databricks

Databricks Exam Guide

# Databricks Certified Data Engineer Associate

Provide Exam Guide Feedback

## Purpose of this Exam Guide

The purpose of this exam guide is to give you an overview of the exam and what is covered on the exam to help you determine your exam readiness. This document will get updated anytime there are any changes to an exam (and when those changes will take effect on an exam) so that you can be prepared. **This version covers the currently live exam as of July 25th, 2025. Please check back two weeks before you take your exam to make sure you have the most current version.**

## Audience Description

The Databricks Certified Data Engineer Associate certification exam assesses an individual's ability to use the Databricks Data Intelligence Platform to complete introductory data engineering tasks. This includes an understanding of the Data Intelligence Platform and its workspace, its architecture, and its capabilities. It also assesses the ability to perform ETL tasks using Apache Spark SQL or PySpark, covering extraction, complex data handling and User defined functions. Finally, the exam assesses the tester's ability to  deploy and orchestrate workloads with Databricks workflows configuring and scheduling jobs effectively.
Individuals who pass this certification exam can be expected to complete basic data engineering tasks using Databricks and its associated tools.

## About the Exam

- Number of scored items: 45 multiple-choice questions
- Time limit: 90 minutes
- Registration fee: USD 200, plus applicable taxes as required per local law
- Delivery method: Online Proctored
- Test aides: none allowed.
- Prerequisite: None required; course attendance and six months of hands-on experience in Databricks is highly recommended
- Validity: 2 years
- Recertification: Recertification is required every two years to maintain your certified status. To recertify, you must take the full exam that is currently live. Please review the "Getting Ready for the Exam" section on the exam webpage to prepare for taking the exam again.

- Unscored Content: Exams may include unscored items to gather statistical information for future use. These items are not identified on the form and do not impact your score. Additional time is factored into account for this content.

## Recommended Training

- Instructor-led: [Data Engineering with Databricks]
- Self-paced (available in Databricks Academy):
    - Data Ingestion with Lakeflow Connect
    - Deploy Workloads with LakeFlow Jobs
    - Build Data Pipelines with Lakeflow Declarative pipeline
    - Data Management and Governance with Unity Catalog

## Exam outline

**Section 1: Databricks Intelligence Platform**

- Enable features that simplify data layout decisions and optimize query performance.
- Explain the value of the Data Intelligence Platform.
- Identify the applicable compute to use for a specific use case.

**Section 2: Development and Ingestion**

- Use Databricks Connect in a data engineering workflow
- Determine the capabilities of Notebooks functionality
- Classify valid Auto Loader sources and use cases
- Demonstrate knowledge of Auto Loader syntax
- Use Databricks' built-in debugging tools to troubleshoot a given issue

**Section 3: Data Processing & Transformations**

- Describe the three layers of the Medallion Architecture and explain the purpose of each layer in a data processing pipeline.
- Classify the type of the cluster and configuration for optimal performance based on the scenario on which cluster is used.
- Emphasize the advantages of DLT (for ETL process in Databricks).
- Implement data pipelines using DLT..
- Identify DDL (Data Definition Language)/DML features.
- Compute complex aggregations and Metrics with PySpark Dataframes.

**Section 4: Productionizing Data Pipelines**

- Identify the difference between DAB and traditional deployment methods.
- Identify the structure of Asset Bundles.
- Deploy a workflow, repair, and rerun a task in case of failure.
- Use serverless for a hands-off, auto-optimized compute managed by Databricks.
- Analyzing the Spark UI to optimize the query.

**Section 5: Data Governance & Quality**

- Explain the difference between managed and external tables.
- Identify the grant of permissions to users and groups within UC.
- Identify key roles in UC.
- Identify how audit logs are stored.
- Use lineage features in Unity Catalog.
- Use the Delta Sharing feature available with Unity Catalog to share data.
- Identify the advantages and limitations of Delta sharing.
- Identify types of delta sharing- Databricks vs external system.
- Analyze the cost considerations of data sharing across clouds
- Identify Use cases of Lakehouse Federation when connected to external sources.

## Sample Questions

These questions are retired from a previous version of the exam. The purpose is to show you the objectives as they are stated on the exam guide, and give you a sample question that aligns to the objective. The exam guide lists the objectives that could be covered on an exam. The best way to prepare for a certification exam is to review the exam outline in the exam guide.

**Question 1**

A data engineer has created a Delta table as part of a data pipeline. Downstream data analysts now need SELECT permission on the Delta table.

Which part of the Databricks Lakehouse Platform can the data engineer use to grant the data analysts the appropriate access?

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

**Question 2**

A dataset has been defined using Delta Live Tables and includes an expectations clause:

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

What is the expected behavior when a batch of data containing data that violates these constraints is processed?

A. Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.
B. Records that violate the expectation are added to the target dataset and recorded as invalid in the event log.
C. Records that violate the expectation cause the job to fail.
D. Records that violate the expectation are added to the target dataset and flagged as invalid in a field added to the target dataset.

**Question 3**

A Delta Live Table pipeline includes two datasets defined using **STREAMING LIVE TABLE**. Three datasets are defined against Delta Lake table sources using **LIVE TABLE**.

The table is configured to run in Development mode using the Triggered Pipeline Mode.

Considering previously unprocessed data exists and all definitions are valid, what is the expected outcome after clicking Start to update the pipeline?

A. All datasets will be updated at set intervals until the pipeline is shut down.The compute resources will persist after the pipeline is stopped to allow for additional testing.
B. All datasets will be updated once and the pipeline will shut down. The compute resources will be terminated.
C. All datasets will be updated at set intervals until the pipeline is shut down.The compute resources will be deployed for the update and terminated when the pipeline is stopped.
D. All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.

**Question 4**

A Delta Live Table pipeline includes two datasets defined using STREAMING LIVE TABLE. Three datasets are defined against Delta Lake table sources using LIVE TABLE.

The table is configured to run in Development mode using the Continuous Pipeline Mode.

Assuming previously unprocessed data exists and all definitions are valid, what is the expected outcome after clicking Start to update the pipeline?

A. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist until the pipeline is shut down.
B. All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.
C. All datasets will be updated once and the pipeline will shut down.The compute resources will be terminated.
D. All datasets will be updated at set intervals until the pipeline is shut down.The compute resources will persist to allow for additional testing.

**Question 5**

A new data engineering team **team** has been assigned to an ELT project. The new data engineering team will need full privileges on the table **sales** to fully manage the project.

Which command can be used to grant full permissions on the database to the new data engineering team?

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;`

**Answers**
Question 1: C
Question 2: B
Question 3: D

Question 4: A
Question 5: D