databricks

Databricks Exam Guide

Databricks Certified Data Engineer Professional

[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 September 30th, 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 Engineering Professional exam validates a candidate's advanced skills in building, optimising, and maintaining production-grade data engineering solutions on the Databricks Lakehouse Platform. Successful candidates demonstrate expertise across core platform features such as Delta Lake, Unity Catalog, Auto Loader, Lakeflow Definitive Pipelines, Databricks Compute (including serverless) Lakeflow Jobs and the Medallion Architecture. This Certification assesses the ability to design secure, reliable, and cost-effective ETL Pipelines, process complex data from diverse sources using Python and SQL, and apply best practices in schema management, observability, governance, and performance optimization. Candidates are also tested on implementing streaming workloads, orchestrating workflows, leveraging DevOps & CI/CD, and deploying with tools like the Databricks CLI, REST API, and Asset Bundles. Professionals who earn this certification are proven to have the knowledge and hands-on experience required to deliver production-ready data engineering solutions on Databricks, with one or more years of experience on the Lakehouse Platform is strongly recommended.

About the Exam

- Number of items: 59 scored multiple-choice questions
- Time limit: 120 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 1 year 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, and additional time is factored into account for this content.

Recommended Training

- Instructor led Advanced Data Engineering With Databricks
- Self-paced (available in Databricks Academy):
    - Databricks Streaming and LakeFlow Declarative Pipelines nm
    - Databricks Data Privacy
    - Databricks Performance Optimization
    - Automated Deployment with Databricks Asset Bundle

Exam outline

Section 1: Developing Code for Data Processing using Python and SQL

- Using Python and Tools for development
    - Design and implement a scalable Python project structure optimized for Databricks Asset Bundles (DABs), enabling modular development, deployment automation, and CI/CD integration.
    - Manage and troubleshoot external third-party library installations and dependencies in Databricks, including PyPI packages, local wheels, and source archives.
    - Develop User-Defined Functions (UDFs) using Pandas/Python UDF.
- Building and Testing an ETL pipeline with Lakeflow Declarative Pipelines, SQL, and Apache

  Spark on the Databricks platform
    - Build and manage reliable, production-ready data pipelines, for batch and streaming data using Lakeflow Declarative Pipelines and Autoloader.
    - Create and Automate ETL workloads using Jobs via UI/APIs/CLI.
    - Explain the advantages and disadvantages of streaming tables compared to materialized views.
    - Use APPLY CHANGES APIs to simplify CDC in Lakeflow Declarative Pipelines.
    - Compare Spark Structured Streaming and Lakeflow Declarative Pipelines to determine the optimal approach for building scalable ETL pipelines.
    - Create a pipeline component that uses control flow operators (e.g. if/else, foreach, etc.)
    - Choose the appropriate configs for environments and dependencies, high memory for notebook tasks, and auto-optimization to disallow retries.
    - Develop unit and integration tests using assertDataFrameEqual, assertSchemaEqual, DataFrame.transform, and testing frameworks, to ensure code correctness, including a built-in debugger.

Section 2: Data Ingestion & Acquisition:

- Design and implement data ingestion pipelines to efficiently ingest a variety of data

formats including Delta Lake, Parquet, ORC, AVRO, JSON, CSV, XML, Text and Binary from diverse sources such as message buses and cloud storage.

- Create an append-only data pipeline capable of handling both batch and streaming data using Delta.

Section 3: Data Transformation, Cleansing, and Quality

- Write efficient Spark SQL and PySpark code to apply advanced data transformations, including window functions, joins, and aggregations, to manipulate and analyze large Datasets.
- Develop a quarantining process for bad data with Lakeflow Declarative Pipelines or autoloader in classic jobs.

Section 4: Data Sharing and Federation

- Demonstrate delta sharing securely between Databricks deployments using Databricks to Databricks Sharing(D2D) or to external platforms using open sharing protocol(D2O).
- Configure Lakehouse Federation with proper governance across supported source Systems.
- Use Delta Share to share live data from Lakehouse to any computing platform.

Section 5: Monitoring and Alerting

- Monitoring
    - Use system tables for observability over resource utilization, cost, auditing and workload monitoring.
    - Use Query Profiler UI and Spark UI to monitor workloads.
    - Use the Databricks REST APIs/Databricks CLI for monitoring jobs and pipelines.
    - Use Lakeflow Declarative Pipelines Event Logs to monitor pipelines.
- Alerting
    - Use SQL Alerts to monitor data quality.
    - Use the Workflows UI and Jobs API to set up job status and performance issue notifications.

Section 6:Cost & Performance Optimisation

- Understand how / why using Unity Catalog managed tables reduces operation Overhead and maintenance burden.
- Understand delta optimization techniques, such as deletion vectors and liquid clustering.
- Understand the optimization techniques used by Databricks to ensure the performance of queries on large datasets (data skipping, file pruning, etc).
- Apply Change Data Feed (CDF) to address specific limitations of streaming tables and enhance latency.
- Use query profile to analyze the query and identify bottlenecks, such as bad data kipping, inefficient types of joins, data shuffling.

Section 7: Ensuring Data Security and Compliance

- Applying Data Security mechanisms.
    - Use ACLs to secure Workspace Objects, enforcing the principle of least privilege, including enforcing principles like least privilege, policy enforcement.
    - Use row filters and column masks to filter and mask sensitive table data.
    - Apply anonymization and pseudonymization methods such as Hashing, Tokenization, Suppression, and Generalization to confidential data**.**
- Ensuring Compliance
    - Implement a compliant batch & streaming pipeline that detects and applies masking of PII to ensure data privacy.
    - Develop a data purging solution ensuring compliance with data retention policies.

Section 8: Data Governance

- Create and add descriptions/metadata about enterprise data to make it more discoverable.
- Demonstrate understanding of Unity Catalog permission inheritance model.

Section 9: Debugging and Deploying

- Debugging and Troubleshooting
    - Identify pertinent diagnostic information using Spark UI, cluster logs, system tables, and query profiles to troubleshoot errors.
    - Analyze the errors and remediate the failed job runs with job repairs and parameter overrides.
    - Use Lakeflow Declarative Pipelines event logs & the Spark UI to debug Lakeflow Declarative Pipelines and Spark pipelines.
- Deploying CI/CD
    - Build and Deploy Databricks resources using Databricks Asset Bundles.
    - Configure and integrate with Git-based CI/CD workflows using Databricks Git Folders for notebook and code deployment.

Section 10: Data Modelling

- Design and implement scalable data models using Delta Lake to manage large datasets.
- Simplify data layout decisions and optimize query performance using Liquid Clustering.
- Identify the benefits of using liquid Clustering over Partitioning and ZOrder.
- Design Dimensional Models for analytical workloads, ensuring efficient querying and aggregation.

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 with 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
Objective: Understand Delta Lake's catalog- metastore operations and ACID compliance behaviour.

A Delta Lake table was created with the below query:

```
CREATE TABLE prod.sales_by_stor
USING DELTA
LOCATION "/mnt/prod/sales_by_store"
```

Realizing that the original query had a typographical error, the below code was executed:

```
ALTER TABLE prod.sales_by_stor RENAME TO prod.sales_by_store
```

Which result will occur after running the second command?

A. All related files and metadata are dropped and recreated in a single ACID transaction.
B. The table name change is recorded in the Delta transaction log.
C. A new Delta transaction log is created for the renamed table..
D. The table reference in the metastore is updated.

Question 2
Objective: Understand Spark Structured Streaming behaviour and determine the optimal approach for production SLA-ready pipelines.

A Structured Streaming job deployed to production has been experiencing delays during peak hours of the day. At present, during normal execution, each microbatch of data is processed in less than 3 seconds. During peak hours of the day, execution time for each microbatch becomes very inconsistent, sometimes exceeding 30 seconds. The streaming write is currently configured with a trigger interval of 10 seconds.

Holding all other variables constant and assuming records need to be processed in less than 10 seconds, which adjustment will meet the requirement?

A. Use the trigger once option and configure a Databricks job to execute the query every 8 seconds; this ensures all backlogged records are processed with each batch.
B. Decrease the trigger interval to 5 seconds; triggering batches more frequently may prevent records from backing up and large batches from causing spill.
C. Decrease the trigger interval to 5 seconds; triggering batches more frequently allows idle executors to begin processing the next batch while longer running tasks from previous batches finish.
D. The trigger interval cannot be modified without modifying the checkpoint directory; to maintain the current stream state, increase the number of shuffle partitions to maximize parallelism.

Question 3

Objective:Apply anonymization and pseudonymization methods such as Hashing, Tokenization, Suppression, and Generalization to confidential data

The data engineering team is migrating an enterprise system with thousands of tables and views into the Lakehouse. They plan to implement the target architecture using a series of bronze, silver, and gold tables. Bronze tables will almost exclusively be used by production data engineering workloads, while silver tables will be used to support both data engineering and machine learning workloads. Gold tables will largely serve business intelligence and reporting purposes. While personal identifying information (PII) exists in all tiers of data, pseudonymization and anonymization rules are in place for all data at the silver and gold levels.

The organization is interested in reducing security concerns while maximizing the ability to collaborate across diverse teams.

Which statement exemplifies best practices for implementing this system?

A. Isolating tables in separate databases based on data quality tiers allows for easy permissions management through database ACLs and allows physical separation of default storage locations for managed tables.
B. Storing all production tables in a single database provides a unified view of all data assets available throughout the Lakehouse, simplifying discoverability by granting all users view privileges on this database.
C. Because databases on Databricks are merely a logical construct, choices around database organization do not impact security or discoverability in the Lakehouse.
D. Working in the default Databricks database provides the greatest security when working with managed tables, as these will be created in the DBFS root.

Question 4
Objective: Design and implement scalable data models using Delta Lake to manage large datasets.

A Delta Lake table representing metadata about content posts from users has the following `schema:`

```
user_id LONG, post_text STRING, post_id STRING, longitude FLOAT,
latitude FLOAT, post_time TIMESTAMP, date DATE
```

Based on the above schema, which column is a good candidate for partitioning the Delta Table?

A. `post_id`
B. `post_time`
C. `date`
D. `user_id`

Question 5
Objective: Demonstrate understanding of Unity Catalog permission inheritance model

A table named user_ltv is being used to create a view that will be used by data analysts on various teams. Users in the workspace are configured into groups, which are used for setting up data access using ACLs.

The **user_ltv** table has the following schema:

```
email STRING, age INT, ltv INT
```

The following view definition is executed:

```
CREATE VIEW email_ltv AS
SELECT
CASE WHEN
  is_member('marketing') THEN email
  ELSE 'REDACTED'
END AS email,
ltv
FROM user_ltv
```

An analyst who is not a member of the marketing group executes the following query:

```
SELECT * FROM email_ltv
```

What will be the result of this query?

A. Only the **email** and **ltv** columns will be returned; the email column will contain the string "**REDACTED**" in each row.
B. Three columns will be returned, but one column will be named **"REDACTED"** and contain only null values.
C. Only the **email** and **ltv** columns will be returned; the email column will contain all null values.
D. The **email** and **ltv** columns will be returned with the values in **user_ltv**.

Question 6:
Objective- Choose the appropriate configs for environments and dependencies, high memory for notebook tasks and auto-optimization to disallow retries.

The business reporting team requires that data for their dashboards be updated every hour. The total processing time for the pipeline that extracts, transforms, and loads the data for their pipeline runs in 10 minutes.

Assuming normal operating conditions, which configuration will meet their service-level agreement requirements with the lowest cost?

A. Schedule a job to execute the pipeline once an hour on a dedicated interactive cluster.
B. Schedule a job to execute the pipeline once an hour on a new job cluster.
C. Schedule a Structured Streaming job with a trigger interval of 60 minutes.
D. Configure a job that executes every time new data lands in a given directory.

Question 7:

Objective- Understand the Notebook development environment, variable management and creating secure, configurable code.

The security team is exploring whether or not the Databricks secrets module can be leveraged for connecting to an external database.

After testing the code with all Python variables being defined with strings, they upload the password to the secrets module and configure the correct permissions for the currently active user. They then modify their code to the following (leaving all other variables unchanged).

```
password = dbutils.secrets.get(scope="db_creds", key="jdbc_password")

print(password)

df = (spark
  .read
  .format("jdbc")
  .option("url", connection)
  .option("dbtable", tablename)
  .option("user", username)
  .option("password", password)
  )
```

Which statement describes what will happen when the above code is executed?

A. The connection to the external table will succeed; the string **"REDACTED"** will be printed.
B. The connection to the external table will succeed; the string value of **password** will printed in plain text.
C.. An interactive input box will appear in the notebook; if the right password is provided, the connection will succeed and the password will be printed in plain text.
D. An interactive input box will appear in the notebook; if the right password is provided, the connection will succeed and the encoded password will be saved to DBFS.

Question 8:
Objective:Understand the optimization techniques used by Databricks to ensure performance of queries on large datasets (data skipping, file pruning, etc)

A data ingestion task requires a one-TB JSON dataset to be written out to Parquet with a target part-file size of 512 MB. Because Parquet is being used instead of Delta Lake, built-in file-sizing features such as Auto-Optimize & Auto-Compaction cannot be used.

Which strategy will yield the best performance without shuffling data?

A. Ingest the data, execute the narrow transformations, repartition to 2,048 partitions (1TB*1024*1024/512), and then write to parquet.
B. Set **spark.sql.adaptive.advisoryPartitionSizeInBytes** to 512 MB bytes, ingest the data, execute the narrow transformations, coalesce to 2,048 partitions (1TB*1024*1024/512), and then write to parquet.
C. Set **spark.sql.files.maxPartitionBytes** to 512 MB, ingest the data, execute the narrow transformations, and then write to parquet.
D. Set **spark.sql.shuffle.partitions** to 2,048 partitions (1TB*1024*1024/512), ingest the data, execute the narrow transformations, optimize the data by sorting it (which automatically repartitions the data), and then write to parquet.

Question 9:
Objective: Apply Delta Lake clone to learn how shallow and deep clone interact with source/target tables.

The marketing team is looking to share data in an aggregate table with the sales organization, but the field names used by the teams do not match, and a number of marketing-specific fields have not been approved for the sales org.

Which solution addresses the situation while emphasizing simplicity?

A. Create a view on the marketing table selecting only those fields approved for the sales team; alias the names of any fields that should be standardized to the sales naming conventions.
B. Create a new table with the required schema and use Delta Lake's DEEP CLONE

functionality to sync up changes committed to one table to the corresponding table.

C. Use a CTAS statement to create a derivative table from the marketing table; configure a production job to propagate changes.
D. Add a parallel table write to the current production pipeline, updating a new sales table that varies as required from the marketing table.

Question 10:
Objective: Create a multi-task job with multiple dependencies

A Databricks job has been configured with three tasks, each of which is a Databricks notebook. Task A does not depend on other tasks. Tasks B and C run in parallel, with each having a serial dependency on task A.

What will be the resulting state if tasks A and B complete successfully but task C fails during a scheduled run?

A. All logic expressed in the notebook associated with tasks A and B will have been successfully completed; some operations in task C may have completed successfully.
B. Unless all tasks complete successfully, no changes will be committed to the Lakehouse; because task C failed, all commits will be rolled back automatically.
C. All logic expressed in the notebook associated with tasks A and B will have been successfully completed; any changes made in task C will be rolled back due to task failure.
D. Because all tasks are managed as a dependency graph, no changes will be committed to the Lakehouse until all tasks have successfully been completed.

Answers
Question 1: D
Question 2: B
Question 3: A
Question 4: C
Question 5: A
Question 6: B
Question 7: A
Question 8: C
Question 9: A
Question 10:A