Avaliação generativa da carga de trabalho de IA

AWS Orientação prescritiva

AWS Orientação prescritiva: Avaliação generativa da carga de trabalho de IA

Table of Contents

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.
- Critérios: 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:

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 engenheir os 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 interessa das (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 avaliaçõe s 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 generativ as 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 operacion alização 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 necessári os 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 documento s, 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: atualizaç ões 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: verificaç ão 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 solicitaç ões/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 sequencia mento 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 principai s: 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 principal mente 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 treinamen to 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 individua is no armazenamento de objetos para maior flexibilidade. Grandes conjuntos de dados são armazenados em um banco de dados distribuí do (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 intersecc ionais no modelo generativo de IA?	Usando análise de imparcialidade intersecc ional, 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 conteinerização 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-proce ssamento 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 acelerado s 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 transferê ncia 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 ferrament as 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, frequênci a de atualização e uso. Usamos ferrament as 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 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
Publicação inicial	—	6 de novembro de 2024

AWS Glossário de orientação prescritiva

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

Números

7 Rs

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

- Refatorar/rearquitetar: mova uma aplicação e modifique sua arquitetura aproveitando ao máximo os recursos nativos de nuvem para melhorar a agilidade, a performance e a escalabilidade. Isso normalmente envolve a portabilidade do sistema operacional e do banco de dados. Exemplo: migrar seu banco de dados Oracle on-premises para o Amazon Aurora Edição Compatível com PostgreSQL.
- Redefinir a plataforma (mover e redefinir [mover e redefinir (lift-and-reshape]): mova uma aplicação para a nuvem e introduza algum nível de otimização a fim de aproveitar os recursos da nuvem. Exemplo: migrar seu banco de dados Oracle on-premises para o Amazon Relational Database Service (Amazon RDS) para Oracle na Nuvem AWS.
- Recomprar (drop and shop): mude para um produto diferente, normalmente migrando de uma licença tradicional para um modelo SaaS. Exemplo: migrar seu sistema de gerenciamento de relacionamento com o cliente (CRM) para o Salesforce.com.
- Redefinir a hospedagem (mover sem alterações [lift-and-shift])mover uma aplicação para a nuvem sem fazer nenhuma alteração a fim de aproveitar os recursos da nuvem. Exemplo: migrar seu banco de dados Oracle on-premises para o Oracle em uma instância do EC2 na Nuvem AWS.
- Realocar (mover o hipervisor sem alterações [hypervisor-level lift-and-shift]): mover a infraestrutura para a nuvem sem comprar novo hardware, reescrever aplicações ou modificar suas operações existentes. Você migra servidores de uma plataforma on-premises para um serviço de nuvem para a mesma plataforma. Exemplo: migrar um Microsoft Hyper-V aplicativo para o. AWS
- Reter (revisitar): mantenha as aplicações em seu ambiente de origem. Isso pode incluir aplicações que exigem grande refatoração, e você deseja adiar esse trabalho para um

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

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

A

ABAC

Consulte controle de acesso baseado em atributo.

serviços abstraídos

Veja serviços gerenciados.

ACID

Veja atomicidade, consistência, isolamento, durabilidade.

migração ativa-ativa

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

migração ativa-passiva

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

AGGREGATE FUNCTION

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

AI

Veja inteligência artificial.

AIOps

Veja operações de inteligência artificial.

anonimização

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

antipadrões

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

controle de aplicações

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

portfólio de aplicações

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

inteligência artificial (IA)

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

operações de inteligência artificial (AIOps)

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

criptografia assimétrica

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

atomicidade, consistência, isolamento, durabilidade (ACID)

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

controle de acesso por atributo (ABAC)

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

fonte de dados autorizada

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

Zona de disponibilidade

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

AWS Estrutura de adoção da nuvem (AWS CAF)

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

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

Uma ferramenta que avalia as cargas de trabalho de migração do banco de dados, recomenda estratégias de migração e fornece estimativas de trabalho. AWS O WQF está incluído com AWS

Schema Conversion Tool ()AWS SCT. Ela analisa esquemas de banco de dados e objetos de código, código de aplicações, dependências e características de performance, além de fornecer relatórios de avaliação.

B

bot malicioso

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

BCP

Veja planejamento de continuidade de negócios

gráfico de comportamento

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

sistema big-endian

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

classificação binária

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

filtro de bloom

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

blue/green deployment (implantação azul/verde)

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

bot

Uma aplicação de software que executa tarefas automatizadas na internet e simula a atividade ou interação humana. Alguns bots são úteis ou benéficos, como crawlers da web que indexam informações na internet. Outros bots, conhecidos como bots maliciosos, têm como objetivo causar disrupção ou danos a indivíduos ou organizações.

botnet

Redes de bots infectadas por malware e sob o controle de uma única parte, conhecidas como bot herder ou operador de bots. Os botnets são o mecanismo mais conhecido para escalar bots e seu impacto.

ramo

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

Acesso de emergência

Em circunstâncias excepcionais e por meio de um processo aprovado, um meio rápido para um usuário obter acesso a um Conta da AWS que ele normalmente não tem permissão para acessar. Para obter mais informações, consulte o indicador Implement break-glass procedures nas orientações do AWS Well-Architected.

estratégia brownfield

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

cache do buffer

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

capacidade de negócios

O que uma empresa faz para gerar valor (por exemplo, vendas, atendimento ao cliente ou marketing). As arquiteturas de microsserviços e as decisões de desenvolvimento podem

ser orientadas por recursos de negócios. Para obter mais informações, consulte a seção Organizados de acordo com as capacidades de negócios do whitepaper Executar microsserviços conteinerizados na AWS.

planejamento de continuidade de negócios (BCP)

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

C

CAF

Veja AWS Cloud Adoption Framework.

implantação canário

O lançamento lento e incremental de uma versão para usuários finais. Quando estiver confiante, você implanta a nova versão e substitui a versão atual por completo.

CCoE

Veja Centro de Excelência da Nuvem.

CDC

Veja captura de dados de alteração.

captura de dados de alterações (CDC)

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

engenharia do caos

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

CI/CD

Veja integração e entrega contínuas.

classificação

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

criptografia no lado do cliente

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

Centro de excelência em nuvem (CCoE)

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

computação em nuvem

A tecnologia de nuvem normalmente usada para armazenamento de dados remoto e gerenciamento de dispositivos de IoT. A computação em nuvem é normalmente conectada à tecnologia de computação de borda.

modelo operacional em nuvem

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

estágios de adoção da nuvem

As quatro fases pelas quais as organizações normalmente passam ao migrar para a Nuvem AWS:

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

Esses estágios foram definidos por Stephen Orban na postagem do blog The Journey Toward Cloud-First & the Stages of Adoption no blog de estratégia Nuvem AWS empresarial. Para obter

informações sobre como eles se relacionam com a estratégia de AWS migração, consulte o guia de preparação para migração.

CMDB

Veja banco de dados de gerenciamento de configuração.

repositório de código

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

cache frio

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

dados frios

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

visão computacional (CV)

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

desvio de configuração

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

banco de dados de gerenciamento de configuração (CMDB)

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

pacote de conformidade

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

integração contínua e entrega contínua (CI/CD)

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

CV

Veja visão computacional.

D

dados em repouso

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

classificação de dados

Um processo para identificar e categorizar os dados em sua rede com base em criticalidade e confidencialidade. É um componente crítico de qualquer estratégia de gerenciamento de riscos de segurança cibernética, pois ajuda a determinar os controles adequados de proteção e retenção para os dados. A classificação de dados é um componente do pilar de segurança no AWS Well-Architected Framework. Para obter mais informações, consulte Classificação de dados.

desvio de dados

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

dados em trânsito

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

data mesh

Um framework de arquitetura que fornece propriedade de dados distribuída e descentralizada com gerenciamento e governança centralizados.

minimização de dados

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

perímetro de dados

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

pré-processamento de dados

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

proveniência dos dados

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

titular dos dados

Um indivíduo cujos dados estão sendo coletados e processados.

data warehouse

Um sistema de gerenciamento de dados compatível com business intelligence, como analytics. Os data warehouses geralmente contêm grandes quantidades de dados históricos e geralmente são usados para consultas e análises.

linguagem de definição de dados (DDL)

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

linguagem de manipulação de dados (DML)

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

DDL

Veja linguagem de definição de banco de dados.

deep ensemble

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

Aprendizado profundo

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

defense-in-depth

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

administrador delegado

Em AWS Organizations, um serviço compatível pode registrar uma conta de AWS membro para administrar as contas da organização e gerenciar as permissões desse serviço. Essa conta é chamada de administrador delegado para esse serviço Para obter mais informações e uma lista de serviços compatíveis, consulte Serviços que funcionam com o AWS Organizations na documentação do AWS Organizations .

implantação

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

ambiente de desenvolvimento

Veja ambiente.

controle detectivo

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

mapeamento do fluxo de valor de desenvolvimento (DVSM)

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

gêmeo digital

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

tabela de dimensões

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

desastre

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

Recuperação de desastres (RD)

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

DML

Veja linguagem de manipulação de banco de dados.

design orientado por domínio

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

DR

Veja recuperação de desastres.

Detecção da oscilação

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

DVSM

Veja mapeamento do fluxo de valor de desenvolvimento.

E

EDA

Veja análise exploratória de dados.

EDI

Veja intercâmbio eletrônico de dados.

computação de borda

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

intercâmbio eletrônico de dados (EDI)

A troca automatizada de documentos comerciais entre organizações. Para obter mais informações, consulte O que é EDI (Intercâmbio eletrônico de dados)?.

criptografia

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

chave de criptografia

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

endianismo

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

endpoint

Veja endpoint de serviço.

serviço de endpoint

Um serviço que pode ser hospedado em uma nuvem privada virtual (VPC) para ser compartilhado com outros usuários. Você pode criar um serviço de endpoint com AWS PrivateLink e conceder permissões a outros diretores Contas da AWS ou a AWS Identity and Access Management (IAM). Essas contas ou entidades principais podem se conectar ao serviço de endpoint de maneira privada criando endpoints da VPC de interface. Para obter mais informações, consulte Criar um serviço de endpoint na documentação do Amazon Virtual Private Cloud (Amazon VPC).

planejamento de recursos empresariais (ERP)

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

criptografia envelopada

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

ambiente

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

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

epic

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

ERP

Veja planejamento de recursos empresariais.

análise exploratória de dados (EDA)

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

F

tabela de fatos

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

Antecipar-se à falha

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

delimitação de isolamento contra falhas

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

ramificação de recursos

Veja ramificação.

recursos

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

importância do recurso

O quanto um recurso é importante para as previsões de um modelo. Isso geralmente é expresso como uma pontuação numérica que pode ser calculada por meio de várias técnicas, como Shapley Additive Explanations (SHAP) e gradientes integrados. Para obter mais informações, consulte Interpretabilidade do modelo de aprendizado de máquina com AWS.

transformação de recursos

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

prompt few shot

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

FGAC

Veja controle de acesso refinado.

Controle de acesso refinado (FGAC)

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

migração flash-cut

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

FM

Veja modelo de base.

modelo de base (FM)

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

G

IA generativa

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

bloqueio geográfico

Veja restrições geográficas.

restrições geográficas (bloqueio geográfico)

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

Fluxo de trabalho do GitFlow

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

golden image

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

estratégia greenfield

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

barreira de proteção

Uma regra de alto nível que ajuda a governar recursos, políticas e conformidade em todas as unidades organizacionais ()OUs. Barreiras de proteção preventivas impõem políticas para garantir o alinhamento a padrões de conformidade. Elas são implementadas usando políticas de controle de serviço e limites de permissões do IAM. Barreiras de proteção detectivas detectam violações de políticas e problemas de conformidade e geram alertas para remediação. Eles são implementados usando AWS Config, AWS Security Hub CSPM, Amazon GuardDuty AWS Trusted Advisor, Amazon Inspector e verificações personalizadas AWS Lambda .

H

HA

Veja alta disponibilidade.

migração heterogênea de bancos de dados

Migrar seu banco de dados de origem para um banco de dados de destino que usa um mecanismo de banco de dados diferente (por exemplo, Oracle para Amazon Aurora). A migração heterogênea geralmente faz parte de um esforço de redefinição da arquitetura, e converter

o esquema pode ser uma tarefa complexa. O AWS fornece o AWS SCT para ajudar nas conversões de esquemas.

alta disponibilidade (HA)

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

modernização de historiador

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

dados de hold-out

Uma parte dos dados históricos rotulados que são retidos de um conjunto de dados usado para treinar um modelo de machine learning. Você pode usar dados de hold-out para avaliar a performance do modelo comparando as predições do modelo com os dados de retenção.

migração homogênea de bancos de dados

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

dados quentes

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

hotfix

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

período de hipercuidados

Imediatamente após a substituição, o período em que uma equipe de migração gerencia e monitora as aplicações migradas na nuvem para resolver quaisquer problemas. Normalmente,

a duração desse período é de 1 a 4 dias. No final do período de hipercuidados, a equipe de migração normalmente transfere a responsabilidade pelas aplicações para a equipe de operações de nuvem.

eu

IaC

Veja infraestrutura como código.

Política baseada em identidade

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

aplicação ociosa

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

IIoT

Veja Internet das Coisas Industrial.

infraestrutura imutável

Um modelo que implanta uma nova infraestrutura para workloads de produção em vez de atualizar, aplicar patches ou modificar a infraestrutura existente. Infraestruturas imutáveis são inerentemente mais consistentes, confiáveis e preditivas do que infraestruturas mutáveis. Para obter mais informações, consulte a prática recomendada Implantar usando infraestrutura imutável no AWS Well-Architected Framework.

VPC de entrada (admissão)

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

migração incremental

Uma estratégia de substituição na qual você migra a aplicação em pequenas partes, em vez de realizar uma única substituição completa. Por exemplo, é possível mover inicialmente

apenas alguns microsserviços ou usuários para o novo sistema. Depois de verificar se tudo está funcionando corretamente, mova os microsserviços ou usuários adicionais de forma incremental até poder descomissionar seu sistema herdado. Essa estratégia reduz os riscos associados a migrações de grande porte.

Indústria 4.0

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

infraestrutura

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

Infraestrutura como código (IaC)

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

Internet industrial das coisas (IIoT)

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

VPC de inspeção

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

Internet das coisas (IoT)

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

interpretabilidade

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

IoT

Veja Internet das Coisas.

Biblioteca de informações de TI (ITIL)

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

Gerenciamento de serviços de TI (ITSM)

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

ITIL

Veja biblioteca de informações de TI.

ITSM

Veja gerenciamento de serviços de TI.

L

controle de acesso baseado em etiqueta (LBAC)

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

zona de pouso

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

informações sobre zonas de pouso, consulte Configurar um ambiente da AWS com várias contas seguro e escalável.

grande modelo de linguagem (LLM)

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

migração de grande porte

Uma migração de 300 servidores ou mais.

LBAC

Veja controle de acesso baseado em rótulo.

privilégio mínimo

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

mover sem alterações (lift-and-shift)

Veja 7 Rs.

sistema little-endian

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

LLM

Veja grande modelo de linguagem.

ambientes inferiores

Veja ambiente.

M

machine learning (ML)

Um tipo de inteligência artificial que usa algoritmos e técnicas para reconhecimento e aprendizado de padrões. O ML analisa e aprende com dados gravados, por exemplo, dados da

Internet das Coisas (IoT), para gerar um modelo estatístico baseado em padrões. Para obter mais informações, consulte Machine learning.

ramificação principal

Veja ramificação.

Malware

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

Serviços gerenciados

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

sistema de execução de manufatura (MES)

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

MAP

Veja Programa de Aceleração da Migração.

mecanismo

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

conta de membro

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

MES

Veja sistema de execução de manufatura.

Transporte de Telemetria de Enfileiramento de Mensagens (MQTT)

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

microsserviço

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

arquitetura de microsserviços

> Uma abordagem à criação de aplicações com componentes independentes que executam cada processo de aplicação como um microsserviço. Esses microsserviços se comunicam por meio de uma interface bem definida usando leveza. APIs Cada microsserviço nessa arquitetura pode ser atualizado, implantado e escalado para atender à demanda por funções específicas de uma aplicação. Para obter mais informações, consulte Implementação de microsserviços em. AWS

Programa de Aceleração da Migração (MAP)

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

migração em escala

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

fábrica de migração

> Equipes multifuncionais que simplificam a migração de workloads por meio de abordagens automatizadas e ágeis. As equipes da fábrica de migração geralmente incluem operações,

analistas e proprietários de negócios, engenheiros de migração, desenvolvedores e DevOps profissionais que trabalham em sprints. Entre 20 e 50% de um portfólio de aplicações corporativas consiste em padrões repetidos que podem ser otimizados por meio de uma abordagem de fábrica. Para obter mais informações, consulte discussão sobre fábricas de migração e o guia do Cloud Migration Factory neste conjunto de conteúdo.

metadados de migração

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

padrão de migração

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

Avaliação de Portfólio para Migração (MPA)

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

Avaliação de Preparação para Migração (MRA)

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

estratégia de migração

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

ML

Veja machine learning.

modernização

Transformar uma aplicação desatualizada (herdada ou monolítica) e sua infraestrutura em um sistema ágil, elástico e altamente disponível na nuvem para reduzir custos, ganhar eficiência e aproveitar as inovações. Para obter mais informações, consulte Strategy for modernizing applications in the Nuvem AWS.

avaliação de preparação para modernização

Uma avaliação que ajuda a determinar a preparação para modernização das aplicações de uma organização. Ela identifica benefícios, riscos e dependências e determina o quão bem a organização pode acomodar o estado futuro dessas aplicações. O resultado da avaliação é um esquema da arquitetura de destino, um roteiro que detalha as fases de desenvolvimento e os marcos do processo de modernização e um plano de ação para abordar as lacunas identificadas. Para obter mais informações, consulte Evaluating modernization readiness for applications in the Nuvem AWS.

aplicações monolíticas (monólitos)

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

MPA

Veja Avaliação do Portfólio para Migração.

MQTT

Veja Transporte de Telemetria de Enfileiramento de Mensagens.

classificação multiclasse

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

infraestrutura mutável

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

O

OAC

Veja controle de acesso de origem.

OAI

Veja identidade de acesso de origem.

OCM

Veja gerenciamento de alterações organizacionais.

migração offline

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

OI

Veja integração de operações.

Ola

Veja acordo de nível operacional.

migração online

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

OPC-UA

Veja Open Process Communications - Unified Architecture.

Open Process Communications - Unified Architecture (OPC-UA)

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

acordo de nível operacional (OLA)

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

análise de prontidão operacional (ORR)

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

tecnologia operacional (TO)

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

integração de operações (OI)

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

trilha organizacional

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

gerenciamento de alterações organizacionais (OCM)

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

controle de acesso de origem (OAC)

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

Identidade do acesso de origem (OAI)

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

ORR

Veja análise de prontidão operacional.

OT

Veja tecnologia operacional.

VPC de saída (egresso)

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

P

limite de permissões

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

Informações de identificação pessoal (PII)

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

PII

Veja informações de identificação pessoal.

manual

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

PLC

Veja controlador lógico programável.

PLM

Veja gerenciamento do ciclo de vida do produto.

política

Um objeto que pode definir permissões (veja política baseada em identidade), especificar condições de acesso (veja políta baseada em recurso) ou definir as permissões máximas para todas as contas em uma organização no AWS Organizations (veja política de controle de serviços).

persistência poliglota

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

avaliação do portfólio

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

predicado

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

pushdown de predicados

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

controle preventivo

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

principal (entidade principal)

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

Privacidade por design

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

zonas hospedadas privadas

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

controle proativo

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

gerenciamento do ciclo de vida do produto (PLM)

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

ambiente de produção

Veja ambiente.

controlador lógico programável (PLC)

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

encadeamento de prompts

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

pseudonimização

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

publish/subscribe (pub/sub)

Um padrão que permite comunicações assíncronas entre microsserviços para melhorar a escalabilidade e a capacidade de resposta. Por exemplo, em um MES baseado em microsserviços, um microsserviço pode publicar mensagens de eventos em um canal em que outros microsserviços possam assinar. O sistema pode adicionar novos microsserviços sem alterar o serviço de publicação.

Q

plano de consulta

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

regressão de planos de consultas

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

R

Matriz RACI

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

RAG

Veja geração aumentada via recuperação.

ransomware

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

Matriz RASCI

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

RCAC

Veja controle de acesso por linha e coluna.

réplica de leitura

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

Redefinir arquitetura

Veja 7 Rs.

objetivo de ponto de recuperação (RPO).

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

objetivo de tempo de recuperação (RTO)

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

refatorar

Veja 7 Rs.

Região

Uma coleção de AWS recursos em uma área geográfica. Cada um Região da AWS é isolado e independente dos outros para fornecer tolerância a falhas, estabilidade e resiliência. Para obter informações, consulte Specify which Regiões da AWS your account can use.

regressão

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

redefinir a hospedagem

Veja 7 Rs.

versão

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

realocar

Veja 7 Rs.

redefinir a plataforma

Veja 7 Rs.

recomprar

Veja 7 Rs.

resiliência

A capacidade de uma aplicação de resistir ou se recuperar de interrupções. Alta disponibilidade e recuperação de desastres são considerações comuns ao planejar a resiliência na Nuvem AWS. Para obter mais informações, consulte Nuvem AWS Resilience.

política baseada em recurso

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

matriz responsável, accountable, consultada, informada (RACI)

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

controle responsivo

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

reter

Veja 7 Rs.

Retirada

Veja 7 Rs.

Geração Aumentada de Recuperação (RAG)

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

alternância

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

controle de acesso por linha e coluna (RCAC)

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

RPO

Veja objetivo de ponto de recuperação.

RTO

Veja objetivo de tempo de recuperação.

runbook

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

S

SAML 2.0

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

SCADA

Veja controle de supervisão e aquisição de dados.

SCP

Veja política de controle de serviço.

secret

Em AWS Secrets Manager, informações confidenciais ou restritas, como uma senha ou credenciais de usuário, que você armazena de forma criptografada. Consiste no valor secreto e em seus metadados. O valor secreto pode ser binário, uma única string ou várias strings. Para obter mais informações, consulte What's in a Secrets Manager secret? na documentação do Secrets Manager.

segurança desde a concepção

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

controle de segurança

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

hardening da segurança

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

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

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

automação de resposta de segurança

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

Criptografia do lado do servidor

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

política de controle de serviços (SCP)

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

service endpoint (endpoint de serviço)

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

acordo de serviço (SLA)

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

indicador de nível de serviço (SLI)

Uma avaliação de um aspecto de performance de um serviço, como taxa de erro, disponibilidade ou throughput.

objetivo de nível de serviço (SLO)

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

modelo de responsabilidade compartilhada

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

SIEM

Veja sistema de gerenciamento de eventos e informações de segurança.

ponto único de falha (SPOF)

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

SLA

Veja cacordo de serviço.

SLI

Veja indicador de nível de serviço.

SLO

Veja objetivo de nível de serviço.

split-and-seed modelo

Um padrão para escalar e acelerar projetos de modernização. À medida que novos recursos e lançamentos de produtos são definidos, a equipe principal se divide para criar novas equipes de produtos. Isso ajuda a escalar os recursos e os serviços da sua organização, melhora a produtividade do desenvolvedor e possibilita inovações rápidas. Para obter mais informações, consulte Phased approach to modernizing applications in the Nuvem AWS.

SPOF

Veja ponto único de falha.

esquema em estrela

Uma estrutura organizacional de banco de dados que usa uma grande tabela de fatos para armazenar dados transacionais ou medidos e usa uma ou mais tabelas dimensionais menores para armazenar atributos de dados. Essa estrutura foi projetada para ser usada em um data warehouse ou para fins de inteligência comercial.

padrão strangler fig

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

sub-rede

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

controle supervisório e aquisição de dados (SCADA)

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

symmetric encryption (criptografia simétrica)

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

testes sintéticos

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

prompt do sistema

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

T

tags

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

variável-alvo

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

lista de tarefas

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

ambiente de teste

Veja ambiente.

treinamento

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

gateway de trânsito

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

fluxo de trabalho baseado em troncos

Uma abordagem na qual os desenvolvedores criam e testam recursos localmente em uma ramificação de recursos e, em seguida, mesclam essas alterações na ramificação principal. A ramificação principal é então criada para os ambientes de desenvolvimento, pré-produção e produção, sequencialmente.

Acesso confiável

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

tuning (ajustar)

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

equipe de duas pizzas

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

U

incerteza

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

tarefas indiferenciadas

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

ambientes superiores

Veja ambiente.

V

aspiração

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

controle de versões

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

emparelhamento da VPC

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

Vulnerabilidade

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

W

cache quente

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

dados mornos

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

função de janela

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

workload

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

workstreams

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

WORM

Veja gravação única e várias leituras.

WQF

Veja AWS Workload Qualification Framework.

gravação única e várias leituras (WORM)

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

Z

exploração de dia zero

Um ataque, normalmente malware, que tira proveito de uma vulnerabilidade zero-day.

vulnerabilidade de dia zero

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

prompt zero shot

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

eficácia dos prompts zero-shot depende da complexidade da tarefa e da qualidade do prompt. Veja também prompts few-shot.

aplicação zumbi

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

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