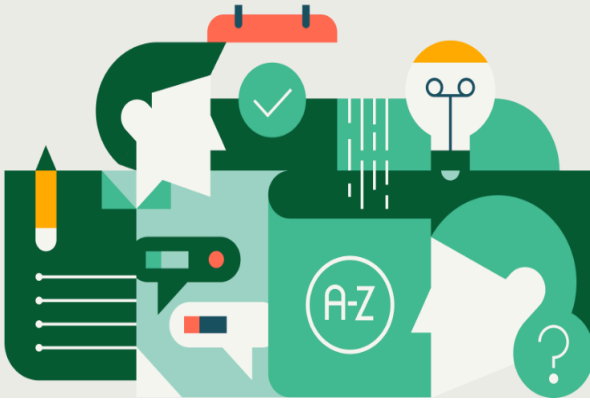
databricks

Guia de Exame Databricks

Databricks Certified Data Engineer Professional



Enviar feedback sobre o guia do exame

Objetivo deste guia do exame

Este guia do exame oferece uma visão geral do exame Databricks Certified Data Engineer Professional e do que ele abrange, para ajudá-lo a determinar sua prontidão 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 se preparar. **Verifique novamente duas semanas antes do seu exame para garantir que você tenha a versão mais atual. Esta versão abrange a versão atualmente em vigor em 30 de novembro de 2025.**

Descrição do público

O exame Databricks Certified Data Engineering Professional valida as habilidades avançadas do candidato na criação, otimização e manutenção de soluções de engenharia de dados de nível de produção na Databricks Data Intelligence Platform. Os candidatos aprovados demonstram expertise nos principais recursos da plataforma, como Delta Lake, Unity Catalog, Auto Loader, Lakeflow Spark Declarative Pipelines, Databricks Compute (incluindo serverless), Lakeflow Jobs e a Medallion Architecture. Esta certificação avalia a capacidade de projetar pipelines de ETL seguros, confiáveis e econômicos, processar dados complexos de diversas fontes usando Python e SQL, e aplicar as melhores práticas em gerenciamento de esquema, observabilidade, governança e otimização de desempenho. Os candidatos também são avaliados quanto à implementação de cargas de trabalho de streaming, à orquestração de fluxos de trabalho, ao uso de DevOps e CI/CD, e à implantação com ferramentas como a Databricks CLI, a REST API e os Asset Bundles. Espera-se que os indivíduos aprovados neste exame de certificação sejam capazes de concluir tarefas avançadas de engenharia de dados usando o Databricks e suas ferramentas associadas.

Sobre o exame

- Número de itens: 59 questões de múltipla escolha pontuadas
- Tempo limite: 120 minutos
- Taxa de inscrição: USD 200, mais os impostos aplicáveis conforme exigido pela legislação local
- Método de aplicação: Supervisionado on-line e supervisionado em centro de testes

- Materiais de apoio: Nenhum material de apoio é fornecido, incluindo documentação de API.
- Pré-requisito: Nenhum é obrigatório; recomenda-se fortemente a participação em cursos relacionados e um ano de experiência prática nas tarefas de engenharia de dados descritas no guia do exame.
- Validade: 2 anos
- Recertificação: A recertificação é necessária a cada dois anos para manter seu status de certificado. Para se recertificar, você deve fazer a versão atual do exame. Consulte a seção "Preparando-se para o exame" abaixo para se preparar para o seu exame de recertificação.
- Conteúdo não pontuado: Os exames podem incluir itens não pontuados para coletar informações estatísticas para uso futuro. Esses itens não são identificados no formulário e não afetam sua pontuação, e um tempo adicional é considerado para este conteúdo.

Treinamento recomendado

- Conduzido por instrutor: Advanced Data Engineering With Databricks
- Autoguiado (disponível na Databricks Academy):
    - Databricks Streaming and Lakeflow Spark Declarative Pipelines
    - Databricks Data Privacy
    - Databricks Performance Optimization
    - Automated Deployment with Databricks Asset Bundle

Estrutura do exame

Seção 1: Desenvolvimento de código para processamento de dados usando Python e SQL

- Uso de Python e ferramentas para desenvolvimento
- Projetar e implementar uma estrutura de projeto Python escalável e otimizada para Databricks Asset Bundles (DABs), possibilitando o desenvolvimento modular, a automação de implantação e a integração de CI/CD.
- Gerenciar e solucionar problemas de instalações e dependências de bibliotecas externas de terceiros no Databricks, incluindo pacotes PyPI, wheels locais e arquivos de origem.
- Desenvolver funções definidas pelo usuário (UDFs) usando Pandas/Python UDF.
- Criação e teste de um pipeline de ETL com Lakeflow Spark Declarative Pipelines, SQL e Apache Spark na Databricks Platform
- Criar e gerenciar pipelines de dados confiáveis e prontos para produção para dados em lote e streaming usando Lakeflow Spark Declarative Pipelines e Autoloader.
- Criar e automatizar cargas de trabalho de ETL usando Jobs via UI/APIs/CLI.
- Explicar as vantagens e desvantagens das tabelas de streaming em comparação com as visões materializadas.
- Usar as APIs APPLY CHANGES para simplificar o CDC no Lakeflow Spark Declarative Pipelines.
- Comparar o Spark Structured Streaming e o Lakeflow Spark Declarative Pipelines para

determinar a abordagem ideal para a criação de pipelines de ETL escaláveis.
- Criar um componente de pipeline que use operadores de fluxo de controle (por exemplo, if/else, for/each, etc.).
- Escolher as configurações apropriadas para ambientes e dependências, alta memória para tarefas de notebook e otimização automática para não permitir novas tentativas.
- Desenvolver testes de unidade e de integração usando assertDataFrameEqual, assertSchemaEqual, DataFrame.transform e frameworks de teste, para garantir a exatidão do código, incluindo um depurador integrado.

Seção 2: Ingestão e aquisição de dados:

- Projetar e implementar pipelines de ingestão de dados para ingerir com eficiência uma variedade de dados
- formatos, incluindo Delta Lake, Parquet, ORC, AVRO, JSON, CSV, XML, Text e Binary a partir de
- diversas fontes, como barramentos de mensagens e armazenamento em nuvem.
- Criar um pipeline de dados somente de acréscimo (append-only) capaz de lidar com dados em lote e streaming usando Delta.

Seção 3: Transformação, limpeza e qualidade de dados

- Escrever código Spark SQL e PySpark eficiente para aplicar transformações de dados avançadas, incluindo funções de janela, junções e agregações, para manipular e analisar grandes conjuntos de dados.
- Desenvolver um processo de quarentena para dados incorretos com Lakeflow Spark Declarative Pipelines ou com autoloader em jobs clássicos.

Seção 4: Compartilhamento e federação de dados

- Demonstrar o compartilhamento delta com segurança entre implantações do Databricks usando o Databricks to Databricks Sharing (D2D) ou para plataformas externas usando o protocolo de compartilhamento aberto (D2O).
- Configurar o Lakehouse Federation com a governança adequada nos sistemas de origem suportados.
- Usar o Delta Share para compartilhar dados ao vivo do Lakehouse para qualquer plataforma de computação.

Seção 5: Monitoramento e alertas

- Monitoramento
    - Usar tabelas do sistema para observabilidade sobre utilização de recursos, custo, auditoria e monitoramento de cargas de trabalho.
    - Usar a Query Profiler UI e a Spark UI para monitorar cargas de trabalho.
    - Usar as Databricks REST APIs/Databricks CLI para monitorar jobs e pipelines.
    - Usar os Event Logs do Lakeflow Spark Declarative Pipelines para monitorar pipelines.

- Alertas
    - Usar os SQL Alerts para monitorar a qualidade dos dados.
    - Usar a Lakeflow Jobs UI e a Jobs API para configurar notificações sobre o status do job e problemas de desempenho.

Seção 6: Otimização de custo e desempenho

- Entender como/por que o uso de tabelas gerenciadas do Unity Catalog reduz as operações
- sobrecarga e o ônus de manutenção.
- Entender as técnicas de otimização delta, como deletion vectors e liquid clustering.
- Entender as técnicas de otimização usadas pelo Databricks para garantir o desempenho de consultas em grandes conjuntos de dados (data skipping, file pruning, etc.).
- Aplicar o Change Data Feed (CDF) para resolver limitações específicas das tabelas de streaming e melhorar a latência.
- Usar o perfil de consulta para analisar a consulta e identificar gargalos, como data skipping inadequado, tipos ineficientes de junções e embaralhamento de dados (data shuffling).

Seção 7: Garantia de segurança e conformidade de dados

- Aplicação de mecanismos de segurança de dados.
    - Usar ACLs para proteger objetos do Workspace, aplicando o princípio de privilégio mínimo, incluindo a aplicação de princípios como privilégio mínimo e imposição de políticas.
    - Usar filtros de linha e máscaras de coluna para filtrar e mascarar dados sensíveis de tabelas.
    - Aplicar métodos de anonimização e pseudonimização, como Hashing, Tokenização, Supressão e generalização, a dados confidenciais.
- Garantia de conformidade
    - Implementar um pipeline de lote e streaming em conformidade que detecte e aplique o mascaramento de PII para garantir a privacidade dos dados.
    - Desenvolver uma solução de expurgo de dados que garanta a conformidade com as políticas de retenção de dados.

Seção 8: Governança de dados

- Criar e adicionar descrições/metadados sobre dados corporativos para torná-los mais detectáveis.
- Demonstrar a compreensão do modelo de herança de permissões do Unity Catalog.

Seção 9: Depuração e implantação

- Depuração e solução de problemas
    - Identificar informações de diagnóstico pertinentes usando a Spark UI, logs de cluster, tabelas do sistema e perfis de consulta para solucionar erros.
    - Analisar os erros e corrigir as execuções de jobs com falha por meio de reparos de

jobs e substituições de parâmetros.
    - Usar os logs de eventos do Lakeflow Spark Declarative Pipelines e a Spark UI para depurar o Lakeflow Spark Declarative Pipelines e os pipelines do Spark.
- Implantação de CI/CD
    - Criar e implantar recursos do Databricks usando os Databricks Asset Bundles.
    - Configurar e integrar com fluxos de trabalho de CI/CD baseados em Git usando os Databricks Git Folders para implantação de notebooks e código.

Seção 10: Modelagem de dados

- Projetar e implementar modelos de dados escaláveis usando o Delta Lake para gerenciar grandes conjuntos de dados.
- Simplificar as decisões de layout de dados e otimizar o desempenho das consultas usando o Liquid Clustering.
- Identificar os benefícios de usar o liquid Clustering em vez de Partitioning e ZOrder.
- Projetar modelos dimensionais para cargas de trabalho analíticas, garantindo consultas e agregações eficientes.

Perguntas de exemplo

Estas perguntas foram aposentadas de uma versão anterior do exame. O objetivo é mostrar a você os objetivos conforme declarados no guia do exame e fornecer uma pergunta de exemplo alinhada 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 a estrutura do exame no guia do exame.

Pergunta 1

Objetivo: Entender as operações de catalog-metastore do Delta Lake e o comportamento de conformidade com ACID.

Uma tabela Delta Lake foi criada com a seguinte consulta:

```
CREATE TABLE prod.sales_by_stor
USING DELTA
LOCATION "/mnt/prod/sales_by_store"
```

Percebendo que a consulta original tinha um erro tipográfico, o código abaixo foi executado: ALTER TABLE prod.sales_by_stor RENAME TO prod.sales_by_store Qual resultado ocorrerá após a execução do segundo comando?

A. Todos os arquivos e metadados relacionados são descartados e recriados em uma única transação ACID.
B. A alteração do nome da tabela é registrada no log de transações do Delta.

C. Um novo log de transações do Delta é criado para a tabela renomeada.
D. A referência da tabela no metastore é atualizada.

Pergunta 2
Objetivo: Entender o comportamento do Spark Structured Streaming e determinar a abordagem ideal para pipelines prontos para o SLA de produção.

Um job do Spark Structured Streaming implantado em produção vem apresentando atrasos durante os horários de pico. Atualmente, durante a execução normal, cada microbatch de dados é processado em menos de 3 segundos. Durante os horários de pico, o tempo de execução de cada microbatch se torna
muito inconsistente, às vezes excedendo 30 segundos. A gravação de streaming está atualmente configurada com um intervalo de gatilho de 10 segundos.

Mantendo todas as outras variáveis constantes e supondo que os registros precisem ser processados em menos de 10 segundos, qual ajuste atenderá ao requisito?

A. Usar a opção trigger once e configurar um job do Databricks para executar a consulta a cada 8 segundos; isso garante que todos os registros acumulados sejam processados a cada lote.
B. Diminuir o intervalo de gatilho para 5 segundos; acionar lotes com mais frequência pode evitar que os registros se acumulem e que lotes grandes causem um spill.
C. Diminuir o intervalo de gatilho para 5 segundos; acionar lotes com mais frequência permite que executores ociosos comecem a processar o próximo lote enquanto as tarefas de execução mais longa dos lotes anteriores terminam.
D. O intervalo de gatilho não pode ser modificado sem modificar o diretório de checkpoint; para manter o estado atual do stream, aumente o número de partições de shuffle para maximizar o paralelismo.

Pergunta 3
Objetivo: Projetar e implementar modelos de dados escaláveis usando o Delta Lake para gerenciar grandes conjuntos de dados.

Uma tabela Delta Lake que representa metadados sobre postagens de conteúdo de usuários tem o seguinte esquema:

```
user_id LONG, post_text STRING, post_id STRING, longitude FLOAT, latitude FLOAT, post_time TIMESTAMP, date DATE
```

Com base no esquema acima, qual coluna é uma boa candidata para o particionamento da Delta Table?

A. post_id
B. post_time
C. date
D. user_id

Pergunta 4
Objetivo: Demonstrar a compreensão do modelo de herança de permissões do Unity Catalog

Uma tabela chamada user_ltv está sendo usada para criar uma visão que será utilizada por analistas de dados de várias equipes. Os usuários no workspace são configurados em grupos, que são usados para configurar o acesso a dados por meio de ACLs.
A tabela user_ltv tem o seguinte esquema:
`email STRING, age INT, ltv INT`
A seguinte definição de visão é executada:

```
CREATE VIEW email_ltv AS
SELECT
CASE WHEN
is_member('marketing') THEN email
ELSE 'REDACTED'
END AS email,
ltv
FROM user_ltv
```

Um analista que não é membro do grupo marketing executa a seguinte consulta: SELECT * FROM email_ltv
Qual será o resultado desta consulta?

A. Somente as colunas email e ltv serão retornadas; a coluna email conterá a string "REDACTED" em cada linha.
B. Três colunas serão retornadas, mas uma coluna será nomeada "REDACTED" e conterá apenas valores nulos.
C. Somente as colunas email e ltv serão retornadas; a coluna email conterá apenas valores nulos.
D. As colunas email e ltv serão retornadas com os valores de user_ltv.

Pergunta 5
Objetivo- Escolher as configurações apropriadas para ambientes e dependências, alta memória para tarefas de notebook e otimização automática para não permitir novas tentativas.

A equipe de negócios exige que os dados de seus dashboards sejam atualizados a cada hora. O tempo total de processamento do pipeline, que extrai, transforma e carrega dados em suas execuções, é de 10 minutos.

Supondo condições normais de operação, qual configuração atenderá aos requisitos do acordo de nível de serviço (SLA) com o menor custo?

A. Agendar um job para executar o pipeline uma vez por hora em um cluster interativo dedicado.
B. Agendar um job para executar o pipeline uma vez por hora em um novo job cluster.
C. Agendar um job do Spark Structured Streaming com um intervalo de gatilho de 60 minutos.
D. Configurar um job que seja executado sempre que novos dados chegarem a um determinado diretório.

Pergunta 6
Objetivo- Entender o ambiente de desenvolvimento do Notebook, o gerenciamento de variáveis e a criação de código seguro e configurável.

A equipe de segurança está avaliando se o módulo de secrets do Databricks pode ser aproveitado para se conectar a um banco de dados externo.
Após testar o código com todas as variáveis Python definidas com strings, eles fazem upload da senha para o módulo de secrets e configuram as permissões corretas para o usuário atualmente ativo. Eles modificam seu código para o seguinte (mantendo todas as outras variáveis inalteradas).

```
password = dbutils.secrets.get(scope="db_creds", key="jdbc_password")
print(password)
df = (spark
.read
.format("jdbc")
.option("url", connection)
.option("dbtable", tablename)
.option("user", username)
.option("password", password)
)
```

O que acontecerá quando este código for executado?

A. A conexão com a tabela externa será bem-sucedida; a string "REDACTED" será impressa. B. A conexão com a tabela externa será bem-sucedida; o valor da string da senha será impresso em texto simples.
C.. Uma caixa de entrada interativa aparecerá no notebook; se a senha correta for fornecida, a conexão será bem-sucedida e a senha será impressa em texto simples.
D. Uma caixa de entrada interativa aparecerá no notebook; se a senha correta for fornecida, a conexão será bem-sucedida e a senha codificada será salva no DBFS.

Pergunta 7
Objetivo: Entender as técnicas de otimização usadas pelo Databricks para garantir o desempenho de consultas em grandes conjuntos de dados (data skipping, file pruning, etc)

Uma tarefa de ingestão de dados exige que um conjunto de dados JSON de 1 TB seja gravado em Parquet com um tamanho de arquivo de 512 MB. Como o Parquet está sendo usado em vez do Delta Lake, recursos integrados de dimensionamento de arquivos, como Auto-Optimize e Auto-Compaction, não podem ser usados.

Qual abordagem funcionará sem reorganizar os dados?

A. Ingerir os dados, executar as transformações estreitas (narrow), reparticionar para 2.048 partições (1TB*1024*1024/512) e, em seguida, gravar em Parquet.
B. Definir spark.sql.adaptive.advisoryPartitionSizeInBytes como 512 MB, ingerir os dados, executar as transformações estreitas (narrow), fazer coalesce para 2.048 partições (1TB*1024*1024/512) e, em seguida, gravar em Parquet.
C. Definir spark.sql.files.maxPartitionBytes como 512 MB, ingerir os dados, executar as transformações estreitas (narrow) e, em seguida, gravar em Parquet.
D. Definir spark.sql.shuffle.partitions como 2.048 partições (1TB*1024*1024/512), ingerir os dados, executar as transformações estreitas (narrow), otimizar os dados ordenando-os (o que reparticiona automaticamente os dados) e, em seguida, gravar em Parquet.

Pergunta 8
Objetivo: Aplicar o clone do Delta Lake para aprender como os shallow clones e deep clones interagem com as tabelas de origem/destino.

A equipe de marketing deseja compartilhar dados em uma tabela agregada com a organização de vendas, mas os nomes dos campos usados pelas equipes não coincidem, e vários campos específicos de marketing não foram aprovados para a organização de vendas.
Qual solução aborda a situação enfatizando a simplicidade?

A. Criar uma visão na tabela de marketing selecionando apenas os campos aprovados para a equipe de vendas; usar alias nos nomes de quaisquer campos que devam ser padronizados de acordo com as convenções de nomenclatura de vendas.
B. Criar uma nova tabela com o esquema necessário e usar a funcionalidade DEEP CLONE do Delta Lake para sincronizar as alterações confirmadas em uma tabela com a tabela correspondente.
C. Usar uma instrução CTAS para criar uma tabela derivada da tabela de marketing e, em seguida, configurar um job de produção para propagar as alterações.
D. Adicionar uma gravação de tabela paralela ao pipeline de produção atual, atualizando uma nova tabela de vendas que varia conforme necessário em relação à tabela de marketing.

Pergunta 9
Objetivo: Criar um job de múltiplas tarefas com várias dependências.

Um job do Databricks foi configurado com três tarefas, cada uma das quais é um notebook do Databricks. A tarefa A não depende de outras tarefas. As tarefas B e C são executadas em paralelo, cada uma com uma dependência serial da tarefa A.

Qual será o estado resultante se as tarefas A e B forem concluídas com sucesso, mas a tarefa C falhar durante uma execução agendada?

A. Toda a lógica expressa no notebook associado às tarefas A e B terá sido concluída com sucesso; algumas operações na tarefa C podem ter sido concluídas com sucesso.
B. A menos que todas as tarefas sejam concluídas com sucesso, nenhuma alteração será confirmada no Lakehouse; como a tarefa C falhou, todas as confirmações serão revertidas automaticamente.
C. Toda a lógica expressa no notebook associado às tarefas A e B terá sido concluída com sucesso; quaisquer alterações feitas na tarefa C serão revertidas devido à falha da tarefa.
D. Como todas as tarefas são gerenciadas como um grafo de dependências, nenhuma alteração será confirmada no Lakehouse até que todas as tarefas tenham sido concluídas com sucesso.

Answers:

Pergunta 1: D
Pergunta 2: B
Pergunta 3: C
Pergunta 4: A
Pergunta 5: B
Pergunta 6: A
Pergunta 7: C
Pergunta 8: A
Pergunta 9 :A