databricks

Databricks Exam Guide

Databricks Certified Data Analyst 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 version as of August 1, 2024. 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 Analyst Associate certification exam assesses an individual's ability r to use the Databricks SQL service to complete introductory data analysis tasks. This includes an understanding of the Databricks SQL service and its capabilities, an ability to manage data with Databricks tools following best practices, using SQL to complete data tasks in the Lakehouse, creating production-grade data visualizations and dashboards, and developing analytics applications to solve common data analytics problems. Individuals who pass this certification exam can be expected to complete basic data analysis tasks using Databricks SQL and its associated capabilities.

About the Exam

- Number of items: 45 scored 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, and additional time is factored into account for this content.

Recommended Training

- Instructor-led: Data Analysis with Databricks
- Self-paced (available in Databricks Academy): Data Analysis with Databricks

Exam Outline

Section 1: Databricks SQL

- Describe the key audience and side audiences for Databricks SQL.
- Describe that a variety of users can view and run Databricks SQL dashboards as stakeholders.
- Describe the benefits of using Databricks SQL for in-Lakehouse platform data processing.
- Describe how to complete a basic Databricks SQL query.
- Identify Databricks SQL queries as a place to write and run SQL code.
- Identify the information displayed in the schema browser from the Query Editor page.
- Identify Databricks SQL dashboards as a place to display the results of multiple queries at once.
- Describe how to complete a basic Databricks SQL dashboard.
- Describe how dashboards can be configured to automatically refresh.
- Describe the purpose of Databricks SQL endpoints/warehouses.
- Identify Serverless Databricks SQL endpoint/warehouses as a quick-starting option.
- Describe the trade-off between cluster size and cost for Databricks SQL endpoints/warehouses.
- Identify Partner Connect as a tool for implementing simple integrations with a number of other data products.
- Describe how to connect Databricks SQL to ingestion tools like Fivetran.
- Identify the need to be set up with a partner to use it for Partner Connect.
- Identify small-file upload as a solution for importing small text files like lookup tables and quick data integrations.
- Import from object storage using Databricks SQL.
- Identify that Databricks SQL can ingest directories of files of the files are the same type.
- Describe how to connect Databricks SQL to visualization tools like Tableau, Power BI, and Looker.
- Identify Databricks SQL as a complementary tool for BI partner tool workflows.
- Describe the medallion architecture as a sequential data organization and pipeline system of progressively cleaner data.
- Identify the gold layer as the most common layer for data analysts using Databricks SQL.
- Describe the cautions and benefits of working with streaming data.
- Identify that the Lakehouse allows the mixing of batch and streaming workloads.

Section 2: Data Management

- Describe Delta Lake as a tool for managing data files.
- Describe that Delta Lake manages table metadata.
- Identify that Delta Lake tables maintain history for a period of time.
- Describe the benefits of Delta Lake within the Lakehouse.
- Describe persistence and scope of tables on Databricks.
- Compare and contrast the behavior of managed and unmanaged tables.
- Identify whether a table is managed or unmanaged.
- Explain how the LOCATION keyword changes the default location of database contents.
- Use Databricks to create, use, and drop databases, tables, and views.
- Describe the persistence of data in a view and a temp view
- Compare and contrast views and temp views.
- Explore, preview, and secure data using Data Explorer.
- Use Databricks to create, drop, and rename tables.
- Identify the table owner using Data Explorer.
- Change access rights to a table using Data Explorer.
- Describe the responsibilities of a table owner.
- Identify organization-specific considerations of PII data

Section 3: SQL in the Lakehouse

- Identify a query that retrieves data from the database with specific conditions
- Identify the output of a SELECT query
- Compare and contrast MERGE INTO, INSERT TABLE, and COPY INTO.
- Simplify queries using subqueries.
- Compare and contrast different types of JOINs.
- Aggregate data to achieve a desired output.
- Manage nested data formats and sources within tables.
- Use cube and roll-up to aggregate a data table.
- Compare and contrast roll-up and cube.
- Use windowing to aggregate time data.
- Identify a benefit of having ANSI SQL as the standard in the Lakehouse.
- Identify, access, and clean silver-level data.
- Utilize query history and caching to reduce development time and query latency.
- Optimize performance using higher-order Spark SQL functions.
- Create and apply UDFs in common scaling scenarios.

Section 4: Data Visualization and Dashboarding

- Create basic, schema-specific visualizations using Databricks SQL.
- Identify which types of visualizations can be developed in Databricks SQL (table, details, counter, pivot).

- Explain how visualization formatting changes the reception of a visualization
- Describe how to add visual appeal through formatting
- Identify that customizable tables can be used as visualizations within Databricks SQL.
- Describe how different visualizations tell different stories.
- Create customized data visualizations to aid in data storytelling.
- Create a dashboard using multiple existing visualizations from Databricks SQL Queries.
- Describe how to change the colors of all of the visualizations in a dashboard.
- Describe how query parameters change the output of underlying queries within a dashboard
- Identify the behavior of a dashboard parameter
- Identify the use of the "Query Based Dropdown List" as a way to create a query parameter from the distinct output of a different query.
- Identify the method for sharing a dashboard with up-to-date results.
- Describe the pros and cons of sharing dashboards in different ways
- Identify that users without permission to all queries, databases, and endpoints can easily refresh a dashboard using the owner's credentials.
- Describe how to configure a refresh schedule
- Identify what happens if a refresh rate is less than the Warehouse's "Auto Stop"
- Describe how to configure and troubleshoot a basic alert
- Describe how notifications are sent when alerts are set up based on the configuration

Section 5: Analytics applications

- Compare and contrast discrete and continuous statistics.
- Describe descriptive statistics.
- Describe key moments of statistical distributions.
- Compare and contrast key statistical measures.
- Describe data enhancement as a common analytics application.
- Enhance data in a common analytics application.
- Identify a scenario in which data enhancement would be beneficial.
- Describe the blending of data between two source applications.
- Identify a scenario in which data blending would be beneficial.
- Perform last-mile ETL as project-specific data enhancement.

Sample Questions

These questions are retired from a previous version of the exam. The purpose is to show you 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

Objective: *Identify the benefits of using Databricks SQL for business intelligence (BI) analytics projects over using third-party BI tools?*

A data analyst is trying to determine whether to develop their dashboard in Databricks SQL or a partner business intelligence (BI) tool like Tableau, Power BI, or Looker.

When is it advantageous to use Databricks SQL instead of using third-party BI tools to develop the dashboard?

A. When the data being transformed as part of the visualizations is very large
B. When the visualizations require custom formatting
C. When the visualizations require production-grade, customizable branding
D. When the data being transformed is in table format

Question 2
Objective: Aggregate data columns using SQL functions to answer defined business questions.

A data analyst has been asked to count the number of customers in each region and has written the following query:

```
SELECT region, count(*) AS number_of_customers
FROM customers
ORDER BY region;
```

What is the mistake in the query?

A. The query is selecting region, but region should only occur in the **`ORDER BY`** clause.
B. The query is missing a **`GROUP BY region`** clause.
C. The query is using **`ORDER BY`**, which is not allowed in an aggregation.
D. The query is using **`count(*)`**, which will count all the customers in the customers table, no matter the region.

Question 3
Objective: Identify code blocks that can be used to create user-defined functions

A data analyst has created a user-defined function using the following line of code:

```
CREATE FUNCTION price(spend DOUBLE, units DOUBLE)
RETURNS DOUBLE
RETURN spend / units;
```

Which code block can be used to apply this function to the **`customer_spend`** and **`customer_units`** columns of the table **`customer_summary`** to create column **`customer_price`**?

A.
```
SELECT function(price(customer_spend, customer_units)) AS
customer_price
FROM customer_summary
```
B.
```
SELECT double(price(customer_spend, customer_units)) AS
customer_price
FROM customer_summary
```
C.
```
SELECT price
FROM customer_summary
```
D.
```
SELECT PRICE customer_spend, customer_units AS customer_price
FROM customer_summary
```
E.
```
SELECT price(customer_spend, customer_units) AS customer_price
FROM customer_summary
```

Question 4

Objective: Automate a refresh schedule for a query.

Where in the Databricks SQL workspace can a data analyst configure a refresh schedule for a query when the query is not attached to a dashboard or alert?

A. The Dashboard Editor
B. The Visualization Editor
C. The Query Editor
D. SQL Warehouse
E. Data Explorer

Question 5

Objective: Define different types of data augmentation.

A data analyst is working with gold-layer tables to complete an ad-hoc project. A stakeholder has provided the analyst with an additional dataset that can be used to augment the gold-layer tables already in use.

Which term is used to describe this data augmentation?

A. Data testing
B. Last-mile ETL
C. Ad-hoc improvements
D. Data enhancement
E. Last-mile dashboarding

Answers

Question 1: A

Question 2: B

Question 3: E

Question 4: C
Question 5: D