databricks

Databricks Exam Guide

Databricks Certified Data Engineer Associate



[Fornecer Feedback sobre o Guia do Exame]

Atualizar Informação

Atualizações	Data
Exame publicado	4 de maio de 2026
Guia do exame atualizado	6 de agosto de 2026

Verifique o guia do exame em nossas páginas da web duas semanas antes da data do seu exame para garantir que você está se preparando para a versão atual do exame.

Objetivo deste Guia de Exame

O objetivo deste guia de exame é fornecer uma visão geral do exame e do que ele abrange, para ajudá-lo a determinar sua prontidão para o exame. Este documento é atualizado sempre que há alterações no exame (e indica quando essas alterações entram em vigor) para que você possa se preparar com confiança. Este guia abrange a versão do exame que está ativa a partir da data de vigência listada acima.

Descrição do Público-Alvo

O exame de certificação Databricks Certified Data Engineer Associate avalia a capacidade do indivíduo de usar o Databricks Platform para realizar tarefas fundamentais de engenharia de dados. Isso inclui a compreensão do Databricks Platform, seu workspace, arquitetura e capacidades, juntamente com as tarefas relacionadas à ingestão de dados, carregamento de dados e transformação e modelagem de dados, como a capacidade de realizar tarefas de Extract, Transform, Load (ETL) usando PySpark, trabalhando com Lakeflow Jobs e CI/CD. O exame também avalia a compreensão de técnicas de solução de problemas, monitoramento e otimização, bem como o conhecimento de como alcançar governança e segurança dentro do Databricks Platform.

Sobre o Exame

Item	Detalhe
Número de questões pontuadas	45 questões pontuadas de múltipla escolha
Possíveis questões não pontuadas	Podem ser incluídas até 10 questões não pontuadas (consulte "Conteúdo não pontuado" abaixo)
Número total de questões que podem aparecer	45 questões pontuadas, além das questões não pontuadas
Limite de tempo	90 minutos (o tempo adicional para as questões não pontuadas já está incluído)
Taxa de inscrição	US$ 200, mais impostos aplicáveis, conforme exigido pela legislação local
Forma de aplicação	On-line com supervisão ou em centro de testes
Idiomas disponíveis	Inglês; japonês; português brasileiro (pt-BR); coreano
Materiais de consulta	Não permitidos
Pré-requisito	Não há pré-requisito obrigatório, mas o treinamento relacionado é altamente recomendado
Experiência recomendada	É altamente recomendável ter seis meses de experiência prática na plataforma Databricks, alinhada à estrutura do exame listada
Validade	2 anos

Recertificação: A recertificação é necessária a cada dois anos para manter seu estado de certificado. Para se recertificar, você pode fazer o Data Engineer Associate Renewal Exam ou o exame completo atual.

Conteúdo não pontuado: Os exames podem incluir itens não pontuados para coletar informação estatística para uso futuro. Esses itens não são identificados no exame e não afetam sua pontuação, e o tempo adicional é considerado no Limite de tempo para levá-los em conta.

Preparação Recomendada

Treinamento

- **Conduzido por instrutor: Engenharia de Dados com Databricks**
- **No seu próprio ritmo (disponível na Databricks Academy):**

- Ingestão de Dados com LakeFlow Connect
- Implantar cargas de trabalho com Lakeflow Jobs
- DevOps Essentials para Engenharia de Dados
- Construa pipelines de dados com Apache Spark™ Declarative Pipelines
- Introdução à Governança

Guia de Preparação para IA - Qualquer Exame Databricks

O Databricks AI Prep Guide mostra como transformar o chatbot de IA que você já usa em um tutor personalizado para qualquer certificação Databricks: configure-o com o guia oficial do exame, identifique as respostas desatualizadas que a IA tende a fornecer, descubra seus pontos fracos e combine tudo isso com a prática hands-on que todo exame Databricks exige.

Prática Hands-On

Não sabe o que praticar? O Guia de Preparação de IA (consulte a seção Guia de Preparação de IA acima) inclui um prompt em "o passo 5 — Mínimo prático" que faz um chatbot de IA gerar uma lista de verificação personalizada com 6–10 tarefas práticas mapeadas para os objetivos deste exame, cada uma pequena o suficiente para ser concluída no Databricks Free Edition.

Documentação da Databricks

Ao se preparar, estude a documentação da Databricks oficial para as áreas da plataforma abordadas neste exame. Ela é a fonte mais atual e autoritativa. Use-a para julgar o quanto estudar cada tópico e para garantir que você está aprendendo recursos atualizados e não obsoletos, em vez de material desatualizado.

Documentação oficial: documentação da Databricks

Estrutura do Exame

Seção 1: Databricks Intelligence Platform (6%)

- Compreender os componentes principais da Databricks Platform, como sua arquitetura, Delta Lake e Unity Catalog.
- Compreender os serviços de compute da Databricks Platform, incluindo suas características, limitações e modelos de custo, e selecionar a opção mais adequada para cada caso de uso de workload.

Seção 2: Ingestão de Dados e Carregamento (21%)

- Habilitar e detalhar padrões de ingestão de dados, incluindo lotes, transmissão e carregamento incremental, e importar dados de fontes como arquivos locais, conectores padrão do LakeFlow Connect e conectores gerenciados do LakeFlow Connect.
- Use o comando COPY INTO para carregar incrementalmente arquivos do cloud object storage (ADLS/S3/GCS) em tabelas gerenciadas pelo Unity Catalog.
- Use o Auto Loader com imposição de esquema e evolução do esquema nos modos de lotes (por exemplo, listagem de diretório ou notificação de arquivo) para carregar dados em tabelas governadas pelo Unity Catalog.
- Configurar o Lakeflow Connect para ingerir dados de forma confiável a partir de diversas fontes empresariais em tabelas gerenciadas pelo Unity Catalog.
- Use clientes JDBC/ODBC ou REST em Notebooks para carregar dados no armazenamento cloud ou diretamente em tabelas governadas pelo Unity Catalog, geralmente orquestrado e agendado com Lakeflow Jobs.
- Priorize entre Auto Loader, Lakeflow Connect (conectores padrão e gerenciados), conectores de parceiros e outros métodos de ingestão com base em requisitos técnicos como volume de dados, frequência de ingestão, tipos de dados e necessidades de governança com Unity Catalog.
- Ingerir dados semiestruturados e não estruturados (por exemplo, JSON e dados aninhados) via LakeFlow Connect e outros conectores gerenciados em tabelas Delta governadas pelo Unity Catalog.

Seção 3: Transformação e Modelagem de Dados (22%)

- Implemente a limpeza de dados lendo tabelas Bronze com PySpark/SQL, tratando valores nulos, padronizando tipos de dados e gravando em novas tabelas Prata.
- Combine DataFrames com operações como Inner join, left join, broadcast join, múltiplas chaves, cross join, union e union all.
- Manipule colunas, linhas e estruturas de tabela adicionando, removendo, dividindo, renomeando nomes de colunas, aplicando filtros e expandindo arrays.
- Realizar operações de deduplicação de dados e operações de agregação em DataFrames, como count, approximate count distinct e mean, summary.
- Entenda os parâmetros básicos de ajuste (spark.sql.shuffle.partitions, spark.default.parallelism, spark.executor/driver.memory, spark.sql.autoBroadcastJoinThreshold) e meça novamente o desempenho.
- Entender a diferença entre objetos da camada ouro, como Materialized Views, views, Streaming Tables e tabelas para equipes de BI e analítica no Unity Catalog, e saber como criá-los.
- Aplicar verificações de qualidade de dados e regras de validação para garantir conjuntos de dados Prata e ouro confiáveis.

Seção 4: Trabalhando com Lakeflow Jobs (16%)

- Implementar fluxos de controle (tentativas de repetição e tarefas condicionais como ramificação e loop) usando Lakeflow Jobs para orquestração de pipeline
- Configurar tarefas comuns (notebook, SQL query, dashboard e tarefas de pipeline) e suas dependências usando Lakeflow Jobs e seu gráfico de tarefas baseado em DAG
- Implementar agendamentos de job usando Lakeflow Jobs com compreensão dos tipos de gatilho (agendado, chegada de arquivo e atualização de tabela)
- Escolha entre gatilhos baseados em tempo e data-driven com base na disponibilidade dos dados e nas dependências do pipeline.

Seção 5: Implementing CI/CD (10%)

- Gerencie seu fluxo de trabalho de desenvolvimento de código na interface do Databricks workspace, incluindo criação e alternância entre branches no Databricks Git Folders (anteriormente Databricks Repos), confirmação e envio de alterações, e criação de pull requests usando a integração Git do Databricks.
- Entender a configuração específica por ambiente usando variáveis e substituições do Declarative Automation Bundle (anteriormente Databricks Asset Bundles) ao promover a mesma base de código entre os ambientes dev, test e prod.
- Implante Declarative Automation Bundles (anteriormente Databricks Asset Bundles) para empacotar, configurar e promover Lakeflow Jobs, Lakeflow Declarative Pipelines e outros ativos do workspace nos ambientes de dev, test e prod.
- Entender o Databricks CLI para validar, implantar e gerenciar Declarative Automation Bundles (anteriormente Databricks Asset Bundles) e outros ativos de workspace em Workflows de CI/CD automatizados.

Seção 6: Solução de Problemas, Monitoramento e Otimização (10%)

- Identifique tendências no desempenho de jobs usando o Lakeflow Jobs run history view para comparar os tempos de execução atuais com as linhas de base históricas.
- Use a interface do Lakeflow Jobs para monitorar a integridade do Pipeline interpretando os status do job, visualizando grafos de tarefas baseados em DAG para identificar bloqueadores upstream e fazendo o acompanhamento dos tempos de execução e das taxas de falha do Pipeline.
- Identifique gargalos de desempenho comuns, como distorção de dados, shuffling e disk spilling, interpretando as métricas de nível de estágio no Spark UI.
- Entenda os recursos do liquid clustering e do Predictive Optimization.
- Diagnosticar falhas de startup de cluster, conflitos de biblioteca e problemas de falta de memória.

Seção 7: Governança e Segurança (15%)

- Diferenciar entre tabelas gerenciadas e externas no Unity Catalog e realizar operações básicas (criar, modificar, excluir e converter entre tabelas gerenciadas e externas) nelas.
- Configurar controles de acesso usando a UI e SQL, aplicando privilégios GRANT, REVOKE e DENY a entidades de segurança (usuários, grupos e service principals) nos níveis apropriados da hierarquia de segurança.
- Entender o mascaramento em nível de coluna e a segurança em nível de linha para restringir a visibilidade dos dados com base em grupos de usuários.
- Compreender as políticas ABAC do Unity Catalog para controlar centralmente a filtragem em nível de linha e o mascaramento de colunas para dados sensíveis.

Perguntas de Exemplo

O objetivo das questões de exemplo é mostrar a você os objetivos conforme são descritos no guia do exame e fornecer uma questão de exemplo alinhada a cada objetivo. O guia do exame lista os objetivos que podem ser abordados em um exame; a melhor forma de se preparar é revisar o esboço completo do exame para a sua data de exame acima.

Questão 1

Objetivo: Identificar gargalos de desempenho comuns, como distorção de dados, shuffling e disk spilling, interpretando as métricas de nível de estágio no Spark UI.

Um engenheiro de dados percebe que a duração de um batch job dobrou após a integração de uma nova fonte de dados. O job realiza uma agregação groupBy ampla em uma key customer_id. No Spark UI, o estágio mais longo mostra que a maioria das tarefas é concluída em menos de 30 segundos, mas uma tarefa leva mais de 10 minutos; o resumo de tarefas do estágio mostra leitura aleatória Min/Mediana próxima de 400 MB, enquanto a leitura aleatória Máx excede 5 GB. O Adaptive Query Execution já está habilitado com as configurações default, mas a tarefa distorcida persiste.

Qual solução reduz o tempo de execução do job?

A. Aumentar o tamanho do cluster para adicionar mais executores para que a tarefa lenta termine mais rapidamente
B. Depender do tratamento de join com skew do AQE para dividir automaticamente a partição superdimensionada em tempo de execução
C. Reduza spark.sql.shuffle.partitions para consolidar mais trabalho em menos tarefas
D. Adicionar uma chave salt à chave de agregação para que o valor distorcido seja distribuído entre múltiplas tarefas

Questão 2

Objetivo: Compreender os componentes principais da Databricks Platform, como sua arquitetura, Delta Lake e Unity Catalog.

Um engenheiro de dados precisa de iteração rápida de pipeline mantendo rollbacks confiáveis após ingestões incorretas, garantindo trilhas de auditoria para compliance regulatório e fornecendo acesso consistente a uma única fonte da verdade para cargas de trabalho de IA e BI.

Qual estratégia o engenheiro de dados deve usar para atender a esses requisitos?

A. Armazenamento DBFS CSV com versionamento manual de arquivos e cópias noturnas para reversão.
B. Transações ACID do Delta Lake e viagem do tempo, governadas pelo Unity Catalog para acesso consistente e linhagem.
C. Somente armazenamento de objetos cloud, com consultas SQL ad hoc para recuperação e governança.
D. DataFrames efêmeros em memória para trilhas de auditoria e distribuição de BI.

Questão 3

Objetivo: Habilitar e detalhar padrões de ingestão de dados, incluindo lotes, transmissão e carregamento incremental, e importar dados de fontes como arquivos locais, conectores padrão do LakeFlow Connect e conectores gerenciados do LakeFlow Connect.

Um engenheiro de dados está construindo pipelines downstream para consumir logs de auditoria do Databricks a partir de um bucket S3 de propriedade do cliente. Antes de implementar a inferência de esquema e o checkpointing, ele deseja entender o formato de entrega, a latência de ingestão típica e se os arquivos podem ser sobrescritos.

Qual é o comportamento de armazenamento do audit log do Databricks?

A. Os arquivos são entregues como JSON com registro de eventos típico em menos de 15 minutos após o início da entrega, e novas entregas podem sobrescrever arquivos existentes
B. Os arquivos são entregues como CSV com garantias de latência abaixo de um minuto, e substituições nunca ocorrem após a gravação de um arquivo para preservar a imutabilidade
C. Os arquivos são entregues como Parquet com consistência eventual além de 24 horas, e as substituições são desabilitadas para simplificar a ingestão de transmissão
D. Os arquivos são entregues como JSON com entrega em uma cadência de lotes semanal, e as substituições sobrescrevem o conteúdo anterior completamente sem acrescentar dados

Questão 4

Objetivo: Compreender os serviços de compute da Databricks Platform, incluindo suas características, limitações e modelos de custo, e selecionar a opção mais adequada para cada caso de uso de workload.

Uma equipe de engenharia de dados oferece suporte a vários analistas de negócios que executam queries SQL ad hoc ao longo do dia em tabelas Delta organizadas. A equipe precisa garantir um bom desempenho de query, startup rápido do cluster e suporte a múltiplos usuários simultâneos, ao mesmo tempo em que gerencia os custos evitando o dimensionamento desnecessário para clusters muito grandes.

Qual configuração de cluster atende a esses requisitos?

A. Um job cluster com autoscaling projetado para workflows ETL agendados
B. Um cluster de uso geral configurado com um número fixo de worker nodes
C. Um Serverless SQL Warehouse com escalonamento automático habilitado
D. Um cluster de nó único configurado para tarefas leves de desenvolvimento

Questão 5

Objetivo: Implantar Declarative Automation Bundles (anteriormente Databricks Asset Bundles) para empacotar, configurar e promover Lakeflow Jobs, Lakeflow Declarative Pipelines e outros ativos de workspace nos ambientes de dev, test e prod.

Uma equipe deseja uma forma modular de implantar, versionar e orquestrar pipelines de ETL no Databricks, habilitando CI/CD e repetibilidade.

Qual recurso oferece suporte a esse requisito?

A. Use modelos no Unity Catalog para representar jobs de ETL, onde cada modelo armazena o artefato de código do pipeline e o CI/CD promove versões atualizando aliases de modelos vinculados a tarefas do job.
B. Empacote a lógica de transformação como bibliotecas wheel armazenadas em Unity Catalog Volumes e vincule-as às tarefas do Lakeflow Jobs para garantir uma implantação determinística entre ambientes.
C. Encapsule a lógica da API em um notebook montado em Volume e use a Jobs API v2 para acionar o notebook, dependendo da história de revisões do notebook como sistema de versionamento.
D. Use Declarative Automation Bundles (DABs) para definir recursos e ativos de código, versioná-los no Git e promover implantações entre ambientes por meio de ações automatizadas de CI/CD.

Questão 6

Objetivo: Entender o uso do Databricks CLI para validar, implantar e gerenciar DABs e outros ativos do workspace em um fluxo de trabalho automatizado de CI/CD

Um engenheiro de dados está usando Declarative Automation Bundles para implantar um job que se conecta a um banco de dados externo. Para seguir as melhores práticas de segurança, a senha do banco de dados não deve ser codificada diretamente no arquivo databricks.yml nem no código-fonte do notebook.

Como o engenheiro de dados deve configurar o bundle para lidar com segurança com esse segredo durante uma implantação automatizada?

A. Defina uma variável no bloco de variáveis do databricks.yml e forneça seu valor por meio de uma variável de ambiente com o prefixo BUNDLE_VAR_
B. Armazene a senha em um arquivo de texto simples chamado passwords.txt na raiz do bundle e inclua-o na implantação
C. Use o comando Databricks bundle encrypt para ofuscar a string antes de fazer o commit no Git
D. Use uma configuração padrão do Spark spark.conf.set("db_pass", "...") diretamente dentro do notebook.

Questão 7

Objetivo: Aplicar mascaramento em nível de coluna e segurança em nível de linha para restringir a visibilidade dos dados com base no grupo do usuário.

Um engenheiro de dados precisa garantir que apenas os membros do grupo 'finance_team' possam ver o valor completo da coluna 'credit_card_number'. Todos os outros usuários devem ver uma versão mascarada.

Qual recurso e configuração do Unity Catalog devem ser usados para implementar esse requisito?

A. Criar uma Tabela Delta separada para cada grupo de usuários com valores redigidos
B. Habilitar funções IAM da AWS na coluna específica dentro das propriedades da Tabela Delta
C. Aplicar uma função de mascaramento de coluna usando a lógica is_account_group_member('finance_team')
D. Definir uma permissão DENY na coluna específica para todos os usuários, exceto a equipe de finanças

Questão 8

Objetivo: Diferenciar entre tabelas gerenciadas e tabelas externas no Unity Catalog e realizar operações básicas.

Um engenheiro de dados converte uma tabela externa em uma tabela gerenciada. Logo após, a equipe descobre um problema e precisa desfazer a conversão.

Qual ação restaura a Tabela ao seu tipo anterior em runtimes mais recentes?

A. Execute ALTER TABLE ... UNSET MANAGED para reverter para externo
B. Remova a tabela gerenciada e recrie-a como externa usando o LOCATION original, pois rollback não é suportado
C. Execute ALTER TABLE ... SET EXTERNAL para alterar o tipo da tabela de volta sem mover os dados
D. Restaure a tabela a partir da história de viagem do tempo do Delta usando RESTORE TABLE ... TO VERSION AS OF

Questão 9

Objetivo: Usar a UI do Lakeflow Jobs para monitorar a integridade do pipeline interpretando os status do job e visualizando os grafos de tarefas

Uma tarefa de job do Databricks falha com o estado 'Timed Out' após ser executada por 45 minutos. O engenheiro de dados determina que a tarefa pode levar até 60 minutos para ser concluída com êxito devido ao aumento no volume de dados a ser processado e deseja que execuções futuras terminem sem serem encerradas automaticamente.

Qual configuração o engenheiro de dados deve modificar para atingir o objetivo?

A. A condição de dependência na tarefa upstream
B. O limite para a duração da execução de tarefas
C. A configuração de execuções concorrentes máximas no job
D. A configuração de encerramento automático do cluster no Recurso de Compute

Questão 10

Objetivo: Implementar schedules de job usando Lakeflow Jobs com compreensão dos tipos de Gatilho

Uma equipe de engenharia de dados recebe arquivos de exportação de clickstream de um fornecedor terceiro de analítica. O fornecedor faz upload de arquivos JSON compactados em um local de armazenamento cloud externo em intervalos imprevisíveis — às vezes vários arquivos em uma hora, às vezes nada durante a maior parte de um dia. Um Lakeflow Job executa um notebook de lotes que lê novos arquivos com o Auto Loader, os enriquece e mescla os resultados em uma Delta Table. A equipe quer que cada novo conjunto de arquivos seja processado logo após chegar, sem que operadores fiquem fazendo monitoramento do local de armazenamento, e sem incorrer em custos de compute durante longos períodos parados ou iniciar execuções que não encontrem dados novos.

Qual tipo de Gatilho deve ser configurado para este Lakeflow Job?

A. Gatilho agendado a cada 5 minutos para verificar novos arquivos de exportação
B. Gatilho de chegada de arquivo no local de armazenamento externo do fornecedor
C. Gatilho contínuo para manter o job sempre em execução
D. Gatilho manual iniciado por um operador sempre que uma notificação de upload de fornecedor chega

Gabarito & Explicações

#	Resposta	Explicação
1	D	Um valor de Max shuffle read de 5 GB contra uma mediana de ~400 MB é um caso clássico de skew de chave (key skew) — um único customer_id com volume desproporcional domina uma única partição. O tratamento de skew do AQE otimiza apenas joins com distorção (sort-merge/shuffle-hash), e não agregações groupBy; portanto, não alivia esse estágio (stage) mesmo com o AQE ativado, o que elimina a opção B. Adicionar executores mantém a única task gigante em execução isolada (A), e reduzir shuffle.partitions torna as partições ainda maiores, agravando o skew (C). Aplicar a técnica de salting na chave de agregação distribui a chave de alto volume entre várias tasks, seguido por uma segunda etapa para consolidar os resultados parciais.
2	B	O Delta Lake fornece transações ACID e time travel, permitindo rollback confiável após uma ingestão com erro e um histórico completo de auditoria, enquanto o Unity Catalog governa o acesso e a linhagem para que as cargas de trabalho de IA e BI leiam uma única fonte da verdade. As outras opções dependem de cópias manuais ou armazenamento não governado.
3	A	O Databricks entrega logs de auditoria como arquivos JSON, normalmente em até 15 minutos após o evento assim que a entrega se inicia, e uma entrega posterior pode sobrescrever um arquivo existente. Portanto, os pipelines de ingestão devem tolerar sobrescritas em vez de presumir que os arquivos sejam imutáveis.
4	C	Um SQL warehouse serverless com escalonamento automático (autoscaling) foi desenvolvido para atender a vários usuários interativos de SQL simultâneos, garantindo inicialização rápida e isolamento de queries enquanto escala apenas conforme a demanda exige. Clusters de job atendem a ETLs agendados, e clusters fixos ou de nó único (single-node) não conseguem equilibrar concorrência e custo.
5	D	Os Declarative Automation Bundles declaram jobs, pipelines e ativos de código como configurações versionadas no Git, garantindo que a mesma definição seja promovida entre os ambientes de dev, test e prod por meio de CI/CD automatizado. As outras opções dependem de repositórios de artefatos ou do histórico de revisões de notebooks, nenhum dos quais versiona a orquestração em si.
6	A	Os DABs oferecem suporte à parametrização por meio de um bloco variables. Ao usar o prefixo BUNDLE_VAR_<variable_name>, a CLI injeta automaticamente o valor do ambiente (por exemplo, do GitHub Secrets) no bundle no momento da implantação, garantindo que o segredo (secret) nunca toque o código-fonte nem o arquivo YAML.
7	C	O mascaramento de coluna (column masking) usando a função is_account_group_member() é a escolha correta porque permite segurança refinada e dinâmica (dynamic fine-grained security), em que a mesma tabela pode ser consultada por diferentes usuários enquanto oculta seletivamente dados confidenciais com base na associação a grupos.
8	A	UNSET MANAGED é o rollback documentado. Ele atualiza os metadados da tabela para apontar de volta ao local externo (external location) original e está disponível apenas

		dentro de uma janela de 14 dias após a conversão. Gravações feitas no local gerenciado após a conversão são preservadas; commits entre a conversão e o rollback oferecem suporte a time travel por versão, mas não por timestamp. O Databricks exclui os dados no local gerenciado sete dias após o rollback, e jobs de streaming precisam ser reiniciados.
9	B	O status "Timed Out" indica que a execução excedeu o limite de tempo limite (timeout) configurado na task — neste caso, 45 minutos. Aumentá-lo para mais de 60 minutos (com margem de segurança, por exemplo, 90 minutos) permite que as próximas execuções sejam concluídas em vez de serem canceladas.
10	B	Gatilhos de chegada de arquivo (file arrival triggers) monitoram um local externo e iniciam uma execução apenas quando novos arquivos aparecem — de forma automatizada, com baixa latência e sem consumo de computação enquanto o local permanecer inalterado.