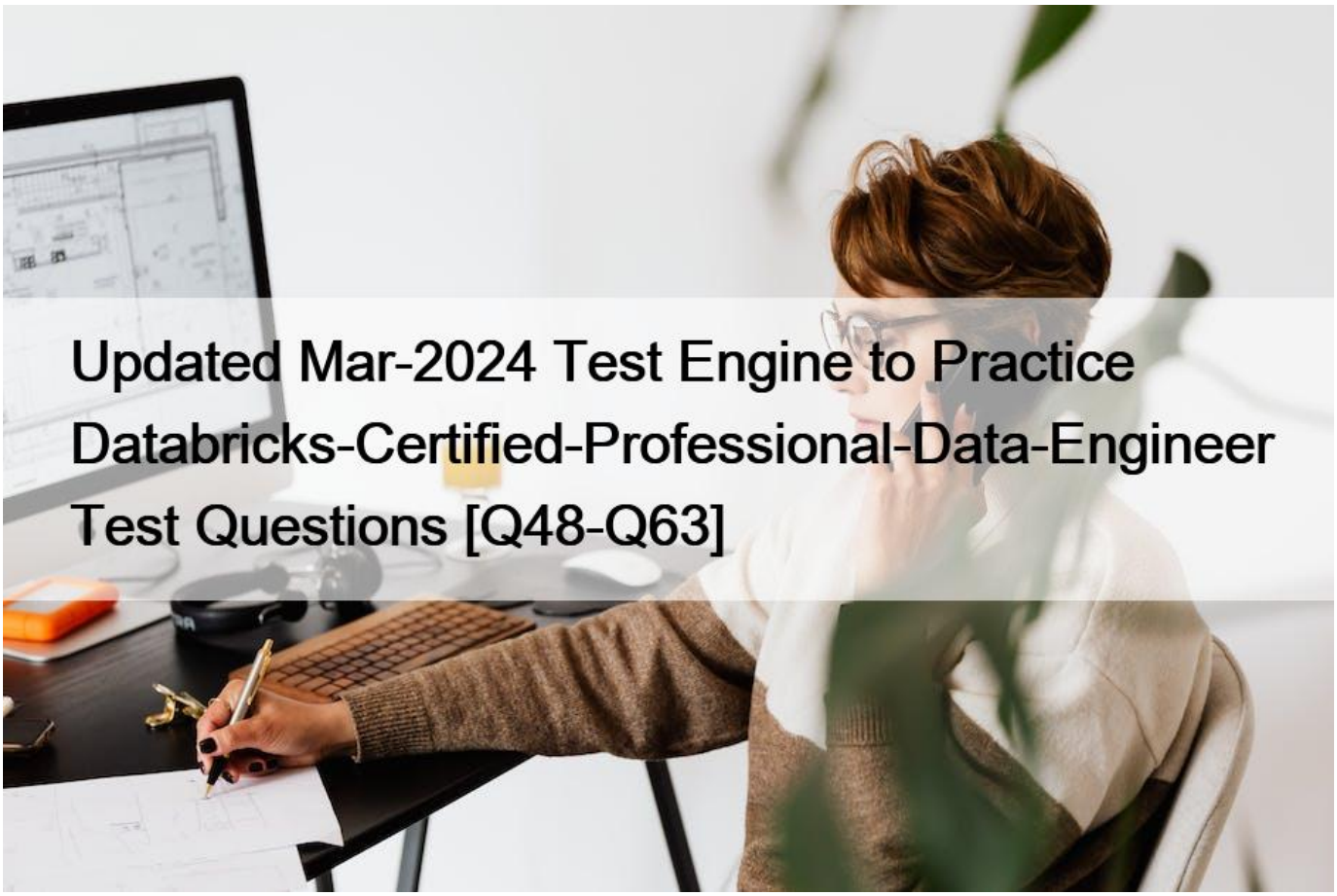
Updated Mar-2024 Test Engine to Practice Databricks-Certified-Professional-Data-Engineer Test Questions [Q48-Q63



Updated Mar-2024 Test Engine to Practice Databricks-Certified-Professional-Data-Engineer Test Questions
Databricks-Certified-Professional-Data-Engineer Real Exam Questions Test Engine Dumps Training With 84 Questions

Databricks Certified Professional Data Engineer exam is a rigorous and comprehensive assessment of a candidate's skills in designing, building, and maintaining data pipelines on the Databricks platform. Databricks-Certified-Professional-Data-Engineer exam covers a wide range of topics, including data storage and retrieval, data processing, data transformation, and data visualization. Candidates are tested on their ability to design and implement scalable and reliable data architectures, as well as their proficiency in troubleshooting and optimizing data pipelines.

Q48. In order to use Unity catalog features, which of the following steps needs to be taken on man-aged/external tables in the Databricks workspace?
* Enable unity catalog feature in workspace settings
* Migrate/upgrade objects in workspace managed/external tables/view to unity catalog
* Upgrade to DBR version 15.0
* Copy data from workspace to unity catalog

* Upgrade workspace to Unity catalog
Explanation

Upgrade tables and views to Unity Catalog – Azure Databricks | Microsoft Docs Managed table: Upgrade a managed to Unity Catalog External table: Upgrade an external table to Unity Catalog

Q49. Which of the following is true of Delta Lake and the Lakehouse?
* Because Parquet compresses data row by row. strings will only be compressed when a character is repeated multiple times.
* Delta Lake automatically collects statistics on the first 32 columns of each table which are leveraged in data skipping based on query filters.
* Views in the Lakehouse maintain a valid cache of the most recent versions of source tables at all times.
* Primary and foreign key constraints can be leveraged to ensure duplicate values are never entered into a dimension table.
* Z-order can only be applied to numeric values stored in Delta Lake tables
Explanation

https://docs.delta.io/2.0.0/table-properties.html

Delta Lake automatically collects statistics on the first 32 columns of each table, which are leveraged in data skipping based on query filters1. Data skipping is a performance optimization technique that aims to avoid reading irrelevant data from the storage layer1. By collecting statistics such as min/max values, null counts, and bloom filters, Delta Lake can efficiently prune unnecessary files or partitions from the query plan1. This can significantly improve the query performance and reduce the I/O cost.

The other options are false because:

Parquet compresses data column by column, not row by row2. This allows for better compression ratios, especially for repeated or similar values within a column2.

Views in the Lakehouse do not maintain a valid cache of the most recent versions of source tables at all times3. Views are logical constructs that are defined by a SQL query on one or more base tables3. Views are not materialized by default, which means they do not store any data, but only the query definition3. Therefore, views always reflect the latest state of the source tables when queried3.

However, views can be cached manually using the CACHE TABLE or CREATE TABLE AS SELECT commands.

Primary and foreign key constraints can not be leveraged to ensure duplicate values are never entered into a dimension table. Delta Lake does not support enforcing primary and foreign key constraints on tables. Constraints are logical rules that define the integrity and validity of the data in a table. Delta Lake relies on the application logic or the user to ensure the data quality and consistency.

Z-order can be applied to any values stored in Delta Lake tables, not only numeric values. Z-order is a technique to optimize the layout of the data files by sorting them on one or more columns. Z-order can improve the query performance by clustering related values together and enabling more efficient data skipping. Z-order can be applied to any column that has a defined ordering, such as numeric, string, date, or boolean values.

References: Data Skipping, Parquet Format, Views, [Caching], [Constraints], [Z-Ordering]

Q50. You are working on a marketing team request to identify customers with the same information between two tables CUSTOMERS_2021 and CUSTOMERS_2020 each table contains 25 columns with the same schema, You are looking to identify rows that match between two tables across all columns, which of the following can be used to perform in SQL
* 1.SELECT * FROM CUSTOMERS_2021

2. UNION

3.SELECT * FROM CUSTOMERS_2020
* 1.SELECT * FROM CUSTOMERS_2021

2. UNION ALL

3.SELECT * FROM CUSTOMERS_2020
* 1.SELECT * FROM CUSTOMERS_2021 C1

2.INNER JOIN CUSTOMERS_2020 C2

3.ON C1.CUSTOMER_ID = C2.CUSTOMER_ID
* 1.SELECT * FROM CUSTOMERS_2021

2. INTERSECT

3.SELECT * FROM CUSTOMERS_2020
* 1.SELECT * FROM CUSTOMERS_2021

2.EXCEPT

3.SELECT * FROM CUSTOMERS_2020
Explanation

Answer is,

1.SELECT * FROM CUSTOMERS_2021

2. INTERSECT

3.SELECT * FROM CUSTOMERS_2020

To compare all the rows between both the tables across all the columns using intersect will help us achieve that, an inner join is only going to check if the same column value exists across both the tables on a single column.

INTERSECT [ALL | DISTINCT]

*Returns the set of rows which are in both subqueries.

If ALL is specified a row that appears multiple times in the subquery1 as well as in subquery will be returned multiple times.

If DISTINCT is specified the result does not contain duplicate rows. This is the default.

Q51. A data engineering team needs to query a Delta table to extract rows that all meet the same condi-tion.

However, the team has noticed that the query is running slowly. The team has already tuned the size of the

data files. Upon investigating, the team has concluded that the rows meeting the condition are sparsely located

throughout each of the data files.

Based on the scenario, which of the following optimization techniques could speed up the query?
* Tuning the file size
* Bin-packing
* Data skipping
* Write as a Parquet file
* Z-Ordering

Q52. The data architect has mandated that all tables in the Lakehouse should be configured as external Delta Lake tables.

Which approach will ensure that this requirement is met?
* When the workspace is being configured, make sure that external cloud object storage has been mounted.
* Whenever a table is being created, make sure that the location keyword is used.
* Whenever a database is being created, make sure that the location keyword is used
* When tables are created, make sure that the external keyword is used in the create table statement.
* When configuring an external data warehouse for all table storage. leverage Databricks for all ELT.

Q53. Which of the following commands can be used to query a delta table?
* 1.%python

2.spark.sql(“select * from table_name”)
* 1.%sql

2.Select * from table_name
* Both A & B

(Correct)
* 1.%python

2.execute.sql(“select * from table”)
* 1.%python

2.delta.sql(“select * from table”)
Explanation

The answer is both options A and B

Options C and D are incorrect because there is no command in Spark called execute.sql or delta.sql

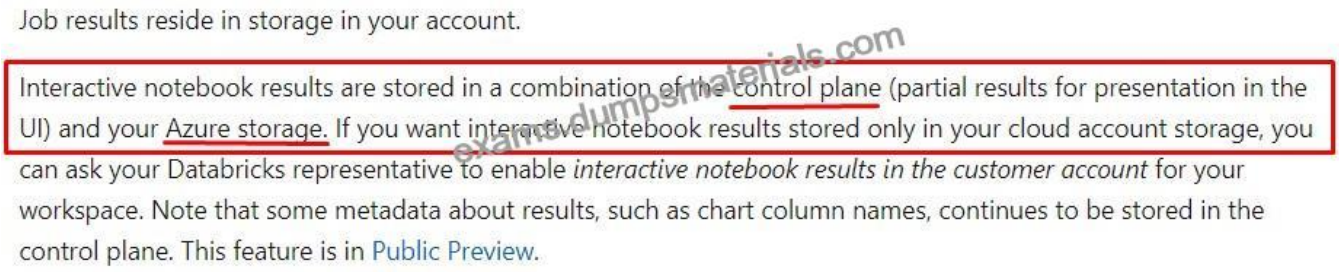
Q54. Where are Interactive notebook results stored in Databricks product architecture?
* Data plane
* Control plane
* Data and Control plane
* JDBC data source
* Databricks web application
Explanation

The answer is Data and Control plane,

Only Job results are stored in Data Plane(your storage), Interactive notebook results are stored in a combination of the control plane

(partial results for presentation in the UI) and customer storage.

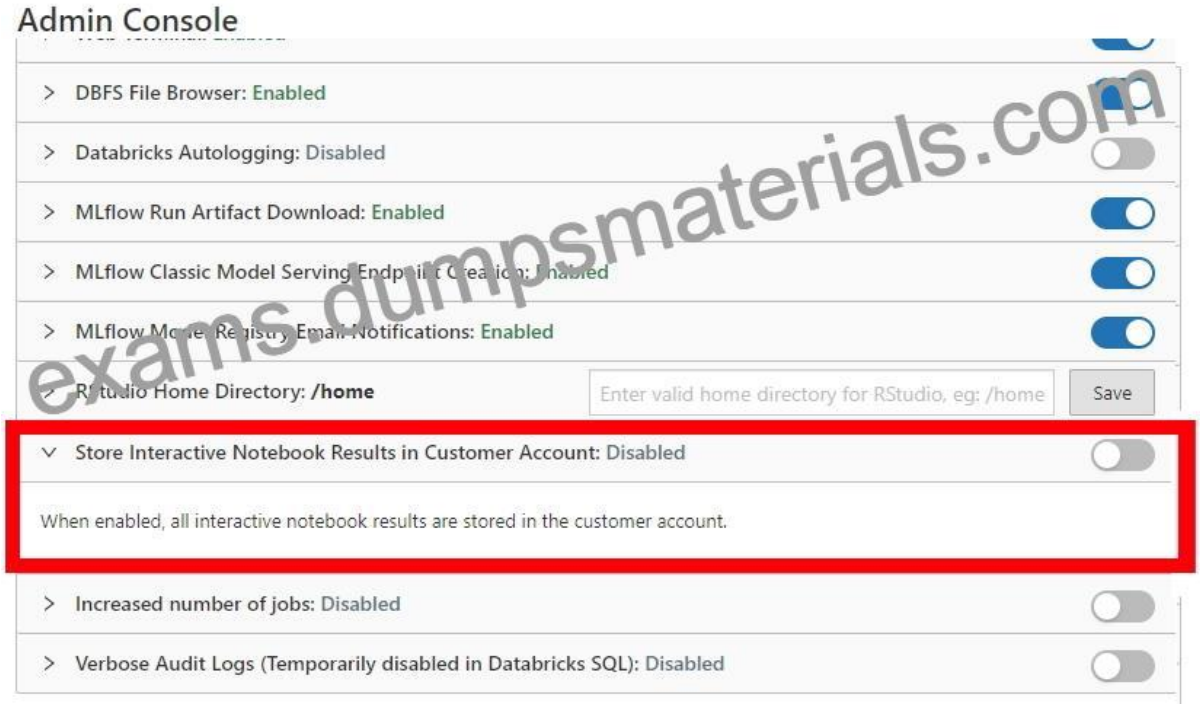
https://docs.microsoft.com/en-us/azure/databricks/getting-started/overview#–high-level-architecture Snippet from the above documentation, Graphical user interface, application Description automatically generated



How to change this behavior?

You can change this behavior using Workspace/Admin Console settings for that workspace, once enabled all of the interactive results are stored in the customer account(data plane) except the new notebook visualization feature Databricks has recently introduced, this still stores some metadata in the control pane irrespective of the below settings. please refer to the documentation for more details.

Graphical user interface, text, application, email Description automatically generated



Why is this important to know?

I recently worked on a project where we had to deal with sensitive information of customers and we had a security requirement that

all of the data need to be stored in the data plane including notebook results.

Q55. Question-26. There are 5000 different color balls, out of which 1200 are pink color. What is the maximum

likelihood estimate for the proportion of “pink” items in the test set of color balls?
* 2.4
* 24 0
* .24
* .48
* 4.8

Explanation

Given no additional information, the MLE for the probability of an item in the test set is exactly its frequency

in the training set. The method of maximum likelihood corresponds to many well-known estimation methods

in statistics. For example, one may be interested in the heights of adult female penguins, but be unable to

measure the height of every single penguin in a population due to cost or time constraints. Assuming that the

heights are normally (Gaussian) distributed with some unknown mean and variance, the mean and variance

can be estimated with MLE while only knowing the heights of some sample of the overall population. MLE

would accomplish this by taking the mean and variance as parameters and finding particular parametric values

that make the observed results the most probable (given the model).

In general, for a fixed set of data and underlying statistical model the method of maximum likelihood selects

the set of values of the model parameters that maximizes the likelihood function. Intuitively, this maximizes

the “agreement” of the selected model with the observed data, and for discrete random variables it indeed

maximizes the probability of the observed data under the resulting distribution. Maximum-likelihood

estimation gives a unified approach to estimation, which is well-defined in the case of the normal distribution

and many other problems. However in some complicated problems, difficulties do occur: in such problems,

maximum-likelihood estimators are unsuitable or do not exist.

Q56. The data engineering team is using a SQL query to review data completeness every day to monitor the ETL job, and query output is being used in multiple dashboards which of the following ap-proaches can be used to set up a schedule and automate this process?
* They can schedule the query to run every day from the Jobs UI.
* They can schedule the query to refresh every day from the query’s page in Databricks SQL
* They can schedule the query to run every 12 hours from the Jobs UI.
* They can schedule the query to refresh every day from the SQL endpoint’s page in Databricks SQL.
* They can schedule the query to refresh every 12 hours from the SQL endpoint’s page in Databricks SQL

Explanation

The answer is They can schedule the query to refresh every 12 hours from the SQL endpoint’s page in Databricks SQL, The query pane view in Databricks SQL workspace provides the ability to add or edit and schedule individual queries to run.

You can use scheduled query executions to keep your dashboards updated or to enable routine alerts. By default, your queries do not have a schedule.

Note

If your query is used by an alert, the alert runs on its own refresh schedule and does not use the query schedule.

To set the schedule:

* Click the query info tab.

* Graphical user interface, text, application, email Description automatically generated

* Click the link to the right of Refresh Schedule to open a picker with schedule intervals.

* Graphical user interface, application Description automatically generated

* 3.Set the schedule.

* The picker scrolls and allows you to choose:

* *An interval: 1-30 minutes, 1-12 hours, 1 or 30 days, 1 or 2 weeks

* *A time. The time selector displays in the picker only when the interval is greater than 1 day and the day selection is greater than 1 week. When you schedule a specific time, Databricks SQL takes input in your computer’s timezone and converts it to UTC. If you want a query to run at a certain time in UTC, you must adjust the picker by your local offset. For example, if you want a query to execute at 00:00 UTC each day, but your current timezone is PDT (UTC-7), you should select 17:00 in the picker:

* Graphical user interface Description automatically generated

Q57. A data engineering manager has noticed that each of the queries in a Databricks SQL dashboard takes a few

minutes to update when they manually click the “Refresh” button. They are curious why this might be

occurring, so a team member provides a variety of reasons on why the delay might be occurring.

Which of the following reasons fails to explain why the dashboard might be taking a few minutes to update?
* The queries attached to the dashboard might take a few minutes to run under normal circumstances
* The queries attached to the dashboard might all be connected to their own, unstarted Databricks clusters
* The Job associated with updating the dashboard might be using a non-pooled endpoint
* The SQL endpoint being used by each of the queries might need a few minutes to start up
* The queries attached to the dashboard might first be checking to determine if new data is available

Q58. Direct query on external files limited options, create external tables for CSV files with header and pipe delimited CSV files, fill in the blanks to complete the create table statement CREATE TABLE sales (id int, unitsSold int, price FLOAT, items STRING)

_____

_____

LOCATION “dbfs:/mnt/sales/*.csv”
* FORMAT CSV

OPTIONS (“true”,”|”)
* USING CSV

TYPE (“true”,”|”)
* USING CSV

OPTIONS (header =”true”, delimiter = “|”)

(Correct)
* FORMAT CSV

FORMAT TYPE (header =”true”, delimiter = “|”)
* FORMAT CSV

TYPE (header =”true”, delimiter = “|”)
Explanation

Answer is

USING CSV

OPTIONS (header =”true”, delimiter = “|”)

Here is the syntax to create an external table with additional options

CREATE TABLE table_name (col_name1 col_typ1,..)

USING data_source

OPTIONS (key=’value’, key2=vla2)

LOCATION = “/location”

Q59. Suppose there are three events then which formula must always be equal to P(E1|E2,E3)?
* P(E1,E2,E3)P(E1)/P(E2:E3)
* P(E1,E2;E3)/P(E2,E3)
* P(E1,E2|E3)P(E2|E3)P(E3)
* P(E1,E2|E3)P(E3)
* P(E1,E2,E3)P(E2)P(E3)
Explanation

This is an application of conditional probability: P(E1,E2)=P(E1|E2)P(E2). so

P(E1|E2) = P(E1.E2)/P(E2)

P(E1,E2,E3)/P(E2,E3)

If the events are A and B respectively, this is said to be “the probability of A given B”

It is commonly denoted by P(A|B):or sometimes PB(A). In case that both “A” and “B” are categorical

variables, conditional probability table is typically used to represent the conditional probability.

Q60. A junior member of the data engineering team is exploring the language interoperability of Databricks notebooks. The intended outcome of the below code is to register a view of all sales that occurred in countries on the continent of Africa that appear in thegeo_lookuptable.

Before executing the code, runningSHOWTABLESon the current database indicates the database contains only two tables:geo_lookupandsales.

Which statement correctly describes the outcome of executing these command cells in order in an interactive notebook?
* Both commands will succeed. Executing show tables will show that countries at and sales at have been registered as views.
* Cmd 1 will succeed. Cmd 2 will search all accessible databases for a table or view named countries af: if this entity exists, Cmd 2 will succeed.
* Cmd 1 will succeed and Cmd 2 will fail, countries at will be a Python variable representing a PySpark DataFrame.
* Both commands will fail. No new variables, tables, or views will be created.
* Cmd 1 will succeed and Cmd 2 will fail, countries at will be a Python variable containing a list of strings.
Explanation

This is the correct answer because Cmd 1 is written in Python and uses a list comprehension to extract the country names from the geo_lookup table and store them in a Python variable named countries af. This variable will contain a list of strings, not a PySpark DataFrame or a SQL view. Cmd 2 is written in SQL and tries to create a view named sales af by selecting from the sales table where city is in countries af. However, this command will fail because countries af is not a valid SQL entity and cannot be used in a SQL query. To fix this, a better approach would be to use spark.sql() to execute a SQL query in Python and pass the countries af variableas a parameter. Verified References: [Databricks Certified Data Engineer Professional], under

“Language Interoperability” section; Databricks Documentation, under “Mix languages” section.

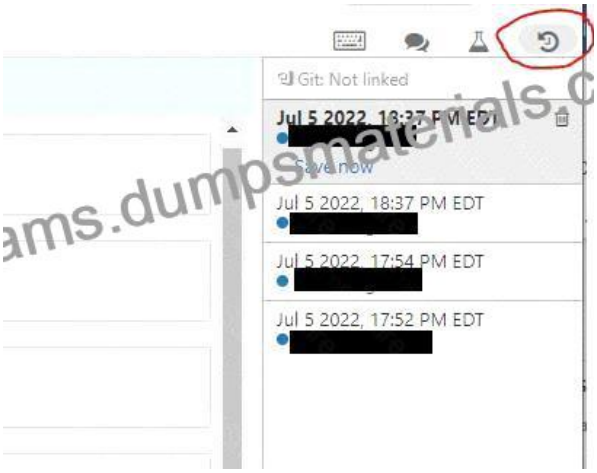
Q61. You noticed that colleague is manually copying the notebook with _bkp to store the previous ver-sions, which of the following

feature would you recommend instead.
* Databricks notebooks support change tracking and versioning
* Databricks notebooks should be copied to a local machine and setup source control locally to version the notebooks
* Databricks notebooks can be exported into dbc archive files and stored in data lake
* Databricks notebook can be exported as HTML and imported at a later time

Explanation

Answer is Databricks notebooks support automatic change tracking and versioning.

When you are editing the notebook on the right side check version history to view all the changes, every change you are making is captured and saved.



Q62. A data engineer has configured a Structured Streaming job to read from a table, manipulate the data, and then

perform a streaming write into a new table. The code block used by the data engineer is below:

1. (spark.table(“sales”)

2. .withColumn(“avg_price”, col(“sales”) / col(“units”))

3. .writeStream

4. .option(“checkpointLocation”, checkpointPath)

5. .outputMode(“complete”)

6. ._____

7. .table(“new_sales”)

8.)

If the data engineer only wants the query to execute a single micro-batch to process all of the available data,

which of the following lines of code should the data engineer use to fill in the blank?
* .processingTime(1)
* .processingTime(“once”)
* .trigger(processingTime=”once”)
* .trigger(once=True)
* .trigger(continuous=”once”)

Q63. Which of the following benefits does Delta Live Tables provide for ELT pipelines over standard data pipelines

that utilize Spark and Delta Lake on Databricks?
* The ability to write pipelines in Python and/or SQL
* The ability to declare and maintain data table dependencies
* The ability to automatically scale compute resources
* The ability to access previous versions of data tables
* The ability to perform batch and streaming queries

Databricks Certified Professional Data Engineer Certification Exam can be attempted by professionals and students who have experience in data engineering, data management, ETL, and data processing. The preparation for the exam can be done via online training courses such as the Databricks Data Engineering Certification Preparation Course, the online Databricks Documentation, and different study materials such as books and videos from verified training providers.

Databricks-Certified-Professional-Data-Engineer Actual Questions Answers PDF 100% Cover Real Exam Questions:
https://www.dumpsmaterials.com/Databricks-Certified-Professional-Data-Engineer-real-torrent.html]