databricks

Guia do exame Databricks

Databricks Certified Generative AI Engineer Associate

Forneça Feedback sobre o guia de exame

Finalidade deste Guia do Exame

Este guia de exame fornece uma visão geral do exame e o que ele abrange para ajudá-lo a determinar sua preparação para o exame. Este documento será atualizado sempre que houver alterações em um exame (e quando essas alterações entrarem em vigor) para que você possa estar preparado. **Esta versão cobre a versão atualmente ativa até e incluindo 18 de março de 2026. Por favor, verifique novamente duas semanas antes da data do exame para garantir que tem a versão mais atualizada.**

Descrição do público

O exame de certificação Databricks Certified Generative AI Engineer Associate avalia a capacidade de um indivíduo de projetar e implementar soluções habilitadas para LLM usando a Databricks. Isso inclui a decomposição de problemas para dividir requisitos complexos em tarefas gerenciáveis, bem como a escolha de modelos, ferramentas e abordagens apropriadas do cenário atual de generative AI para o desenvolvimento de soluções abrangentes. Ele também avalia ferramentas específicas do Databricks, como Vector Search para pesquisas de similaridade semântica, Model Serving para implantar modelos e soluções, MLflow para gerenciar o ciclo de vida da solução e Unity Catalog para governança de dados. Espera-se que as pessoas aprovadas neste exame criem e implantem aplicativos RAG de alto desempenho e cadeias LLM que aproveitem ao máximo a Databricks e seu conjunto de ferramentas.

Sobre o Exame

- Número de questões pontuadas: 45 itens de múltipla escolha ou seleção múltipla*
  Limite de tempo: 90 minutos
- Taxa de inscrição: $200
- Método de entrega: Prova supervisionada online
- Pré-requisito: Nenhum é necessário; participação em cursos relacionados e seis meses de experiência prática são altamente recomendados.
- Validade: 2 anos.

- Recertificação: A recertificação é necessária a cada dois anos para manter seu estado certificado. Para se re-certificar, é preciso fazer o exame completo que está vivo. Consulte a seção "Preparando-se para o exame" na página do exame para se preparar para fazer o exame novamente.
- Questões não pontuadas: Os exames podem incluir questões não pontuadas para coletar informações estatísticas para uso futuro. Essas questões não são identificadas no formulário e não afetam sua pontuação. Um tempo adicional é considerado nos exames para dar conta dessas questões.

Preparação recomendada

- Todos os cursos atuais da Databricks Academy ILT relacionados ao papel de aprendiz de Generative AI, especificamente Engenharia de Generative AI com Databricks.
- Self-Paced (disponível na Databricks Academy): Engenharia de Generative AI com Databricks, com esses cursos:
    - Construir Agents de Recuperação em Databricks
    - Construir aplicativos de Agent único em Databricks
    - Avaliação e governança de aplicativos de Generative AI
    - Implantação e monitoramento de aplicativos de Generative AI
- Conhecimento dos LLMs atuais e suas capacidades
- Conhecimento de engenharia de prompts, geração de prompts e avaliação
- Conhecimento de ferramentas e serviços online relacionados atuais como LangChain, Hugging Face Transformers, etc.
- Conhecimento prático de Python e suas bibliotecas que dão suporte a aplicativos RAG e ao desenvolvimento de cadeias de LLM
- Conhecimento prático de APIs atuais para preparação de dados, encadeamento de modelos, etc.
- Recursos de documentação relevante do Databricks

Estrutura do exame

Seção 1: Projetar Aplicativos

- Desenhe um prompt que obtenha uma resposta especificamente formatada
- Selecione tarefas do modelo para atender a um determinado requisito de negócios
- Selecione componentes da cadeia para entrada e saída desejadas do modelo
- Traduz metas de casos de uso de negócios em uma descrição das entradas e saídas desejadas para o pipeline de IA
- Defina e ordenar ferramentas que coletam conhecimento ou tomem ações para raciocínio em vários estágios

- Determina como e quando utilizar os Blocos de Agent (Assistente de Conhecimento, Supervisor Multiagent, Extração de Informação) para resolver problemas.

Seção 2: Preparação de dados

- Aplique uma estratégia de chunking para uma determinada estrutura de documento e restrições de modelo.
- Filtre conteúdo desnecessário em documentos de origem que prejudicam a qualidade de um aplicativo RAG.
- Escolha o pacote Python apropriado para extrair o conteúdo de documentos a partir dos dados e formato de origem fornecidos.
- Defina operações e sequência para gravar determinado texto segmentado em tabelas Delta Lake no Unity Catalog
- Identifique os documentos de origem necessários que fornecem o conhecimento e a qualidade essenciais para um determinado aplicativo RAG .
- Use ferramentas e métricas para avaliar o desempenho da recuperação
- Projete sistemas de recuperação usando estratégias avançadas de chunking.
- Explique o papel da reclassificação no processo de recuperação de informação.

Seção 3: Desenvolvimento de Aplicativos

- Selecione Langchain/ferramentas semelhantes para uso em um aplicativo de Generative AI.
- Avalie qualitativamente as respostas para identificar problemas comuns, como qualidade e segurança.
- Selecione a estratégia de chunking com base no modelo e na avaliação de recuperação.
- Aprimore um prompt com contexto adicional a partir da entrada de um usuário com base em principais campos, termos e intenções.
- Crie um prompt que ajuste a resposta de um LLM de uma resposta padrão para uma saída desejada.
- Implemente LLM guardrails para evitar resultados negativos.
- Selecione o melhor LLM com base nos atributos do aplicativo a ser desenvolvido
- Selecione um modelo de incorporação de tamanho de contexto com base nos documentos de origem, nas queries esperadas e na estratégia de otimização
- Selecione um modelo de um hub ou marketplace de modelos para uma tarefa, com base em metadados/model cards
- Selecione o melhor modelo para uma determinada tarefa com base em métricas comuns geradas em experimentos
- Utilize o MLflow e Agent Framework para desenvolver sistemas agentic

- Compare as fases de avaliação e monitoramento do ciclo de vida do aplicativo Gen AI
- Permitir que os sistemas multi-agents utilizem o Genie Spaces ou a API conversacional para recuperar dados.

Seção 4: Montagem e implantação de aplicativos

- Codifique uma cadeia usando um modelo pyfunc com pré e pós-processamento
- Controle o acesso aos recursos de model serving endpoints
- Codifique uma cadeia simples de acordo com os requisitos
- Escolha os elementos básicos necessários para criar um aplicativo RAG: model flavor, modelo de incorporação, recuperador, dependências, exemplos de entrada, assinatura de modelo
- Registre o modelo no Unity Catalog usando o MLflow
- Crie e query um índice de Vector Search
- Identifique como servir um aplicativo LLM que utiliza Foundation Model APIs
- Explique os principais conceitos e componentes do Mosaic AI Vector Search
- Identifique cargas de trabalho de inferência em lote e aplica `ai_query()` adequadamente
- Configure o vector search para uma solução específica com base no número de incorporações, frequência de atualização, latência e requisitos de custo.
- Configure um armazenamento de dados persistente para armazenar e recuperar informações intermédias na memória ou informações estruturadas.
- Aplique as melhores práticas de CI/CD, como a atualização de um índice do Vector Search, a promoção de prompts em diferentes ambientes e o teste de componentes individuais de um agent.
- Integre servidores MCP geridos, externos e personalizados com base nos requisitos de um aplicativo específico.
- Aplique o controle de versão de prompts e gerencie o ciclo de vida dos prompts.
- Desenvolva uma interface de utilizador interativa apropriada para um cenário de utilização de agents (aplicações, Slack, Teams etc.)

Seção 5: Governança

- Use técnicas de mascaramento como guard rails para atender a um objetivo de desempenho
- Selecione técnicas de guardrail para proteger contra entradas de usuários maliciosos para um aplicativo de Gen AI
- Use requisitos legais de licenciamento para fontes de dados para evitar risco legal
- Recomende uma alternativa para mitigação de texto problemático em uma fonte de dados que alimenta um aplicativo GenAI

Seção 6: Avaliação e Monitoramento

- Selecione uma LLM (tamanho e arquitetura) com base em um conjunto de métricas de avaliação quantitativa
- Selecione as principais métricas a serem monitoradas para um cenário específico de implantação de LLM
- Avalie o desempenho do modelo em um aplicativo RAG usando o MLflow
- Use o log de inferência para avaliar o desempenho do aplicativo RAG implantado
- Use os recursos do Databricks para controlar os custos de LLM
- Use tabelas de inferência e monitoramento de agents para acompanhar um LLM endpoint em tempo real
- Identifique judges de avaliação que exigem ground truth
- Compare as fases de avaliação e monitoramento do ciclo de vida do aplicativo GenAI
- Utilize o AI Gateway (Tabelas de Inferência, Tabelas de Utilização e limitação de taxa) para acompanhar um LLM ou agent implementado através do Agent Framework.
- Utilize os Databricks custom Scorers para avaliar agents e LLMs.
- Incorpore o feedback dos SMEs para melhorar o desempenho do agent.

Exemplos de questões

Essas questões são semelhantes aos itens de questões reais e dão a você uma noção geral de como as questões são feitas neste exame. Eles incluem os objetivos do exame, como estão indicados no guia do exame, e fornecem um exemplo de questões que se alinham ao objetivo. O guia do exame lista todos os objetivos que podem ser abordados em um exame. A melhor maneira de se preparar para um exame de certificação é revisar o resumo no guia do exame.

Questão 1

Objetivo: Aplique uma estratégia de chunking para uma determinada estrutura de documento e restrições de modelo

Um engenheiro de Generative AI está carregando 150 milhões de incorporações em um vector database que suporta no máximo 100 milhões.

Quais DUAS ações eles podem tomar para reduzir a contagem de registros?

A. Aumentar o tamanho do documento chunk
B. Diminuir a sobreposição entre chunks
C. Diminuir o tamanho do documento chunk
D. Aumentar a sobreposição entre chunks
E. Usar um modelo de incorporação menor

Questão 2

Objetivo: Identifique os documentos de origem que fornecem o conhecimento e a qualidade necessários para um aplicativo RAG.

Um engenheiro de Generative AI está avaliando as respostas de um aplicativo GenAI voltado para o cliente que eles estão desenvolvendo para ajudar na venda de peças automotivas. O aplicativo exige que o cliente insira `account_id` e `transaction_id` explicitamente para responder às questões. Após o lançamento inicial, o feedback dos clientes foi de que o aplicativo se saiu bem em responder aos detalhes do pedido e do faturamento, mas não respondeu com precisão às perguntas sobre data de envio e prevista de chegada.

Qual das abordagens a seguir melhoraria a capacidade do aplicativo de responder a essas questões?

A. Criar um vector store que inclua as políticas de envio da empresa e as condições de

pagamento para todas as peças automotivas

B. Criar uma tabela feature store com `transaction_id` como primary key, que é preenchida com dados de fatura e data de entrega esperada
C. Fornecer dados de exemplo para datas de chegada esperadas como um dataset de ajuste e, em seguida, ajuste periodicamente o modelo para que ele tenha informações de envio atualizadas
D. Alterar o prompt de bate-papo para inserir quando o pedido foi feito e instrua o modelo a adicionar 14 dias a isso, pois nenhum método de envio deve exceder 14 dias

Questão 3

Objetivo: Escolha o pacote Python apropriado para extrair o conteúdo do documento dos dados e formato de origem fornecidos.

Um engenheiro de Generative AI está criando um aplicativo RAG que contará com o contexto recuperado de documentos de origem que foram digitalizados e salvos como arquivos de imagem em formatos como .jpeg ou .png. Eles querem desenvolver uma solução usando a menor quantidade de linhas de código.

Qual pacote Python deve ser usado para extrair o texto dos documentos de origem?

A. beautifulsoup
B. scrapy
C. pytesseract
D. pyquery

Questão 4

Objetivo: Selecione um comprimento de contexto de modelo de incorporação com base nos documentos de origem, nas queries esperadas e na estratégia de otimização

Um engenheiro de Generative AI está criando um aplicativo baseado em LLM. Os documentos para seu recuperador foram divididos em partes de até 512 tokens cada. O engenheiro de Generative AI sabe que o custo e a latência são mais importantes do que a qualidade para esse aplicativo. Eles têm vários níveis de comprimento de contexto para escolher.

Qual suprirá sua necessidade?

A. Comprimento do contexto 512: o menor modelo é de 0,13 GB e dimensão de incorporação 384

B. Comprimento de contexto 514: o menor modelo é de 0,44 GB e dimensão de incorporação 768
C. Comprimento de contexto 2048: o menor modelo é de 11 GB e dimensão de incorporação 2560
D. Comprimento de contexto 32768: o menor modelo tem 14 GB e dimensão de incorporação 4096

Questão 5

Objetivo: Selecione o melhor LLM com base nos atributos do aplicativo a ser desenvolvido

Um engenheiro de Generative AI gostaria de criar um aplicativo que possa atualizar um campo de memorando com cerca de um parágrafo de comprimento para apenas uma única frase que mostre a intenção do campo de memorando, mas que se encaixe no front-end do aplicativo.

Com qual categoria de tarefa de Processamento de Linguagem Natural eles devem avaliar possíveis LLMs para esse aplicativo?

A. text2text Generation
B. Sentencizer
C. Classificação de Texto
D. Resumo

Questão 6

Objetivo: Configurar a *vector search* para uma solução específica com base no número de incorporações, frequência de actualização, latência e requisitos de custo.

Um engenheiro de Generative AI que trabalha para um retalhista online está a tentar melhorar a sua funcionalidade de pesquisa com vector search e filtragem de metadados. As pesquisas podem chegar aos 80 por segundo e a latência é a métrica mais crítica. Não se importam com os custos iniciais de desenvolvimento se isso melhorar a precisão sem prejudicar a latência. O inventário é composto por 100 milhões de artigos em todo o país.

Como deve o engenheiro configurar isso?

A. Utilizar o modelo de incorporação GTE Large, utilizar o vector search padrão com a pesquisa híbrida e a reclassificação ativadas.
B. Utilizar o modelo de incorporação GTE Large, utilize a pesquisa vetorial otimizada para armazenamento com a pesquisa híbrida e a reclassificação ativadas.
C. Ajustar um modelo de incorporação personalizado, utilizar o vector search predefinido e

manter a pesquisa híbrida e a reclassificação desativadas.
D. Ajustar um modelo de incorporação personalizado, utilizar o vector search otimizado para armazenamento e manter a pesquisa híbrida e a reclassificação desativadas.

Questão 7

Objectivo: Aplicar as melhores práticas de CI/CD, tais como a atualização de um índice de Vector Search, a promoção de prompts entre ambientes e o teste de componentes individuais de um agent.

Um Engenheiro de Generative AI necessita de gerir modelos de prompt para um agent nos ambientes de desenvolvimento, homologação e produção. A equipa requer um processo de lançamento controlado: os prompts são atualizados em desenvolvimento, validados em homologação com testes automatizados e promovidos para produção apenas após aprovação. A solução deve preservar o histórico de versões e permitir o regresso a uma versão anterior do prompt, se necessário.

Qual a abordagem que suporta este fluxo de trabalho de promoção?

A. Armazenar os modelos de prompts no repositório da aplicação e promova-os, fundindo a ramificação de teste na ramificação de produção após a aprovação dos testes.
B. Acompanhar os avisos como versões do MLflow e promova-os utilizando aliases após a aprovação.
C. Guardar os prompts num ficheiro JSON no executor de CI e sobrescrever o aviso de produção a cada execução.
D. Colocar os prompts em tabelas Delta e sobrescrever a tabela em produção a cada implementação para garantir a consistência.

Questão 8

Objectivo: Desenvolver uma interface de utilizador interactiva apropriada para um cenário de utilização por agents (aplicativos, Slack, Teams, etc.).

Um Engenheiro de Generative AI está a criar um aplicativo Databricks que permite aos agents de apoio ao cliente fazer perguntas e receber respostas baseadas em PDFs internos. Requisitos: os utilizadores devem autenticar-se com a sua identidade corporativa, a aplicação deve chamar um Mosaic AI Agent endpoint sem expor tokens de longa duração no navegador e o acesso às respostas deve respeitar as permissões de cada utilizador.

Qual a abordagem que atende a esses requisitos?

A. Utilizar um backend de aplicativo Databricks para chamar o Agent endpoint com as credenciais do aplicativo e aplicar a identidade/permissões do utilizador através do contexto autenticado do aplicativo.
B. Armazenar um Databricks personal access token (PAT) no JavaScript da aplicação e chamar o Agent endpoint diretamente a partir do separador.
C. Publicar o Agent endpoint publicamente e proteger-o com uma API key incorporada no frontend do aplicativo.
D. Exportar os PDFs para um bucket público para que o Agent os possa ler sem verificações de identidade.

Questão 9

Objectivo: Integrar servidores MCP geridos, externos e personalizados com base nos requisitos de um aplicativo específico.

Um Engenheiro de Generative AI está a desenvolver um agente assistente de pesquisa que necessita de aceder a informações factuais de uma fonte de dados da internet e realizar pesquisas na web utilizando uma API externa. A Databricks fornece um servidor MCP gerido para esta fonte de dados da internet, e um servidor MCP externo está disponível para a API externa que requer uma key. O aplicativo deve minimizar a sobrecarga de manutenção, garantindo um acesso fiável a ambas as fontes de dados em produção.

Que duas ações deve o engenheiro tomar para integrar estas fontes de dados no agent?

A. Criar um servidor MCP personalizado que encapsule tanto o recurso da internet como as APIs externas numa única interface unificada para o agent chamar.
B. Utilizar o servidor MCP gerido do navegador da Web para navegar programaticamente até aos recursos da Internet para recuperar informações.
C. Configurar as tabelas externas do Unity Catalog para armazenar em cache o conteúdo dos recursos da internet e também os resultados da pesquisa para o acesso offline pelo agente.
D. Configurar o servidor MCP gerido através da configuração do servidor MCP do agent, especificando o tipo de servidor como "gerido" e fornecendo o identificador do servidor do recurso da Internet.
E. Implementar o servidor MCP externo fornecendo os seus detalhes de ligação, armazenando a API key nos Databricks Secrets e referenciando-a na configuração do servidor MCP.

Questão 10

Objetivo: Incorporar o feedback dos SMEs para melhorar o desempenho do agent.

Um Engenheiro de Generative AI é responsável por avaliar um assistente RAG de apoio ao cliente utilizado internamente pelas equipes de operações. Quatro especialistas da área analisam semanalmente as respostas amostradas no MLflow, utilizando dimensões como a precisão factual, a completude e a utilidade. Após várias rondas, o engenheiro percebe que as avaliações dos peritos variam muito para as mesmas respostas, tornando os dados de avaliação pouco fiáveis para acompanhar as melhorias do modelo ao longo do tempo. O engenheiro precisa de criar um processo de avaliação fiável que reduza a inconsistência dos avaliadores e, ao mesmo tempo, apoie melhorias iterativas de qualidade.

O que deve fazer o engenheiro?

A. Utilizar um LLM-as-a-judge para reavaliar as respostas passadas e futuras e tratar as avaliações geradas pelo modelo como a principal fonte de verdade, em vez de conciliar a discordância entre especialistas.
B. Definir rubricas claras, calibrar os SMEs com base nos critérios e utilizar os julgamentos alinhados em `mlflow.genai.evaluate` () para uma avaliação consistente do agent.
C. Calcular a média das pontuações de todos os peritos do domínio para cada resposta e utilizar a pontuação combinada diretamente como o benchmark definitivo para o ajuste do modelo.
D. Construir o benchmark apenas com base nas respostas em que todas as especialistas já concordam e exclua os casos contestados do conjunto de avaliação para melhorar a consistência.

Respostas

Questão 1: A, B

Questão 2: B

Questão 3: C

Questão 4: A

Questão 5: D

Questão 6: C

Questão 7: B

Questão 8: A

Questão 9: D, E

Questão 10: B