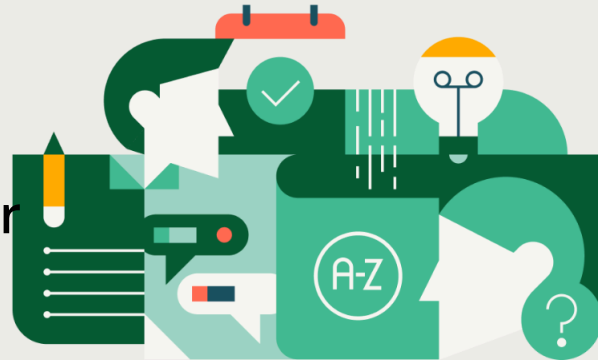
databricks

Databricks Guia do exame

Databricks Certified Data Engineer Associate



[Fornecer feedback sobre o guia de exame]

Finalidade deste Guia do Exame

O objetivo deste guia de exame é fornecer uma visão geral do exame e o que é abordado no exame 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 o exame atualmente ativo a partir de 25 de julho de 2025. Volte duas semanas antes de fazer o exame para ter certeza de que você tem a versão mais atual.**

Descrição do público

O exame de certificação Databricks Certified Data Engineer Associate avalia a capacidade de um indivíduo de usar o Databricks Data Intelligence Platform para concluir tarefas introdutórias de engenharia de dados. Isso inclui uma compreensão da Data Intelligence Platform e seu espaço de trabalho, sua arquitetura e seus recursos. IT também avalia a capacidade de executar tarefas ETL usando Apache Spark SQL ou PySpark, abrangendo extração, manipulação de dados complexos e funções definidas pelo usuário. Por fim, o exame avalia a capacidade do testador de implantar e orquestrar cargas de trabalho com Databricks Workflows configurando e agendando trabalhos de forma eficaz.
Espera-se que as pessoas aprovadas neste exame de certificação concluam tarefas básicas de engenharia de dados usando Databricks e suas ferramentas associadas.

Sobre o Exame

- Número de itens pontuados: 45 questões de múltipla escolha
- Prazo: 90 minutos
- Taxa de inscrição: US$ 200, mais impostos aplicáveis, conforme exigido pela legislação local
- Método de entrega: Supervisionado online
- Auxiliares de teste: nenhum é permitido.
- Pré-requisito: Nenhum exigido; participação no curso e seis meses de experiência prática em Databricks são altamente recomendados
- Validade: 2 anos
- Recertificação: A recertificação é necessária a cada dois anos para manter seu status certificado. Para se recertificar, é preciso fazer o exame completo que está no ar. Consulte

a seção "Preparando-se para o exame" na página do exame para se preparar para fazer o exame novamente.
- Conteúdo sem pontuação: Os exames podem incluir itens não pontuados para coletar informações estatísticas para uso futuro. Esses itens não são identificados no formulário e não impactam sua pontuação. O tempo adicional é considerado para esse conteúdo.

Treinamento recomendado

- Conduzido por instrutor: Engenharia de Dados com Databricks
- Auto-cadenciado (disponível em Databricks Academy):
    - Ingestão de dados com o Lakeflow Connect
    - Implantar cargas de trabalho com LakeFlow Jobs
    - Criar pipelines de dados com o Lakeflow Declarative pipeline
    - Fundamentos de Devops para Engenharia de Dados

Resumo do exame

Secção 1: Databricks Intelligence Platform

- Habilite recursos que simplificam as decisões de layout de dados e otimizam o desempenho da consulta.
- Explicar o valor da Plataforma de Inteligência de Dados.
- Identifique a computação aplicável a ser usada para um caso de uso específico.

Seção 2: Desenvolvimento e Ingestão

- Usar Databricks Connect em um fluxo de trabalho de engenharia de dados
- Determinar os recursos da funcionalidade de Notebooks
- Classificar fontes Auto Loader válidas e casos de uso
- Demonstrar conhecimento da sintaxe Auto Loader
- Use as ferramentas de depuração integradas do Databricks para solucionar um determinado problema

Seção 3: Processamento de Dados e Transformações

- Descreva as três camadas da Arquitetura Medallion e explique a finalidade de cada camada em um pipeline de processamento de dados.
- Classifique o tipo de cluster e a configuração para obter o desempenho ideal com base no cenário em que o cluster é usado.
- Enfatize as vantagens do DLT (para o processo ETL em Databricks).
- Implemente pipelines de dados usando DLT..
- Identifique recursos DDL (Data Definition Language)/DML.
- Calcule agregações e métricas complexas com PySpark DataFrames.

Secção 4: Pipelines de dados para produção

- Identifique a diferença entre o DAB e os métodos de implantação tradicionais.
- Identifique a estrutura dos Asset Bundles.
- Implante um fluxo de trabalho, repare e execute novamente uma tarefa em caso de falha.
- Use serverless para uma computação prática e otimizada automaticamente gerenciada por Databricks.
- Analisando o Spark UI para otimizar a consulta.

Seção 5: Governança de dados e qualidade

- Explicar a diferença entre tabelas gerenciadas e externas.
- Identifique a concessão de permissões a usuários e grupos no UC.
- Identifique papéis-chave no UC.
- Identifique como os logs de auditoria são armazenados.
- Use recursos de linhagem em Unity Catalog.
- Use o recurso Delta Sharing disponível com Unity Catalog para compartilhar dados.
- Identifique as vantagens e limitações do Delta Sharing.
- Identifique os tipos de Delta Sharing - Databricks vs sistema externo.
- Analisar as considerações de custo do compartilhamento de dados entre nuvens
- Identifique casos de uso de Lakehouse Federation quando conectado a fontes externas.

Exemplos de perguntas

Essas perguntas foram retiradas de uma versão anterior do exame. O objetivo é mostrar os objetivos como eles estão declarados no guia do exame e fornecer um exemplo de pergunta que se alinhe ao objetivo. O guia do exame lista os objetivos que podem ser abordados em um exame. A melhor maneira de se preparar para um exame de certificação é revisar o resumo do exame no guia do exame.

Pergunta 1

Um engenheiro de dados criou uma tabela Delta como parte de uma pipeline de dados. Os analistas de dados downstream agora precisam da permissão SELECT na tabela Delta.

Qual parte da Plataforma Databricks Lakehouse o engenheiro de dados pode usar para conceder aos analistas de dados o acesso apropriado?

A. Jobs
B. Dashboards
C. Data Explorer
D. Repos

Pergunta 2

Um conjunto de dados foi definido usando Delta Live Tables e inclui uma cláusula de expectativas:

CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01')

Qual é o comportamento esperado quando um lote de dados contendo dados que violam essas restrições é processado?

A. Os registros que violam a expectativa são descartados do dataset de destino e registrados como inválidos no log de eventos.
B. Os registros que violam a expectativa são adicionados ao dataset de destino e registrados como inválidos no log de eventos.
C. Registros que violam a expectativa fazem com que o trabalho falhe.
D. Os registros que violam a expectativa são adicionados ao dataset de destino e marcados como inválidos em um campo adicionado ao dataset de destino.

Pergunta 3

Uma Delta Live Table Pipeline inclui dois conjuntos de dados definidos usando **STREAMING LIVE TABLE.** Três conjuntos de dados são definidos em relação a fontes de tabela Delta Lake usando **LIVE TABLE**.

A tabela está configurada para ser executada no modo de desenvolvimento usando o Triggered Pipeline Mode.

Considerando que os dados não processados anteriormente existem e todas as definições são válidas, qual é o resultado esperado depois de clicar em Iniciar para atualizar o Pipeline?

A. Todos os conjuntos de dados serão atualizados em intervalos definidos até que o pipeline seja desligado. Os recursos de computação persistirão depois que o pipeline for interrompido para permitir testes adicionais.
B. Todos os conjuntos de dados serão atualizados uma vez e o pipeline será encerrado. Os recursos de computação serão encerrados.
C. Todos os conjuntos de dados serão atualizados em intervalos definidos até que o pipeline seja desligado. Os recursos de computação serão implantados para a atualização e encerrados quando o pipeline for interrompido.
D. Todos os conjuntos de dados serão atualizados uma vez e o pipeline será encerrado. Os recursos de computação persistirão para permitir testes adicionais.

Pergunta 4

Uma Delta Live Table Pipeline inclui dois conjuntos de dados definidos usando STREAMING LIVE TABLE. Três conjuntos de dados são definidos em relação a fontes de tabela Delta Lake usando LIVE TABLE.

A tabela é configurada para ser executada no modo de desenvolvimento usando o modo Pipeline Contínuo.

Supondo que os dados não processados anteriormente existam e todas as definições sejam válidas, qual é o resultado esperado depois de clicar em Iniciar para atualizar o pipeline?

A. Todos os conjuntos de dados serão atualizados em intervalos definidos até que o pipeline seja desligado. Os recursos de computação persistirão até que o pipeline seja desligado.
B. Todos os conjuntos de dados serão atualizados uma vez e o pipeline será encerrado. Os recursos de computação persistirão para permitir testes adicionais.
C. Todos os conjuntos de dados serão atualizados uma vez e o pipeline será encerrado. Os recursos de computação serão encerrados.
D. Todos os conjuntos de dados serão atualizados em intervalos definidos até que o pipeline seja desligado. Os recursos de computação persistirão para permitir testes adicionais.

Pergunta 5

Uma nova equipe **de engenharia de** dados foi designada para um projeto ELT. A nova equipe de engenharia de dados precisará de privilégios totais na tabela **sales** para gerenciar totalmente o projeto.

Qual comando pode ser usado para conceder permissões completas no banco de dados à nova equipe de engenharia de dados?

A. `GRANT SELECT ON TABLE sales TO team;`
B. `GRANT USAGE ON TABLE sales TO team;`
C. `GRANT ALL PRIVILEGES ON TABLE team TO sales;`
D. `GRANT ALL PRIVILEGES ON TABLE sales TO team;`

Respostas

Pergunta 1: C

Pergunta 2: B

Pergunta 3: D

Pergunta 4: A

Pergunta 5: D