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

In which scenario will a data team want to utilize cluster pools?

A. An automated report needs to be refreshed as quickly as possible.
B. An automated report needs to be made reproducible.
C. An automated report needs to be version-controlled across multiple collaborators.
D. An automated report needs to be runnable by all stakeholders.

Question 2

A data engineer needs to determine whether to use the built-in Databricks Notebooks versioning or version their project using Databricks Repos.

What is an advantage of using Databricks Repos over the Databricks Notebooks versioning?

A. Databricks Repos supports the use of multiple branches
B. Databricks Repos is wholly housed within the Databricks Data Intelligence Platform
C. Databricks Repos allows users to revert to previous versions of a notebook
D. Databricks Repos provides the ability to comment on specific changes

Question 3

What can be used to simplify and unify siloed data architectures that are specialized for specific use cases?

A. Data lakehouse
B. Data warehouse
C. Data lake
D. RDBMS

Question 4

What is stored in a Databricks customer's cloud account?

A. Data
B. Notebooks
C. Databricks web application
D. Cluster management metadata

Question 5

A data engineer needs to use a Delta table as part of a data pipeline, but they do not know if they have the appropriate permissions.

In which location can the data engineer review their permissions on the table?

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

Question 6

Which file format is used for storing Delta Tables?

A. Parquet
B. AVRO
C. CSV
D. JSON

Question 7

A data architect has determined that a table of the following format is necessary:

employeeId	startDate	avgRating
a1	2009-01-06	5.5
a2	2018-11-21	7.1
…	…	…

Which code block is used by SQL DDL command to create an empty Delta table in the above format regardless of whether a table already exists with this name?

A. CREATE OR REPLACE TABLE table_name (employeeId STRING, startDate DATE, avgRating FLOAT)
B. CREATE TABLE IF NOT EXISTS table_name (employeeId STRING, startDate DATE, avgRating FLOAT)
C. CREATE TABLE table_name AS SELECT employeeId STRING, startDate DATE, avgRating FLOAT
D. CREATE OR REPLACE TABLE table_name WITH COLUMNS (employeeId STRING, startDate DATE, avgRating FLOAT) USING DELTA

Question 8

A data engineer has been given a new record of data:

```
id STRING = 'a1'

rank INTEGER = 6

rating FLOAT = 9.4
```

Which SQL command can be used to append the new record to an existing Delta table `my_table`?

A. `INSERT INTO my_table VALUES ('a1', 6, 9.4)`
B. `UPDATE my_table VALUES ('a1', 6, 9.4)`
C. `INSERT VALUES ('a1', 6, 9.4) INTO my_table`
D. `UPDATE VALUES ('a1', 6, 9.4) my_table`

Question 9

Which two components function in the DB platform architecture's control plane? (Choose 2)

A. Compute
B. Compute Orchestration
C. Unity Catalog
D. Serverless Compute
E. Virtual Machines

Question 10

A data engineer only wants to execute the final block of a Python program if the Python variable `day_of_week` is equal to 1 and the Python variable `review_period` is True.

Which control flow statement should the data engineer use to begin this conditionally executed code block?

A. `if day_of_week == 1 and review_period:`
B. `if day_of_week = 1 and review_period:`
C. `if day_of_week = 1 and review_period = "True"`
D. `if day_of_week = 1 & review_period = "True":`

Answers

Question 1: A

Question 2: A

Question 3: A

Question 4: A

Question 5: A

Question 6: A

Question 7: A

Question 8: A

Question 9: B&C

Question 10: A