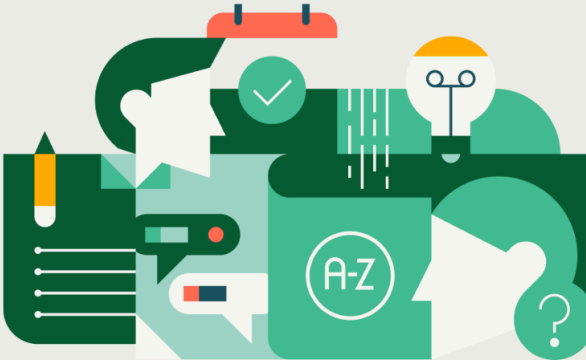
databricks

Databricks 시험 가이드

Databricks Certified Data Engineer Professional



시험 가이드 피드백 제공

이 시험 가이드의 목적

이 시험 가이드의 목적은 시험 개요와 시험 내용을 제공하여 시험 준비 상태를 판단하는 데 도움을 주는 것입니다. 이 문서는 시험에 변경 사항이 있을 때마다 (그리고 해당 변경 사항이 시험에 적용될 때) 업데이트되므로 그에 맞춰 준비할 수 있습니다. 이 버전은 **2025**년 **3**월 **1**일 현재 라이브 버전을 다룹니다**.** 시험을 치르기 **2**주 전에 다시 확인하여 최신 버전을 사용하고 있는지 확인하세요**.**

대상 청중 설명

Databricks Certified Data Engineer Professional 인증 시험은 개인이 Databricks를 사용하여 고급 데이터 엔지니어링 작업을 수행하는 능력을 평가합니다. 여기에는 Databricks 플랫폼과 Apache Spark, Delta Lake, MLflow, Databricks CLI 및 REST API와 같은 개발자 도구에 대한 이해가 포함됩니다. 또한 최적화되고 정리된 ETL 파이프라인을 구축하는 능력도 평가합니다. 또한, 일반 데이터 모델링 개념에 대한 지식을 활용하여 레이크하우스로 데이터를 모델링하는 것도 평가됩니다. 마지막으로, 배포 전에 데이터 파이프라인이 안전하고 안정적이며 모니터링되고 테스트되었는지 확인하는 것도 이 시험에 포함됩니다. 이 인증 시험에 합격한 사람은 Databricks와 관련 도구를 사용하여 고급 데이터 엔지니어링 작업을 완료할 수 있습니다.

시험에 대하여

- 항목 수: 60개의 객관식 문제
- 시간 제한: 120분
- 등록금: USD 200, 현지 법률에 따라 필요한 해당 세금 포함
- 시험 진행 방식: 온라인 감독
- 테스트 보조 도구: 허용되지 않음.
- 필수 조건: 필요한 것은 없습니다. 교육 참석 및 Databricks에서 1년의 실무 경험이 강력히 권장됩니다.
- 유효 기간: 2년
- 재인증: 인증 상태를 유지하려면 2년마다 재인증을 받아야 합니다. 재인증을 받으려면 현재 제공 중인 전체 시험에 응시해야 합니다. 시험에 다시 응시할 준비를 하려면 시험 웹페이지의 '시험 준비하기' 섹션을 검토하세요.

- 점수가 매겨지지 않은 콘텐츠: 시험에는 추후 사용할 통계적 정보를 수집하기 위해 채점되지 않은 항목이 포함될 수 있습니다. 이러한 항목은 양식에 명시되지 않으며 점수에 영향을 미치지 않으며, 이 콘텐츠에는 추가 시간이 고려하여 반영됩니다.

추천 교육

- 자기 주도 학습 (Databricks Academy에서 이용 가능):
    - Databricks 스트리밍 및 Delta Live Table - Delta Dawn
    - Databricks 데이터 개인정보 보호
    - Databricks 성능 최적화 Delta Dawn
    - Databricks 자산 번들을 사용한 자동화된 테스트 및 배포

시험 개요

섹션 **1: Databricks** 툴링

- Delta Lake가 트랜잭션 로그와 클라우드 개체 스토리지를 사용하여 원자성과 내구성을 보장하는 방법을 설명하세요.
- Delta Lake의 낙관적 동시성 제어가 어떻게 격리를 제공하는지, 어떤 트랜잭션이 충돌할 수 있는지 설명하세요.
- Delta Clone의 기본 기능을 설명하세요.
- 분할, zorder, bloom 필터, 파일 크기를 포함한 일반적인 Delta Lake 인덱싱 최적화를 적용합니다.
- Databricks SQL 서비스에 최적화된 Delta 테이블 구현
- 데이터 분할을 위한 다양한 전략 비교 (예: 사용할 적절한 분할 열 식별)

섹션 **2:** 데이터 처리**(**일괄 처리**,** 증분 처리 및 최적화**)**

- 파티션 힌트를 설명하고 구별하세요. 병합, 재파티션, 범위별 재파티션 및 재조정
- 데이터 분할을 위한 다양한 전략 비교 (예: 사용할 적절한 분할 열 식별)
- 개별 파트 파일의 크기를 수동으로 제어하면서 Pyspark 데이터프레임을 디스크에 쓰는 방법을 설명합니다.
- Spark 테이블에서 1개 이상의 레코드를 업데이트하기 위한 여러 가지 전략을 설명합니다 (유형 1)
- 구조적 스트리밍과 Delta Lake에서 활용된 일반적인 디자인 패턴을 구현합니다.
- 스트림 정적 조인과 Delta Lake를 사용하여 상태 정보를 탐색하고 조정하십시오.
- 스트림 정적 조인을 구현합니다.
- Spark 구조화 스트리밍을 사용하여 중복 제거에 필요한 논리를 구현합니다.
- Delta Lake 테이블에서 CDF를 활성화하고 일반 구조화 스트리밍 읽기에서 증분 피드 대신 CDC를 처리하도록 데이터 처리 단계를 재설계합니다.

- CDF를 활용하여 삭제를 쉽게 전파합니다.
- 적절한 데이터 분할을 통해 데이터의 간단한 보관 또는 삭제가 어떻게 가능한지 설명하다.
- "작은 파일" (스캐닝 오버헤드, 과도한 분할 등)이 Spark 쿼리에 성능 문제를 유발하는 방식을 설명합니다.

섹션 **3:** 데이터 모델링

- 브론즈에서 실버로 승격하는 동안 데이터 변환의 목적을 설명합니다.
- CDF (변경 데이터 피드)가 레이크하우스 아키텍처 내에서 업데이트 및 삭제를 전파하는 데 있어 과거의 어려움을 어떻게 해결하는지 논의합니다.
- Delta Lake 복제본을 적용하여 얕은 복제본과 깊은 복제본이 소스/대상 테이블과 상호 작용하는 방식을 알아보세요.
- 스트리밍 워크로드를 프로덕션 환경에 적용하려고 할 때 흔히 저지르는 함정을 피하기 위해 다중화 브론즈 테이블을 설계합니다.
- 다중화 브론즈 테이블에서 데이터를 스트리밍할 때 모범 사례를 구현합니다.
- 브론즈에서 실버까지의 데이터를 처리하기 위해 증분 처리, 품질 강화 및 중복 제거를 적용합니다.
- Delta Lake의 다양한 접근 방식의 장점과 한계를 기반으로 데이터 품질을 강화하는 방법에 대한 정보에 입각한 결정을 내리십시오.
- 외래 키 제약 조건의 부재로 인한 문제를 피하도록 테이블을 구현하십시오
- 잘못된 데이터가 기록되는 것을 방지하기 위해 Delta Lake 테이블에 제약 조건을 추가하십시오
- 조회 테이블을 구현하고 정규화된 데이터 모델에 대한 상충 관계를 설명하십시오
- 스트리밍 및 배치 작업을 사용하여 Delta Lake를 사용하여 느리게 변경되는 차원 테이블을 구현하는 데 필요한 아키텍처 및 작업을 다이어그램화하십시오.
- SCD 유형 0, 1 및 2 테이블 구현

섹션 **4:** 보안 및 거버넌스

- 데이터 마스킹을 수행하기 위한 동적 뷰 생성
- 동적 뷰를 사용하여 행과 열에 대한 액세스를 제어합니다.

섹션 **5:** 모니터링 및 로깅

- 성능 분석, 애플리케이션 디버깅, 그리고 Spark 애플리케이션의 튜닝을 돕기 위해 Spark UI의 요소를 설명하세요.
- 클러스터에서 수행되는 스테이지와 작업에 대한 이벤트 타임라인과 메트릭을 검사하는 능력,
- Spark UI, Ganglia UI, Cluster UI에서 제시된 정보를 바탕으로 결론을 도출하고,성능 문제를 평가하고 실패한 애플리케이션을 디버깅하는 능력이 평가됩니다.
- 프로덕션 스트리밍 작업에 대한 비용 및 지연 시간 SLA를 제어하는 시스템을 설계합니다.
- 스트리밍 및 일괄 작업 배포 및 모니터링

섹션 **6:** 테스트 및 배포

- Python 파일 종속성을 사용하도록 노트북 종속성 패턴 조정
- Wheels로 유지 관리되는 Python 코드를 상대 경로를 사용하는 직접 import로

조정
- 실패한 작업 복구 및 재실행
- 일반적인 사용 사례 및 패턴을 기반으로한 작업을 생성
- 여러 종속성을 가진 다중 작업을 생성
- 프로덕션 스트리밍 작업에 대한 비용 및 지연 시간 SLA를 제어하는 시스템을 설계합니다.
- Databricks CLI를 구성하고 기본 명령을 실행하여 작업 공간 및 클러스터와 상호 작용합니다.
- CLI에서 명령을 실행하여 Databricks Jobs을 배포하고 모니터링합니다.
- REST API를 사용하여 작업을 복제하고 실행을 트리거하고 실행 출력을 내보냅니다.

샘플 질문

이 문제는 이전 버전의 시험에서 사용 중단되었습니다. 목적은 시험 안내서에 명시된 대로 목표를 보여주고 목표에 맞는 샘플 문제를 제공하는 것입니다.시험 안내서에는 시험에서 다룰 수 있는 목표가 나열되어 있습니다. 인증 시험을 준비하는 가장 좋은 방법은 시험 가이드의 시험 개요를 검토하는 것입니다.

질문 **1**

목표: 쿼리로 생성된 *Delta Lake* 테이블에서 실행한 명령의 결과를 식별합니다.

다음 쿼리를 사용하여 Delta Lake 테이블이 생성되었습니다.

```
CREATE TABLE dev.my_table
USING DELTA
LOCATION "/mnt/dev/my_table"
```

다른 사람들이 테이블을 사용할 필요가 있으며 테이블 이름이 오해의 소지가 있다는 것을 깨닫고 아래 코드를 실행했습니다.

```
ALTER TABLE dev.my_table RENAME TO dev.our_table
```

두 번째 명령을 실행하면 어떤 결과가 나올까요?

A. 테이블 이름 변경은 Delta 트랜잭션 로그에 기록됩니다.

B. 메타스토어의 테이블 참조가 업데이트되고 모든 데이터 파일이 이동됩니다.

C. 메타스토어의 테이블 참조가 업데이트되고 데이터는 변경되지 않습니다.

D. 이름이 바뀐 테이블에 대해 새로운 Delta 트랜잭션 로그가 생성됩니다.

E. 모든 관련 파일과 메타데이터는 단일 ACID 트랜잭션에서 삭제되고 다시 생성됩니다.

질문 **2**

목표: *Delta table*에 삽입될 때 이전에 처리된 레코드에 대해 데이터 중복을 제거합니다.

데이터 엔지니어는 단일 소스에서 늦게 도착하는 중복 레코드를 처리할 수 있는 ETL 워크플로를 개발하고 있습니다. 데이터 엔지니어는 배치 내에서 레코드 중복을 제거할 수 있다는 사실을 알고 있지만 다른 솔루션을 찾고 있습니다.
어떤 접근 방식을 사용하면 데이터 엔지니어가 이전에 처리된 레코드를 Delta table에 삽입할 때 데이터 중복을 제거할 수 있습니까?

A. 각 배치가 완료된 후에는 Delta Table을 **VACUUM** 합니다.
B. 중복 레코드를 방지하기 위해 Delta Lake 스키마 적용을 활용합니다.
C. 구성을 **`delta.deduplicate = true`**로 설정합니다.
D. 고유 키에 대해 전체 외부 조인을 수행하고 기존 데이터를 덮어씁니다.
E. 고유 키에 대한 일치 조건으로 삽입 전용 병합을 수행합니다.

질문 **3**
목표: *Lakehouse*의 모든 테이블을 외부, 관리되지 않는 *Delta Lake* 테이블로 구성하는 방법을 파악하다.

데이터 아키텍트는 Lakehouse의 모든 테이블을 외부의 관리되지 않는 Delta Lake 테이블로 구성하라고 명령했다.

어떤 접근 방식을 사용하면 이 요구 사항을 충족할 수 있을까요?

A. 테이블을 생성할 때마다 **`LOCATION`** 키워드를 사용해야 합니다.
B. 테이블을 생성할 때 **`CREATE TABLE`** 문에서 **`EXTERNAL`** 키워드를 사용해야 합니다.
C. 워크스페이스를 구성할 때 외부 클라우드 개체 스토리지가 마운트되었는지 확인하세요.
D. 데이터베이스를 생성할 때마다 **`LOCATION`** 키워드를 사용해야 합니다.
E. 테이블을 생성할 때마다 **`LOCATION`**과 **`UNMANAGED`** 키워드를 사용해야 합니다.

질문 **4**
목표: *Databricks jobs*에 대한 권한 제어를 설명하세요

데이터 엔지니어링 팀은 팀을 바꾼 개인으로부터 Databricks Workflows의 소유권을 이전하려고 합니다. 하지만 Databricks Jobs에 대한 권한 제어가 구체적으로 어떻게 작동하는지는 잘 모릅니다.

Databricks Jobs에 대한 권한 제어를 올바르게 설명한 것은 무엇입니까?

A. Databricks Job의 생성자는 항상 "소유자" 권한을 가지며, 이 구성은 변경할 수 없습니다.

B. Databricks Jobs에는 정확히 한 명의 소유자가 있어야 합니다. "소유자" 권한은 그룹에 할당할 수 없습니다.

C. 기본 "관리자" 그룹 외에는 개별 사용자에게만 권한이 부여될 수 있습니다.
작업.

D. 워크스페이스 관리자만이 그룹에 "소유자" 권한을 부여할 수 있습니다.E. 사용자는 해당 그룹의 구성원인 경우에만 작업 소유권을 그룹에 이전할 수 있습니다.

질문 **5**

목표*: Python* 패키지를 설치하는 방법을 식별하세요...

데이터 엔지니어는 Python 패키지를 사용하여 데이터를 처리해야 합니다. 따라서 현재 활성화된 클러스터의 모든 노드에 Python 패키지를 설치해야 합니다.

현재 활성화된 클러스터의 모든 노드에 노트북 수준으로 범위가 지정된 Python 패키지를 설치하는 방법을 설명하는 것은 무엇입니까?

A. 노트북 셀에서 **`%pip install`**을 사용합니다.

B. 노트북 셀에서 **`%sh pip install`**을 사용합니다.

C. 노트북 설정 스크립트에서 **`source env/bin/activate`**를 실행하세요.

D. 클러스터 UI를 사용하여 PyPI에서 라이브러리를 설치하세요

E. 노트북 셀에서 **`b`** 를 사용하세요

답변

질문 1: C

질문 2: E

질문 3: A

질문 4: B

질문 5: A