databricks

Databricks 시험 가이드

Databricks 공인 데이터 엔지니어 어소시에이트

시험 가이드 피드백 제공

이 시험 가이드의 목적

이 시험 가이드의 목적은 시험에 대한 개요와 시험에서 다루는 내용을 제공하여 시험 준비 상태를 결정하는 데 도움이 되도록 하는 것입니다. 이 문서는 시험에 변경 사항이 있을 때마다(그리고 이러한 변경 사항이 시험에 적용되는 시기) 업데이트되어 준비할 수 있습니다. 이 버전은 **2025**년 **7**월 **25**일 현재 라이브 시험을 다룹니다**.** 시험을 치르기 **2**주 전에 다시 확인하여 최신 버전을 사용하고 있는지 확인하십시오**.**

대상 설명

Databricks Certified Data Engineer Associate 인증 시험은 Databricks Data Intelligence Platform를 사용하여 입문 데이터 엔지니어링 작업을 완료하는 개인의 능력을 평가합니다. 여기에는 Databricks Data Intelligence Platform과 해당 작업 공간, 아키텍처 및 기능에 대한 이해가 포함됩니다. 시험은 또한 추출, 복잡한 데이터 처리 및 사용자 정의 함수를 다루는 Apache Spark SQL 또는 PySpark를 사용하여 ETL 작업을 수행하는 능력을 평가합니다. 마지막으로 시험은 Databricks Workflows 작업을 효과적으로 구성하고 예약하여 워크로드를 배포하고 오케스트레이션하는 테스터의 능력을 평가합니다.
이 인증 시험에 합격한 개인은 Databricks 및 관련 도구를 사용하여 기본 데이터 엔지니어링 작업을 완료할 것으로 예상할 수 있습니다.

시험 정보

- 채점된 항목 수: 객관식 질문 45개
- 시간 제한: 90분
- 등록비: 미화 200달러, 현지 법률에 따라 요구되는 관련 세금
- 배송 방법: 온라인 감독
- 시험 보조 도구: 허용되지 않습니다.
- 전제 조건: 필요하지 않습니다. 강좌 참석 및 Databricks에서 6개월의 실무 경험을 적극 권장합니다.
- 유효성: 2년
- 재인증: 인증 상태를 유지하려면 2년마다 재인증이 필요합니다. 재인증을 받으려면 현재 진행 중인 전체 시험에 응시해야 합니다. 시험 웹페이지의 "시험 준비" 섹션을 검토하여 시험에 다시 응시할 준비를 하십시오.

- 점수가 매겨지지 않은 콘텐츠: 시험에는 나중에 사용할 수 있도록 통계 정보를 수집하기 위해 채점되지 않은 항목이 포함될 수 있습니다. 이러한 항목은 양식에서 식별되지 않으며 점수에 영향을 미치지 않습니다. 이 콘텐츠에는 추가 시간이 고려됩니다.

권장 교육

- 강사 주도: Databricks를 사용한 데이터 엔지니어링
- 자기 주도형(Databricks 아카데미에서 사용 가능):
    - Lakeflow Connect를 사용한 데이터 수집
    - LakeFlow 작업을 사용하여 워크로드 배포
    - Lakeflow 선언형 파이프라인으로 데이터 파이프라인 구축
    - 데이터 엔지니어링을 위한 DevOps 기본

시험 개요

섹션 **1: Databricks** 인텔리전스 플랫폼
- 데이터 레이아웃 결정을 단순화하고 쿼리 성능을 최적화하는 기능을 활성화합니다.
- 데이터 인텔리전스 플랫폼의 가치를 설명하세요.
- 특정 사용 사례에 사용할 적용 가능한 컴퓨트 자원을 식별합니다.

섹션 **2 :** 개발 및 수집
- 데이터 엔지니어링 워크플로에서 Databricks Connect 사용
- Notebooks 기능의 역량 결정
- 유효한 Auto Loader 소스 및 사용 사례 분류
- Auto Loader 구문에 대한 지식 입증
- Databricks의 내장 디버깅 도구를 사용하여 지정된 문제를 해결합니다

섹션 **3:** 데이터 처리 및 변환
- 메달리온 아키텍처의 세 가지 계층을 설명하고 데이터 처리 파이프라인에서 각 계층의 목적을 설명합니다.
- 클러스터가 사용되는 시나리오에 따라 최적의 성능을 위해 클러스터 및 구성을 분류합니다.
- DLT의 장점을 강조합니다(Databricks의 ETL 프로세스용).
- DLT를 사용하여 데이터 파이프라인을 구현합니다.
- DDL(데이터 정의 언어)/DML 기능을 식별합니다.
- PySpark DataFrames를 사용하여 복잡한 집계 및 메트릭을 계산합니다.

섹션 **4 :** 데이터 파이프라인 프로덕션화
- DAB와 기존 배포 방법의 차이점을 식별합니다.

- 자산 번들의 구조를 식별합니다.
- 워크플로를 배포하고, 복구하고, 실패 시 태스크를 다시 실행합니다.
- Databricks에서 관리하는 무간담 자동 최적화 컴퓨팅을 위해 서버리스를 사용합니다.
- 쿼리를 최적화하기 위해 Spark UI를 분석합니다.

섹션 **5 :** 데이터 거버넌스 및 품질

- 관리되는 테이블과 외부 테이블의 차이점을 설명합니다.
- UC 내의 사용자 및 그룹에 대한 권한 부여를 식별합니다.
- UC의 주요 역할을 식별합니다.
- 감사 로그가 저장되는 방법을 식별합니다.
- Unity Catalog에서 리니지 기능을 사용합니다.
- Unity Catalog와 함께 사용할 수 있는 Delta Sharing 기능을 사용하여 데이터를 공유합니다.
- Delta Sharing의 장점과 한계를 파악하세요.
- Delta 공유 유형(Databricks 대 외부 시스템)을 식별합니다.
- 클라우드 간 Data Sharing의 비용 고려 사항 분석
- 외부 소스에 연결된 경우 Lakehouse Federation의 사용 사례를 식별합니다.

샘플 질문

이러한 문제는 이전 버전의 시험에서 사용 중지되었습니다. 목적은 시험 가이드에 명시된 목표를 보여주고 목표에 맞는 샘플 질문을 제공하는 것입니다. 시험 가이드에는 시험에서 다룰 수 있는 목표가 나열되어 있습니다. 인증 시험을 준비하는 가장 좋은 방법은 시험 가이드의 시험 개요를 검토하는 것입니다.

질문 **1**

데이터 엔지니어가 데이터 파이프라인의 일부로 Delta 테이블을 만들었습니다. 이제 다운스트림 데이터 애널리스트에게는 Delta 테이블에 대한 SELECT 권한이 필요합니다.

데이터 엔지니어가 데이터 애널리스트에게 적절한 액세스 권한을 부여하는 데 사용할 수 있는 Databricks 레이크하우스 플랫폼의 어떤 부분인가요?

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

질문 **2**

데이터세트는 Delta Live Tables를 사용하여 정의되었으며 expects 절을 포함합니다.

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

이러한 제약 조건을 위반하는 데이터가 포함된 데이터 배치가 처리될 때 예상되는 동작은 무엇입니까?

A. 예상을 위반하는 레코드는 대상 데이터세트에서 삭제되고 이벤트 로그에 유효하지 않은 것으로 기록됩니다.
B. 예상을 위반하는 레코드는 대상 데이터세트에 추가되고 이벤트 로그에 유효하지 않은 것으로 기록됩니다.
C. 예상을 위반하는 레코드로 인해 작업이 실패합니다.
D. 예상을 위반하는 레코드는 대상 데이터세트에 추가되고 대상 데이터세트에 추가된 필드에서 유효하지 않은 것으로 플래그가 지정됩니다.

질문 **3**

Delta Live Table 파이프라인에는 STREAMING LIVE TABLE을 사용하여 정의된 두 개의 데이터세트가 포함됩니다. LIVE TABLE 을 사용하여 Delta Lake 테이블 소스를 기반으로 세 개의 데이터세트가 정의됩니다.

테이블은 트리거된 파이프라인 모드(Triggered Pipeline Mode)를 사용하여 개발 모드에서 실행되도록 구성됩니다.

이전에 처리되지 않은 데이터가 존재하고 모든 정의가 유효하다는 점을 고려할 때 시작을 클릭하여 파이프라인을 업데이트한 후 예상되는 결과는 무엇입니까?

A. 모든 데이터세트는 파이프라인이 종료될 때까지 설정된 간격으로 업데이트됩니다. 컴퓨팅 리소스는 추가 테스트를 허용하기 위해 파이프라인이 중지된 후에도 유지됩니다.
B. 모든 데이터세트가 한 번 업데이트되고 파이프라인이 종료됩니다. 컴퓨팅 리소스가 종료됩니다.
C. 모든 데이터세트는 파이프라인이 종료될 때까지 설정된 간격으로 업데이트됩니다. 컴퓨팅 리소스는 업데이트를 위해 배포되고 파이프라인이 중지되면 종료됩니다.
D. 모든 데이터세트가 한 번 업데이트되고 파이프라인이 종료됩니다. 컴퓨팅 리소스는 추가 테스트를 허용하기 위해 유지됩니다.

질문 **4**

Delta Live Table 파이프라인에는 STREAMING LIVE TABLE을 사용하여 정의된 두 개의 데이터 세트가 포함됩니다. LIVE TABLE을 사용하여 Delta Lake 테이블 소스에 대해 세 개의 데이터 세트가 정의됩니다.

테이블은 연속 파이프라인 모드(Continuous Pipeline Mode) 를 사용하여 개발 모드에서 실행되도록 구성됩니다.

이전에 처리되지 않은 데이터가 존재하고 모든 정의가 유효하다고 가정하면 시작을 클릭하여 파이프라인을 업데이트한 후 예상되는 결과는 무엇입니까?

A. 모든 데이터세트는 파이프라인이 종료될 때까지 설정된 간격으로 업데이트됩니다. 컴퓨팅 리소스는 파이프라인이 종료될 때까지 유지됩니다.
B. 모든 데이터세트가 한 번 업데이트되고 파이프라인이 종료됩니다. 컴퓨팅 리소스는 추가 테스트를 허용하기 위해 유지됩니다.
C. 모든 데이터세트가 한 번 업데이트되고 파이프라인이 종료됩니다. 컴퓨팅 리소스가 종료됩니다.
D. 모든 데이터세트는 파이프라인이 종료될 때까지 설정된 간격으로 업데이트됩니다. 컴퓨팅 리소스는 추가 테스트를 허용하기 위해 유지됩니다.

질문 **5**

새로운 데이터 엔지니어링 팀 이 ELT 프로젝트에 할당되었습니다. 새로운 데이터 엔지니어링 팀은 프로젝트를 완전히 관리하기 위해 테이블 판매 에 대한 모든 권한이 필요합니다.

새 데이터 엔지니어링 팀에 데이터베이스에 대한 모든 권한을 부여하는 데 사용할 수 있는 명령은 무엇인가요?

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;`

답변
질문 1 : C
질문 2 : B
질문 3 : D
질문 4: A
질문 5: D