações de produtos. Além disso: para melhorar a satisfação do usuário, a taxa de conclusão da tarefa, a qualidade da resposta e assim por diante.                                                                                                   |
| Como esse caso de uso se alinha às metas estratégicas da sua organização?                                     | Isso se alinha ao nosso objetivo estratégico de aumentar a satisfação do cliente reduzindo os tempos de resposta no atendimento ao cliente.                                                                                                                                                                                                                         |
| Qual é o volume esperado de dados ou solicitações para o caso de uso?                                         | 500 transações por segundo (TPS).                                                                                                                                                                                                                                                                                                                                   |
| Quais tipos de fontes de dados são necessários para dar suporte às suas cargas de trabalho generativas de IA? | Bancos de dados estruturados internos (registros de clientes, dados de vendas etc.); dados de texto não estruturados de documentos, e-mails e mídias sociais; arquivos de áudio e vídeo para tarefas de reconhecimento de voz e imagem; dados de streaming em tempo real de dispositivos e sensores de IoT; conjuntos APIs de dados públicos e para enriquecimento. |

| Pergunta                                                                                                    | Exemplo de resposta                                                                                                                                                                                                                                                                                                                                                                                          |
|-------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Com que frequência você precisa atualizar ou atualizar os dados dessas fontes?                              | Bancos de dados transacionais: atualizações quase em tempo real; repositórios de documentos: atualizações diárias em lote; feeds de mídia social: atualizações de hora em hora; dados do sensor de IoT: streaming contínuo em tempo real; conjuntos de dados públicos: atualizações mensais ou trimestrais.                                                                                                  |
| Quais formatos de dados seus modelos generativos de IA exigem como entrada?                                 | Dados estruturados: tabelas de banco de dados CSV, JSON e SQL; dados de texto: texto simples, PDF e HTML; dados de imagem: JPEG, PNG e TIFF; dados de áudio: WAV e; dados de vídeo: e MP3 AVI. MP4                                                                                                                                                                                                           |
| Quais são suas principais preocupações com a qualidade de dados para cargas de trabalho generativas de IA?  | Completude: garantia de que nenhum campo crítico esteja ausente; precisão: verificação da exatidão dos dados e eliminação de erros; consistência: manutenção de formatos e valores uniformes em todas as fontes; pontualidade: garantia de que os dados estejam atualizados para inferência em tempo real; relevância: confirmação de que os dados estão alinhados com a tarefa específica de IA generativa. |
| Quais são os principais requisitos de desempenho (por exemplo, tempo de resposta, produtividade, precisão)? | 95% de precisão; tempo de resposta de < 500 ms; capacidade de lidar com 1000 solicitações/segundo. Alta precisão (95% +), precisão moderada (80-90%), melhor esforço e assim por diante.                                                                                                                                                                                                                     |
| Você tem algum outro KPIs para medir o sucesso desse caso de uso?                                           | KPIs Os principais incluem redução da taxa de erro, economia de tempo por transação e pontuações de satisfação do cliente.                                                                                                                                                                                                                                                                                   |

| Pergunta                                                                                        | Exemplo de resposta                                                                                                                         |
|-------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| Quanta precisão do modelo é desejada e como ela se equilibra com o custo?                       | Alta precisão (> 90%) com custo moderado, precisão moderada (70-80%) com baixo custo e assim por diante.                                    |
| Quais são os principais casos de uso ou cenários da solução generativa de IA?                   | Chatbot de atendimento ao cliente, geração de conteúdo, recomendação de produtos e assim por diante.                                        |
| Quais são os usuários-alvo ou personas do sistema generativo de IA?                             | Agentes de atendimento ao cliente, equipe de marketing, funcionários, usuários finais e assim por diante.                                   |
| Qual é o volume esperado de solicitações ou usuários?                                           | 1.000 solicitações por dia; 10.000 usuários ativos mensais.                                                                                 |
| Há alguma restrição ou requisito de caso de uso específico?                                     | Resposta em tempo real, suporte multilíngue, privacidade de dados e assim por diante.                                                       |
| Você tem um orçamento alocado para desenvolver e manter a solução generativa de IA?             | O custo inicial de desenvolvimento é estimado em \$200.000, com custos anuais de manutenção de \$50.000.                                    |
| Qual é o retorno do investimento (ROI) projetado e o período de retorno para esse caso de uso?  | ROI esperado de 150% em três anos, com um período de retorno de 18 meses.                                                                   |
| Há algum custo oculto ou economia potencial que deva ser considerada?                           | As economias potenciais incluem a redução dos custos de horas extras. Os custos ocultos podem envolver treinamento adicional para a equipe. |
| Quais são as possibilidades de escalabilidade e futura expansão dessa solução generativa de IA? | A solução foi projetada para se expandir com nossas operações, com a possibilidade de expansão para outros departamentos no futuro.         |

| Pergunta                                                                                                                                                           | Exemplo de resposta                                                                                                                                                                                                                                                                                                                                          |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Como você garante a imparcialidade e mitiga o preconceito em seus modelos generativos de IA?                                                                       | Planejamos mitigar o viés por meio de coleta diversificada de dados, auditorias regulares de preconceito e implementação de técnicas de mitigação de preconceitos.                                                                                                                                                                                           |
| Quais processos você tem em vigor para lidar com questões éticas ou consequências não intencionais?                                                                | Gerenciaremos as preocupações éticas por meio de um plano estabelecido de resposta a incidentes de IA, avaliações regulares de risco ético, um sistema de denúncias anônimas para funcionários, colaboração com especialistas externos em ética e monitoramento e ajuste contínuos dos modelos implantados com base no feedback.                             |
| Como você aborda a priorização e o sequenciamento das avaliações generativas da carga de trabalho de IA em diferentes projetos e departamentos da sua organização? | Realizando uma pesquisa de alto nível em todos os departamentos para identificar possíveis casos de uso generativo de IA e avaliá-los com base em três critérios principais: impacto nos negócios, viabilidade técnica e considerações éticas. Projetos com alto impacto potencial, menores barreiras técnicas e preocupações éticas mínimas têm prioridade. |

## Arquitetura

| Pergunta                                                                   | Exemplo de resposta                                                                                                  |
|----------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|
| Que tipo de modelo ou arquitetura generativa de IA está sendo considerado? | Transformador, rede neural convolucional (CNN), rede neural recorrente (RNN), árvores de decisão e assim por diante. |
| Qual é a escala ou o volume esperado de dados e cálculos?                  | Milhões de usuários, petabytes de dados e assim por diante.                                                          |

| Pergunta                                                                                                        | Exemplo de resposta                                                                              |
|-----------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|
| Quais são os requisitos de hardware (por exemplo, CPUs ou GPUs) para treinamento e inferência?                  | High-end GPUs, clusters de CPU, instâncias de nuvem e assim por diante.                          |
| Como o modelo generativo de IA será atualizado ou retreinado ao longo do tempo?                                 | Por meio de aprendizado contínuo, reciclagem periódica, atualizações manuais e assim por diante. |
| Quais são os requisitos de pré-processamento de dados e engenharia de recursos?                                 | Limpeza de texto, aumento de imagem, seleção de recursos e assim por diante.                     |
| Como o sistema generativo de IA lidará com casos extremos, valores discrepantes ou entradas de baixa confiança? | Por meio do recurso à supervisão humana, solicite esclarecimentos e assim por diante.            |
| Quais são os requisitos de latência para o aplicativo generativo de IA?                                         | Em tempo real, quase em tempo real, processamento em lote e assim por diante.                    |

## Armazenamento

| Pergunta                                                                                                                                                 | Exemplo de resposta                                                                                                                                                            |
|----------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Onde os dados de treinamento serão armazenados?                                                                                                          | No armazenamento em nuvem (por exemplo, Amazon S3, armazenamento de arquivos, armazenamento em blocos ou armazenamento de objetos), no armazenamento local e assim por diante. |
| Quais são os requisitos de armazenamento para os dados de treinamento e os artefatos do modelo (por exemplo, capacidade, durabilidade, disponibilidade)? | Armazenamento em escala de petabytes, alta durabilidade (99,999999999% de durabilidade), alta disponibilidade e assim por diante.                                              |

| Pergunta                                                                                                                                                                                    | Exemplo de resposta                                                                                                                                                                                                                                           |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Quais são os requisitos de retenção e backup de dados para os dados de treinamento e artefatos do modelo?                                                                                   | Retenção de dados por x anos, backups diários, backups externos e assim por diante.                                                                                                                                                                           |
| Quais formatos de arquivo são usados principalmente para armazenar seus conjuntos de dados de treinamento de IA (por exemplo, CSV, JSON, Parquet)? HDF5                                     | Arquivos em parquet para dados estruturados e HDF5 para grandes matrizes multidimensionais e dados não estruturados, como imagens e texto. Usamos formatos especializados, como TFRecord para otimizar o carregamento de dados durante o treinamento.         |
| Como seus conjuntos de dados de treinamento são organizados: como arquivos individuais, em bancos de dados ou usando formatos de dados de IA especializados?                                | Conjuntos de dados pequenos e médios são armazenados como arquivos Parquet individuais no armazenamento de objetos para maior flexibilidade. Grandes conjuntos de dados são armazenados em um banco de dados distribuído (Cassandra) para lidar com a escala. |
| Você usa alguma técnica de compressão ou codificação de dados especificamente para dados generativos de treinamento de IA?                                                                  | Para dados tabulares, usamos técnicas de codificação de dicionário e empacotamento de bits disponíveis no Parquet. Para imagens, usamos compressão JPEG com perdas com configurações de qualidade otimizadas para nossos modelos.                             |
| Como você lida com o controle de versão e o armazenamento de diferentes iterações de conjuntos de dados de treinamento? Que impacto isso tem nas suas necessidades gerais de armazenamento? | Usamos um sistema de controle de versão de dados (DVC) integrado à nossa plataforma de ML.                                                                                                                                                                    |

## Regulamentos e conformidade

| Pergunta                                                                                                                                | Exemplo de resposta                                                                                                                                           |
|-----------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Quais são os regulamentos ou requisitos de conformidade relevantes para a solução generativa de IA (por exemplo, GDPR, HIPAA, PCI-DSS)? | GDPR para lidar com dados pessoais, HIPAA para dados de saúde, PCI-DSS para dados de pagamento e assim por diante.                                            |
| Quais diretrizes ou estruturas éticas de IA generativa sua organização adotou?                                                          | Implementamos nossas próprias diretrizes de IA responsável. Todos os projetos de IA generativa passam por uma revisão ética antes da aprovação e implantação. |
| Quais são os requisitos de segurança para o sistema generativo de IA?                                                                   | Criptografia de dados, comunicação de rede segura, auditorias regulares de segurança.                                                                         |
| Quais são os requisitos para privacidade e proteção de dados?                                                                           | Anonimização de dados, criptografia, controle de acesso e assim por diante.                                                                                   |
| Quais são os requisitos da solução para lidar com dados sensíveis ou confidenciais?                                                     | Controles de acesso rígidos, mascaramento de dados, requisitos de residência de dados e assim por diante.                                                     |
| Como a autenticação e a autorização do usuário serão tratadas?                                                                          | Usando chaves de API OAuth, login único (SSO) e controle de acesso baseado em função (RBAC).                                                                  |
| Como a solução será monitorada e gerenciada na produção?                                                                                | Usando ferramentas de monitoramento como Prometheus e Datadog, ferramentas de registro como ELK Stack, sistemas de alerta e assim por diante.                 |

# Integração

| Pergunta                                                                                                                    | Exemplo de resposta                                                                                                                                            |
|-----------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Quais são os requisitos para integrar a solução de IA generativa com sistemas ou fontes de dados existentes?                | REST APIs, filas de mensagens, conectores de banco de dados e assim por diante.                                                                                |
| Como os dados serão ingeridos e pré-processados para a solução generativa de IA?                                            | Usando processamento em lote, dados de streaming, transformações de dados e engenharia de recursos.                                                            |
| Como a saída da solução generativa de IA será consumida ou integrada aos sistemas downstream?                               | Por meio de endpoints de API, filas de mensagens, atualizações de banco de dados e assim por diante.                                                           |
| Quais padrões de integração orientados por eventos podem ser usados para a solução generativa de IA?                        | Filas de mensagens (como Amazon SQS, Apache Kafka, RabbitMQ), sistemas pub/sub, webhooks, plataformas de streaming de eventos.                                 |
| Quais abordagens de integração baseadas em API podem ser usadas para conectar a solução generativa de IA a outros sistemas? | RESTful APIs, GraphQL APIs, SOAP APIs (para sistemas legados).                                                                                                 |
| Quais componentes da arquitetura de microsserviços podem ser usados para a integração generativa da solução de IA?          | Malha de serviços para comunicação entre serviços, gateways de API, orquestração de contêineres (por exemplo, Kubernetes).                                     |
| Como a integração híbrida pode ser implementada para a solução generativa de IA?                                            | Combinando padrões orientados por eventos para atualizações em tempo real, processamento em lote para dados históricos e integração APIs de sistemas externos. |
| Como a saída da solução de IA generativa pode ser integrada aos sistemas downstream?                                        | Por meio de endpoints de API, filas de mensagens, atualizações de banco de dados, webhooks e exportações de arquivos.                                          |

| Pergunta                                                                                                                                                      | Exemplo de resposta                                                                                                                                                                                                                                                                         |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Quais medidas de segurança devem ser consideradas para integrar a solução generativa de IA?                                                                   | Mecanismos de autenticação (como OAuth ou JWT), criptografia (em trânsito e em repouso), limitação de taxa de API e listas de controle de acesso (ACLs).                                                                                                                                    |
| Como você planeja integrar estruturas de código aberto, como LlamaIndex ou LangChain em seu pipeline de dados existente e fluxo de trabalho generativo de IA? | Planejamos usá-lo LangChain para criar aplicativos complexos de IA generativa, especialmente para seus recursos de gerenciamento de agentes e memória. Nosso objetivo é que 60% de nossos projetos de IA generativa sejam usados LangChain nos próximos 6 meses.                            |
| Como você garantirá a compatibilidade entre as estruturas de código aberto escolhidas e sua infraestrutura de dados existente?                                | Estamos criando uma equipe de integração dedicada para garantir uma compatibilidade perfeita. Até o terceiro trimestre, nossa meta é ter um pipeline totalmente integrado que seja usado LlamaIndex para indexação e recuperação eficientes de dados em nossa estrutura atual de data lake. |
| Como você planeja aproveitar os componentes modulares das estruturas, como LangChain para prototipagem e experimentação rápidas?                              | Estamos configurando um ambiente sandbox em que os desenvolvedores podem criar protótipos rapidamente usando os componentes LangChain da.                                                                                                                                                   |
| Qual é a sua estratégia para acompanhar as atualizações e os novos recursos nessas estruturas de código aberto em rápida evolução?                            | Designamos uma equipe para monitorar GitHub repositórios e fóruns comunitários para LangChain e LlamaIndex. Planejamos avaliar e integrar as principais atualizações trimestralmente, com foco em melhorias de desempenho e novos recursos.                                                 |

# Teste

| Pergunta                                                                                                           | Exemplo de resposta                                                                                                                                                                                                                                                                              |
|--------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Quais são os requisitos de teste (por exemplo, testes unitários, testes de integração, end-to-end testes)?         | Teste unitário para componentes individuais, testes de integração com sistemas externos, end-to-end testes para cenários críticos e assim por diante.                                                                                                                                            |
| Como você garante a qualidade e a consistência dos dados em diferentes fontes para o treinamento generativo de IA? | Mantemos a qualidade dos dados por meio de ferramentas automatizadas de criação de perfil de dados, auditorias regulares de dados e um catálogo de dados centralizado. Implementamos políticas de governança de dados para garantir a consistência entre as fontes e manter a linhagem de dados. |
| Como o modelo generativo de IA será avaliado e validado?                                                           | Usando um conjunto de dados resistente, avaliação humana, testes A/B e assim por diante.                                                                                                                                                                                                         |
| Quais são os critérios para avaliar o desempenho e a precisão do modelo generativo de IA?                          | Precisão, recordação, pontuação F1, perplexidade, avaliação humana e assim por diante.                                                                                                                                                                                                           |
| Como os casos extremos e os casos secundários serão identificados e tratados?                                      | Usando um conjunto de testes abrangente, avaliação humana, testes adversários e assim por diante.                                                                                                                                                                                                |
| Como você testará possíveis preconceitos no modelo generativo de IA?                                               | Usando análise de paridade demográfica, testes de igualdade de oportunidades, técnicas de redução de preconceitos adversários, testes contrafactuais e assim por diante.                                                                                                                         |
| Quais métricas serão usadas para medir a imparcialidade nos resultados do modelo?                                  | Taxa de impacto diferente, probabilidades equalizadas, paridade demográfica, métricas de justiça individual e assim por diante.                                                                                                                                                                  |

| Pergunta                                                                                                       | Exemplo de resposta                                                                                                                                                                             |
|----------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Como você garantirá uma representação diversificada em seus conjuntos de dados de teste para detecção de viés? | Usando amostragem estratificada em grupos demográficos, colaboração com especialistas em diversidade, uso de dados sintéticos para preencher lacunas e assim por diante.                        |
| Qual processo será implementado para o monitoramento contínuo da imparcialidade do modelo após a implantação?  | Auditorias regulares de imparcialidade, sistemas automatizados de detecção de viés, análise de feedback do usuário, reciclagem periódica com conjuntos de dados atualizados e assim por diante. |
| Como você abordará os preconceitos interseccionais no modelo generativo de IA?                                 | Usando análise de imparcialidade interseccional, testes de subgrupos, colaboração com especialistas em interseccionalidade e assim por diante.                                                  |
| Como você testará o desempenho do modelo em diferentes idiomas e contextos culturais?                          | Usando conjuntos de testes multilíngues, colaboração com especialistas culturais, métricas de imparcialidade localizadas, estudos de comparação intercultural e assim por diante.               |

## Implantação e automação

| Pergunta                                                              | Exemplo de resposta                                                                                                                                                                                                                                                                                                |
|-----------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Quais são os requisitos para escalabilidade e balanceamento de carga? | Roteamento inteligente de solicitações; sistema de escalonamento automático; otimização para arranques a frio rápidos empregando técnicas como cache de modelos, carregamento lento e sistemas de armazenamento distribuído; projetando o sistema para lidar com padrões de tráfego intermitentes e imprevisíveis. |

| Pergunta                                                                                                            | Exemplo de resposta                                                                                                                                              |
|---------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Quais são os requisitos para atualizar e lançar novas versões?                                                      | Implantações azul/verdes, lançamentos canários, atualizações contínuas e assim por diante.                                                                       |
| Quais são os requisitos para recuperação de desastres e continuidade de negócios?                                   | Procedimentos de backup e restauração, mecanismos de failover, configurações de alta disponibilidade e assim por diante.                                         |
| Quais são os requisitos para automatizar o treinamento, a implantação e o gerenciamento do modelo generativo de IA? | Pipeline de treinamento automatizado, implantação contínua, escalabilidade automática e assim por diante.                                                        |
| Como o modelo generativo de IA será atualizado e retreinado à medida que novos dados forem disponibilizados?        | Por meio de reciclagem periódica, aprendizado incremental, aprendizado por transferência e assim por diante.                                                     |
| Quais são os requisitos para automatizar o monitoramento e o gerenciamento?                                         | Alertas automatizados, escalabilidade automática, autorrecuperação e assim por diante.                                                                           |
| Qual é o seu ambiente de implantação preferido para cargas de trabalho generativas de IA?                           | Uma abordagem híbrida que usa a AWS para treinamento de modelos e nossa infraestrutura local para inferência para atender aos requisitos de residência de dados. |
| Você prefere alguma plataforma de nuvem específica para implantações generativas de IA?                             | Serviços da AWS, especialmente o Amazon SageMaker AI para desenvolvimento e implantação de modelos, e o Amazon Bedrock para modelos básicos.                     |
| Quais tecnologias de containerização você está considerando para cargas de trabalho generativas de IA?              | Queremos padronizar os contêineres Docker orquestrados com o Kubernetes para garantir portabilidade e escalabilidade em nosso ambiente híbrido.                  |

| Pergunta                                                                                                         | Exemplo de resposta                                                                                                                                                                                         |
|------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Você tem alguma ferramenta preferida para CI/CD em seu pipeline de IA generativa?                                | GitLab para controle de versão e pipelines de CI/CD, integrados ao Jenkins para testes e implantação automatizados.                                                                                         |
| Quais ferramentas de orquestração você está considerando para gerenciar fluxos de trabalho generativos de IA?    | Apache Airflow para orquestração de fluxo de trabalho, especialmente para pré-processamento de dados e pipelines de treinamento de modelos.                                                                 |
| Você tem algum requisito específico de infraestrutura local para suportar cargas de trabalho generativas de IA?  | Estamos investindo em servidores acelerados por GPU e redes de alta velocidade para suportar cargas de trabalho de inferência locais.                                                                       |
| Como você planeja gerenciar o controle de versão e a implantação de modelos em diferentes ambientes?             | Planejamos usá-lo MLflow para rastreamento e controle de versão de modelos e integrá-lo à nossa infraestrutura Kubernetes para uma implantação perfeita em todos os ambientes.                              |
| Quais ferramentas de monitoramento e observabilidade você está considerando para implantações generativas de IA? | Prometheus para coleta de métricas e Grafana para visualização, com soluções adicionais de registro personalizadas para monitoramento específico do modelo.                                                 |
| Como você está lidando com a movimentação e sincronização de dados em um modelo de implantação híbrida?          | Usaremos AWS DataSync para uma transferência eficiente de dados entre o armazenamento local e AWS, com trabalhos de sincronização automatizados que são agendados com base em nossos ciclos de treinamento. |
| Quais medidas de segurança você está implementando para implantações generativas de IA em diferentes ambientes?  | Usaremos o IAM para recursos de nuvem, integrados ao nosso Active Directory local para implementar end-to-end criptografia e segmentação de rede para proteger os fluxos de dados.                          |

## Estratégia de dados

| Pergunta                                                                                                                                         | Exemplo de resposta                                                                                                                                                                                                                                                  |
|--------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Quais tipos de dados específicos são cruciais para suas cargas de trabalho generativas de IA e qual porcentagem delas está acessível atualmente? | Os registros de chamadas de clientes e os dados de avaliações de produtos são cruciais. Atualmente, 85% desses tipos de dados estão acessíveis para nossos projetos de IA generativa.                                                                                |
| Como você garante e mede a qualidade dos seus dados?                                                                                             | Implementamos métricas de qualidade de dados, incluindo integridade, precisão, consistência e pontualidade. Usamos ferramentas automatizadas para avaliar regularmente essas métricas e temos uma equipe dedicada para limpeza e enriquecimento de dados.            |
| Qual porcentagem de seus dados atende aos seus padrões de qualidade para uso generativo de IA?                                                   | Atualmente, 78% dos nossos dados atendem aos nossos padrões de qualidade. Nossa meta é atingir 95% nos próximos 12 meses por meio de processos aprimorados de limpeza de dados.                                                                                      |
| Como você planeja criar confiança sobre o uso de dados em IA generativa entre suas partes interessadas?                                          | Estamos implementando um conselho de ética de IA, fornecendo explicações claras sobre as decisões de IA e conduzindo auditorias trimestrais de IA para garantir transparência e justiça.                                                                             |
| Quão abrangente é sua documentação sobre fontes de dados e linhagem?                                                                             | Mantemos um catálogo de dados detalhado que inclui metadados para todas as nossas fontes de dados, incluindo origem, frequência de atualização e uso. Usamos ferramentas de linhagem de dados para rastrear como os dados fluem e se transformam em nossos sistemas. |

| Pergunta                                                                                                              | Exemplo de resposta                                                                                                                                                                                                                                                                |
|-----------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Como você garante a diversidade em seus conjuntos de dados para evitar preconceitos nos modelos de IA?                | Nós obtemos ativamente dados de diversos grupos demográficos e auditamos regularmente nossos conjuntos de dados em busca de viés representacional. Também usamos técnicas de geração de dados sintéticos para equilibrar categorias sub-representadas.                             |
| Qual é sua taxa de atualização de dados para modelos críticos de IA generativa e como você determina essa frequência? | Os modelos críticos são atualizados semanalmente. Essa frequência é determinada pelas métricas de desempenho dos testes A/B, e nosso objetivo é uma degradação não superior a 2% entre as atualizações.                                                                            |
| Quantas versões de conjuntos de dados essenciais você mantém e por quanto tempo?                                      | Mantemos as últimas cinco versões de cada conjunto de dados crítico, com um período de retenção de 18 meses para cada versão.                                                                                                                                                      |
| Quantas equipes multifuncionais estão envolvidas em suas iniciativas de IA generativa e têm acesso aos seus dados?    | Temos três equipes multifuncionais. Cada equipe inclui cientistas de dados, especialistas de domínio, especialistas em ética e analistas de negócios.                                                                                                                              |
| Quais políticas e práticas de governança de dados você tem em vigor?                                                  | Temos um comitê multifuncional de governança de dados que supervisiona nossas políticas de dados. Implementamos controles de acesso baseados em funções, esquemas de classificação de dados e auditorias regulares para garantir a conformidade com nossa estrutura de governança. |

| Pergunta                                                                                                                                             | Exemplo de resposta                                                                                                                                                                                                                                           |
|------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Quais medidas você tem em vigor para garantir a privacidade dos dados, obter o consentimento adequado e manter a confidencialidade?                  | Implementamos uma estrutura abrangente e de privacidade de dados alinhada com o GDPR e a CCPA. Isso inclui obter consentimento explícito para o uso de dados, implementar técnicas de anonimização de dados e avaliações regulares do impacto na privacidade. |
| Qual porcentagem de seus conjuntos de dados de treinamento de IA foi auditada quanto a preconceitos no último trimestre?                             | 70% de nossos conjuntos de dados de treinamento de IA foram auditados quanto a preconceitos no último trimestre. Estamos implementando ferramentas automatizadas de detecção de viés para alcançar 100% de auditorias trimestrais.                            |
| Qual é a sua capacidade atual de processamento de dados e quanto você projeta que seja necessário para futuras cargas de trabalho de IA generativas? | Nossa capacidade atual é de 10 TB/day. We project needing 30 TB/day em um ano e estamos ampliando nossa infraestrutura para atender a essa demanda.                                                                                                           |
| Qual é sua estratégia para equilibrar a privacidade dos dados com as necessidades de dados dos modelos generativos de IA?                            | Estamos implementando técnicas avançadas de anonimização e geração de dados sintéticos. Nossa meta é aumentar nossos dados utilizáveis para IA em 40% e reduzir os riscos de privacidade em 60% no próximo ano.                                               |
| Qual porcentagem de seus conjuntos de dados de aprendizado de máquina (ML) está rotulada com precisão e qual é sua meta de taxa de precisão?         | Atualmente, 85% dos nossos conjuntos de dados de ML são rotulados com precisão. Nossa meta é uma taxa de precisão de 95% no próximo trimestre, empregando técnicas de etiquetagem humana e automatizada.                                                      |

# Traduzindo insights de avaliação em resultados acionáveis

Esta seção fornece uma estrutura para analisar as respostas do questionário e usar esses insights para moldar a arquitetura-alvo e outros resultados importantes da iniciativa generativa de modernização da IA. Essa estrutura preenche a lacuna entre a coleta e a implementação de dados e garante que a avaliação informe e conduza diretamente sua estratégia de modernização.

## Definição da arquitetura de destino:

- Use as respostas do questionário para informar a seleção de serviços em nuvem e o design de pipelines de dados.
- Certifique-se de que o design da arquitetura ofereça suporte à escalabilidade e à interoperabilidade, conforme destacado no guia.

## Avaliação da prontidão do cliente:

- Analise as respostas do questionário relacionadas à infraestrutura, aos processos e à cultura organizacional atuais.
- Identifique lacunas e crie um plano para resolvê-las. Priorize as lacunas que são essenciais para o sucesso do MVP.

## Caso de uso e metas de expansão:

- Extraia problemas comerciais específicos das respostas do questionário para definir metas claras de casos de uso.
- Estabeleça metas ambiciosas que se alinhem à visão de longo prazo da sua organização para a modernização generativa da IA.

## Estimativa do esforço:

- Use os dados do questionário para estimar recursos, tempo e orçamento para o MVP e para a implementação completa.
- Crie uma abordagem em fases que comece com o MVP e descreva as fases subsequentes.

## Necessidades de capacitação:

- Com base nas respostas do questionário, identifique as lacunas de habilidades e as necessidades de treinamento.
- Desenvolva um plano de treinamento que ofereça suporte às necessidades imediatas de MVP e à adoção generativa de IA a longo prazo.

#### Plano de implementação:

- Crie um roteiro abrangente que comece com o MVP e descreva as etapas para a modernização total da IA generativa.
- Defina marcos e resultados claros para cada fase da implementação.

#### Etapas práticas:

- Matriz de priorização: crie uma matriz que mapeie as respostas do questionário aos [seis resultados](#) para ajudar a priorizar recursos e esforços.
- Abordagem iterativa: projete o MVP para ser a primeira iteração em uma série de lançamentos planejados, em que cada versão se baseia na arquitetura de destino completa.
- Alinhamento das partes interessadas: use os resultados do questionário para alinhar as partes interessadas no escopo do MVP e na abordagem em fases para alcançar todos os resultados.
- Ciclo de feedback contínuo: implemente mecanismos para coletar feedback após a implantação do MVP e use insights para refinar os planos para as fases subsequentes.
- Implementação ágil: adote uma metodologia ágil que permita flexibilidade na abordagem de todos os resultados ao longo do tempo, começando com os resultados mais críticos no MVP.

# Próximas etapas

Depois de concluir a avaliação generativa da carga de trabalho da IA, siga estas etapas:

## 1. Forneça uma arquitetura de destino detalhada

- **Objetivo:** O arquiteto da solução cria uma arquitetura de destino abrangente que se alinha às metas da organização e aos resultados da avaliação.
- **Componentes:** essa arquitetura inclui o design de ingestão de dados, pontos de integração e interoperabilidade do sistema para garantir escalabilidade, confiabilidade e otimização do desempenho.

## 2. Explicar o quão específico é Serviços da AWS adequado ao caso de uso

- **Mapeamento de serviços:** identifique e mapeie os específicos Serviços da AWS que melhor se adequam aos casos de uso identificados.
- **Benefícios:** destaque como esses serviços atendem às necessidades comerciais específicas, aumentam a eficiência e oferecem escalabilidade.

## 3. Forneça soluções alternativas opcionais com prós e contras

- **Alternativas:** apresente soluções alternativas que também possam atender aos requisitos da organização.
- **Análise:** ofereça uma análise detalhada das vantagens e desvantagens de cada alternativa, considerando fatores como custo, complexidade e alinhamento com as metas de negócios.

## 4. Forneça uma estimativa detalhada do preço de Serviços da AWS

- **Análise de custos:** forneça uma estimativa de custo detalhada para a proposta Serviços da AWS, incluindo possíveis cenários de uso e modelos de preços.
- **Alinhamento orçamentário:** certifique-se de que o custo esteja alinhado às restrições orçamentárias da organização e forneça uma compreensão clara das implicações financeiras.

## 5. Obtenha feedback sobre a arquitetura proposta

- **Engajamento das partes interessadas:** envolva-se com as partes interessadas para apresentar a arquitetura proposta e obter feedback.
- **Melhoria iterativa:** use o feedback para refinar e melhorar a solução e confirmar se ela atende às necessidades e expectativas de todas as partes interessadas.

## Perguntas frequentes

### Qual é o objetivo principal da avaliação generativa da carga de trabalho da IA?

O objetivo principal da avaliação é avaliar a prontidão de uma organização para modernizar suas cargas de trabalho generativas de IA, identificar casos de uso e desenvolver uma arquitetura de solução alvo. O objetivo é definir os requisitos de modernização, determinar o escopo da implementação e se preparar para uma modernização generativa bem-sucedida da IA.

### Quem deve usar essa avaliação?

Essa avaliação é para arquitetos de soluções, arquitetos corporativos e arquitetos de aplicativos que desejam avaliar os aspectos técnicos da modernização generativa da IA. Também é útil para gerentes de programas e gerentes de pessoal avaliarem as necessidades gerais de prontidão, alocação de recursos e capacitação.

### Quais são os principais componentes avaliados na avaliação?

A avaliação abrange a prontidão geral, o caso de uso, a arquitetura, o armazenamento, os regulamentos e a conformidade, a integração, os testes, a automação da implantação e a estratégia de dados. Esses componentes são cruciais para determinar a prontidão técnica e organizacional para a adoção generativa da modernização da IA.

### Como a avaliação ajuda a definir a arquitetura de destino?

A avaliação fornece uma abordagem estruturada para avaliar os sistemas atuais e identificar melhorias. Ele ajuda você a selecionar tecnologias apropriadas e projetar arquiteturas escaláveis que se alinham às metas de negócios e aos requisitos de casos de uso.

### Quais são os benefícios de realizar uma avaliação generativa da carga de trabalho de IA?

Os benefícios incluem maior eficiência, melhor tomada de decisão, garantia de conformidade, promoção da inovação e preparação para escalabilidade. A avaliação estabelece uma abordagem

estratégica para a modernização generativa da IA e maximiza os benefícios potenciais e, ao mesmo tempo, reduz os riscos.

## Como as organizações podem garantir uma implementação bem-sucedida após a avaliação?

As organizações devem desenvolver um plano de implementação claro que inclua marcos definidos, engajar as partes interessadas desde o início e adotar uma abordagem iterativa. Estabelecer um Centro de Excelência (CoE) e focar no desenvolvimento de talentos também são as melhores práticas recomendadas.

## Quais desafios as organizações podem enfrentar durante a avaliação?

Os desafios podem incluir resistência a mudanças, problemas de qualidade de dados e complexidades de conformidade. Enfrentar esses desafios exige promover uma cultura de inovação, garantir a prontidão dos dados e implementar medidas de segurança robustas.

## Como a avaliação aborda os requisitos regulatórios e de conformidade?

A avaliação avalia as medidas de conformidade atuais e identifica lacunas. Ele garante que as soluções-alvo cumpram os regulamentos relevantes e as leis de privacidade de dados e incorporem as melhores práticas de segurança para proteger informações confidenciais.

## Qual o papel do engajamento das partes interessadas no processo de avaliação?

O engajamento das partes interessadas é crucial para obter a adesão, alinhar as iniciativas de modernização aos objetivos de negócios e garantir uma implementação bem-sucedida. O envolvimento precoce e a comunicação clara dos benefícios são fundamentais para superar a resistência e promover o apoio.

## Como as organizações podem medir o sucesso de suas iniciativas generativas de modernização da IA após a avaliação?

O sucesso pode ser medido usando indicadores-chave de desempenho (KPIs) que se alinham às metas de negócios. O monitoramento e a avaliação regulares dessas métricas ajudam a orientar a tomada de decisões e a demonstrar o valor da modernização generativa da IA para as partes interessadas.

## Como a abordagem de avaliação difere para organizações de tamanhos variados (pequeno, médio ou corporativo) ou setores?

### Pequenas organizações:

- Pode ter recursos e experiência limitados para avaliações abrangentes
- Provavelmente se concentrará em casos de uso específicos de alto impacto, em vez da adoção em toda a empresa
- Pode depender mais de ferramentas e serviços de terceiros para avaliação
- O processo de avaliação pode ser menos formal e mais ágil

### Organizações de médio porte:

- Muitas vezes têm equipes dedicadas de TI ou de dados, mas podem não ter experiência especializada em IA
- Pode adotar uma abordagem em fases, começando com projetos piloto em departamentos importantes
- Necessidade de equilibrar a inovação com os sistemas e processos existentes
- A avaliação provavelmente envolve equipes multifuncionais

### Organizações corporativas:

- Normalmente, têm AI/ML equipes dedicadas e mais recursos para uma avaliação abrangente
- É necessário considerar integrações complexas com sistemas corporativos existentes
- Pode ter requisitos regulatórios específicos do setor a serem considerados
- A avaliação geralmente envolve processos formais de governança

# Recursos

- [IA generativa ativada AWS](#)
- [AWS oferece novos guias de inteligência artificial, aprendizado de máquina e IA generativa para planejar sua estratégia de IA](#) (postagem AWS no blog)
- [Melhores práticas para criar aplicativos de IA generativos em AWS](#) (postagem AWS do blog)
- [Gerador de aplicativos de IA em AWS](#) (Biblioteca de AWS soluções)
- [Capacidades generativas de IA](#) (arquitetura AWS de referência de segurança)
- [AWS estrutura generativa de melhores práticas de IA](#) (Guia AWS Audit Manager do usuário)
- [Escolha de um serviço generativo de IA](#) (guia de AWS decisão)
- [O que é o Amazon Bedrock?](#) (Guia do usuário do Amazon Bedrock)
- [O que é Amazon SageMaker AI?](#) (Guia do desenvolvedor de SageMaker IA da Amazon)

## Histórico do documento

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

| Alteração                          | Descrição | Data                  |
|------------------------------------|-----------|-----------------------|
| <a href="#">Publicação inicial</a> | —         | 6 de novembro de 2024 |

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