databricks

Databricks 試験ガイド

Databricks Certified Data Engineer Associate

試験ガイドに関するフィードバックの提供

この試験ガイドの目的

この試験ガイドでは、試験の準備に役立てていただくために、試験の概要と試験の対象範囲について説明します。試験に何らかの変更がある場合には(そして、それらの変更が試験に反映される際には)、試験の準備を行えるように、このドキュメントは随時更新されます。このバージョンは、**2025** 年 **7** 月 **25** 日現在の実施試験に対応しています。試験を受ける **2** 週間前に、最新版のドキュメントであることを再度ご確認ください。

対象者についての説明

Databricks Certified Data Engineer Associate 認定試験では、Databricks Data Intelligence Platform を使用して入門的なデータエンジニアリングタスクを完了する個人の能力を評価します。これには、Data Intelligence Platform とそのワークスペース、アーキテクチャ、およびその機能の理解が含まれます。この試験は、Apache Spark SQLまたはPySparkを使用してETLタスクを実行する能力も評価し、抽出、複雑なデータ処理、ユーザー定義関数に関する内容をカバーします。最後に、この試験では、Databricks Workflows ジョブの構成とスケジューリングを効果的に行い、ワークロードをデプロイおよびオーケストレートする受験者の能力を評価します。
この認定試験に合格した個人は、Databricks とその関連ツールを使用して基本的なデータエンジニアリングタスクを完了することが期待されます。

試験について

- 採点項目数: 45 問の多肢選択問題
- 期限： 時間制限:90分
- 受験料: 200 米ドル、現地の法律によって適用される税金が加算されます
- 実施方法: オンライン監督付き
- 試験の補助: 許可されていません。
- 前提: 特になし。Databricksでのコース受講と6か月の実務経験を強くお勧めします
- 有効： 2 年間
- 再認定: 認定資格を維持するためには、2年ごとに再認定を受ける必要があります。再認定を受けるには、現在実施中の完全な試験を受ける必要があります。試験のウェブページの「試験の準備」セクションを確認して、再度試験を受ける準備をしてください。
- 採点対象外の内容: 試験には、将来の使用のために統計情報を収集するために、採点対象外の項目が含まれている場合があります。これらの項目はフォーム上では特定されず、得点には影響しません。この内容については、追加の時間が考慮されています。

推奨されるトレーニング

- インストラクター主導: Databricks を使用したデータエンジニアリング
- 自分のペースで進められる (Databricks Academy で利用可能):
    - Lakeflow Connectによるデータ取り込み
    - Lakeflow ジョブを使用したワークロードのデプロイ
    - Lakeflow宣言型パイプラインでデータパイプラインを構築する
    - Unity Catalogによるデータ管理とガバナンス

試験の概要

セクション **1: Databricks Intelligence Platform**

- データレイアウトの決定を簡素化し、クエリパフォーマンスを最適化する機能を有効化する。
- Data Intelligence Platformの価値を説明する。
- 特定のユースケースに使用する適切なコンピューティングを特定する。

セクション **2:** 開発と取り込み

- データエンジニアリングワークフローでの Databricks Connect の使用
- Notebooksの機能を把握する
- 有効な Auto Loader ソースとユースケースの分類
- Auto Loader構文の知識を示す
- Databricks の組み込みデバッグツールを使用して、特定の問題のトラブルシューティングを行う

セクション **3:** データ処理と変換

- メダリオンアーキテクチャの 3 つのレイヤーについて説明し、データ処理パイプラインにおける各レイヤーの目的を説明する。
- 使用されるシナリオに基づいて最適なパフォーマンスを発揮するように、クラスターの種類と構成を分類する。
- DLT の利点を強調する (Databricks の ETL プロセス用)。
- DLT を使用してデータパイプラインを実装する。
- DDL(データ定義言語)/DMLの機能を特定する。
- PySpark DataFrames を使用して複雑な集計とメトリクスを計算する

セクション **4:** データパイプラインの本番運用

- DAB と従来のデプロイ方法の違いを特定する。
- アセットバンドルの構造を特定する。
- ワークフローをデプロイし、障害が発生した場合にタスクを修復し、再実行する。

-
- Databricksによって管理されるサーバレスを使用し、コンピュートの運用管理を自動最適化する。
- Spark UI を分析してクエリを最適化する。

セクション **5:** データガバナンスと品質

- マネージドテーブルと外部テーブルの違いを説明する。
- UC 内のユーザーおよびグループへのアクセス許可の付与を特定する。
- UC における重要なロールを特定する。
- 監査ログの保存方法を特定する。
- Unity Catalog のリネージ機能を使用する。
- Unity Catalog の Delta Sharing 機能を使用し、データを共有する。
- Delta Sharing の利点と制限を特定する。
- Delta Sharing の種類 - Databricks 対外部システムを特定する。
- クラウド間でのデータ共有のコストに関する考慮事項を分析する
- 外部ソースに接続した場合の Lakehouse Federation の使用事例を特定する。

サンプル問題

これらの問題は、以前のバージョンの試験から廃止されたものです。目的は、試験ガイドに記載されている項目に沿ったサンプル問題を提供することです。試験ガイドには、試験の出題対象となる項目の一覧が記載されています。認定試験の準備を行う最善の方法は、試験ガイドの試験の概要を確認することです。

質問 **1**

データエンジニアが、データパイプラインの一部として Delta テーブルを作成しました。このテーブルを利用するデータアナリストには、Delta テーブルに対する SELECT 権限が必要になりました。

データエンジニアは、Databricks レイクハウスプラットフォームのどの部分を使用して、データアナリストに適切なアクセス権を付与できますか?

A. ジョブ
B. Dashboards
C. データエクスプローラ
D. Repos

質問 **2**

データセットは Delta Live Tables を使用して定義されており、expectations 句が含まれています。

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

これらの制約に違反するデータを含むデータのバッチが処理されると、どのような動作が予想されますか。

A. 期待値に違反するレコードはターゲットデータセットから削除され、イベントログに無効として記録されます。
B. 期待値に違反するレコードはターゲットデータセットに追加され、イベントログに無効として記録されます。
C. 期待値に違反するレコードがあると、ジョブは失敗します。
D. 期待値に違反するレコードはターゲットデータセットに追加され、ターゲットデータセットに追加されるフィールドに無効としてフラグが付けられます。

質問3

Delta Live Table パイプライン には **STREAMING LIVE TABLE**を使用して定義された 2 つのデータセットが含まれます。3 つのデータセットは、**LIVE TABLE** を使用して Delta Lake テーブルソースに対して定義されます。

テーブルは、開発モードを利用し、トリガー実行のパイプラインとして構成されています。

以前に未処理のデータが存在し、すべての定義が有効であることを考えると、[開始] をクリックしてパイプラインを更新した後、予想される結果はどうなるでしょうか。

A. すべてのデータセットは、パイプラインがシャットダウンされるまで、設定された間隔で更新されます。コンピュートリソースは、追加のテストを可能にするためにパイプラインが停止された後も保持されます。
B. すべてのデータセットが 1 回更新され、パイプラインがシャットダウンされます。コンピュートリソースは終了します。
C. すべてのデータセットは、パイプラインがシャットダウンされるまで、設定された間隔で更新されます。コンピュートリソースは更新用にデプロイされ、パイプラインが停止すると終了します。
D. すべてのデータセットが 1 回更新され、パイプラインがシャットダウンされます。コンピュートリソースは、追加のテストを可能にするために保持されます。

質問 4

Delta Live Table パイプライン には、STREAMING LIVE TABLE を使用して定義された 2 つのデータセットが含まれます。LIVE TABLE を使用して、Delta Lake テーブルソースに対して 3 つのデータセットが定義されています。

テーブルは、連続パイプラインモードを使用して開発モードで実行するように構成されています。

以前に未処理のデータが存在し、すべての定義が有効であると仮定すると、[開始] をクリックしてパイプラインを更新した後、予想される結果はどうなるでしょうか。

A. すべてのデータセットは、パイプラインがシャットダウンされるまで、設定された間隔で更新されます。コンピュートリソースは、パイプラインがシャットダウンされるまで保持されます。
B. すべてのデータセットが 1 回更新され、パイプラインがシャットダウンされます。コンピュートリソースは、追加のテストを可能にするために保持されます。
C. すべてのデータセットが 1 回だけ更新され、パイプラインがシャットダウンされます。コンピュートリソースは終了します。
D. すべてのデータセットは、パイプラインがシャットダウンされるまで、設定された間隔で更新されます。コンピュートリソースは、追加のテストを可能にするために保持されます。

質問 5

新しいデータエンジニアリングチーム **team** がELTプロジェクトに割り当てられました。新しいデータエンジニアリングチームには、プロジェクトを完全に管理するために、テーブル **sales** に対する完全な権限が必要です。

新しいデータエンジニアリングチームにデータベースに対するフル権限を付与するには、次のコマンドはどれですか。

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;`

回答
質問 1: C
質問 2: B
質問 3: D

質問4: A
質問5: D