Free Questions for Databricks-Certified-Data-Engineer-Associate by ebraindumps

Shared by Boone on 22-07-2024

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

A data engineer wants to create a new table containing the names of customers who live in France.

They have written the following command:

CREATE TABLE customersInFrance

_____ AS

SELECT id,

firstName,

lastName

FROM customerLocations

WHERE country = 'FRANCE';

A senior data engineer mentions that it is organization policy to include a table property indicating that the new table includes personally identifiable information (Pll).

Which line of code fills in the above blank to successfully complete the task?

Options:

A- COMMENT 'Contains PIT

B- 511

C- 'COMMENT PII'

D- TBLPROPERTIES PII

Answer:

D

Explanation:

To include a property indicating that a table contains personally identifiable information (PII), the TBLPROPERTIES keyword is used in SQL to add metadata to a table. The correct syntax to define a table property for PII is as follows:

CREATE TABLE customersInFrance

USING DELTA

TBLPROPERTIES ('PII' = 'true')

AS

SELECT id,

firstName,

lastName

FROM customerLocations

WHERE country = 'FRANCE';

The TBLPROPERTIES ('PII' = 'true') line correctly sets a table property that tags the table as containing personally identifiable information. This is in accordance with organizational policies for handling sensitive information.

Reference: Databricks documentation on Delta Lake: Delta Lake on Databricks

Question 2

Question Type: **MultipleChoice**

A data engineer needs to create a table in Databricks using data from their organization's existing SQLite database. They run the following command:

CREATE TABLE jdbc_customer360

USING

OPTIONS (

url "jdbc:sqlite:/customers.db", dbtable "customer360"

)

Which line of code fills in the above blank to successfully complete the task?

Options:

A- autoloader

B- org.apache.spark.sql.jdbc

C- sqlite

D- org.apache.spark.sql.sqlite

Answer:

B

Explanation:

To create a table in Databricks using data from an SQLite database, the correct syntax involves specifying the format of the data source. The format in the case of using JDBC (Java Database Connectivity) with SQLite is specified by the org.apache.spark.sql.jdbc format. This format allows Spark to interface with various relational databases through JDBC. Here is how the command should be structured:

```
CREATE TABLE jdbc_customer360

USING org.apache.spark.sql.jdbc

OPTIONS (

url 'jdbc:sqlite:/customers.db',

dbtable 'customer360'

)
```

The USING org.apache.spark.sql.jdbc line specifies that the JDBC data source is being used, enabling Spark to interact with the SQLite database via JDBC.

Reference: Databricks documentation on JDBC: Connecting to SQL Databases using JDBC

Question 3

Question Type: **MultipleChoice**

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

Options:

A- Data

B- Cluster management metadata

C- Databricks web application

D- Notebooks

Answer:

A

Explanation:

In a Databricks customer's cloud account, the primary elements stored include:

Data: This is the central type of content stored in the customer's cloud account. Data might include various datasets, tables, and files that are used and managed through Databricks platforms.

Notebooks: These are also stored within a customer's cloud account. Notebooks include scripts, notes, and other information necessary for data analysis and processing tasks.

Cluster management metadata is indeed managed through the cloud, but it's primarily handled by Databricks rather than stored directly in the customer's account. The Databricks web application itself is not stored within the customer's cloud account; rather, it's a service provided by Databricks.

Reference: Databricks documentation: Data in Databricks

Question 4

Question Type: MultipleChoice

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?

Options:

A- All datasets will be updated once and the pipeline will shut down. The compute resources will be terminated.

B- 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.

C- All datasets will be updated once and the pipeline will persist without any processing. The compute resources will persist but go unused.

D- All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.

E- All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing.

Answer:

B

Explanation:

The Continuous Pipeline Mode for Delta Live Tables allows the pipeline to run continuously and process data as it arrives. This mode is suitable for streaming ingest and CDC workloads that require low-latency updates. The Development mode for Delta Live Tables allows the pipeline to run on a dedicated cluster that is not shared with other pipelines. This mode is useful for testing and debugging the pipeline logic before deploying it to production. Therefore, the correct answer is B, because the pipeline will run continuously on a dedicated cluster until it is manually stopped, and the compute resources will be released only after the pipeline is shut down.Reference:Databricks Documentation - Configure pipeline settings for Delta Live Tables,Databricks Documentation - Continuous vs. triggered pipeline execution,Databricks Documentation - Development vs. production mode.

Question 5

Question Type: MultipleChoice

A data architect has determined that a table of the following format is necessary: Which of the following code blocks uses SQL DDL commands to create an empty Delta table in the above format regardless of whether a table already exists with this name?

A.
```
CREATE TABLE IF NOT EXISTS table_name (
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
)
```

B.
```
CREATE OR REPLACE TABLE table_name AS
SELECT
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
USING DELTA
```

C.
```
CREATE OR REPLACE TABLE table_name WITH COLUMNS (
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
) USING DELTA
```

D.
```
CREATE TABLE table_name AS
SELECT
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
```

E.
```
CREATE OR REPLACE TABLE table_name (
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
)
```

Options:

A- Option A

B- Option B

C- Option C

D- Option D

E- Option E

Answer:

E

Question 6

Question Type: **MultipleChoice**

A data engineer needs to apply custom logic to identify employees with more than 5 years of experience in array column employees in table stores. The custom logic should create a new column exp_employees that is an array of all of the employees with more than 5 years of experience for each row. In order to apply this custom logic at scale, the data engineer wants to use the FILTER higher-order function.

Which of the following code blocks successfully completes this task?

A.
```
SELECT
    store_id,
    employees,
    FILTER (employees, i -> i.years_exp > 5) AS exp_employees
FROM stores;
```

B.
```
SELECT
    store_id,
    employees,
    FILTER (exp_employees, years_exp > 5) AS exp_employees
FROM stores;
```

C.
```
SELECT
    store_id,
    employees,
    FILTER (employees, years_exp > 5) AS exp_employees
FROM stores;
```

D.
```
SELECT
    store_id,
    employees,
    CASE WHEN employees.years_exp > 5 THEN employees
         ELSE NULL
    END AS exp_employees
FROM stores;
```

E.
```
SELECT
    store_id,
    employees,
```

Options:

A- Option A

B- Option B

C- Option C

D- Option D

E- Option E

Answer:

A

Explanation:

Option A is the correct answer because it uses the FILTER higher-order function correctly to filter out employees with more than 5 years of experience from the array column "employees". It applies a lambda functioni -> i.years_exp > 5that checks if the years of experience of each employee in the array is greater than 5. If this condition is met, the employee is included in the new array column "exp_employees".

Question 7

Question Type: MultipleChoice

A data engineer and data analyst are working together on a data pipeline. The data engineer is working on the raw, bronze, and silver layers of the pipeline using Python, and the data analyst is working on the gold layer of the pipeline using SQL. The raw source of the pipeline is a streaming input. They now want to migrate their pipeline to use Delta Live Tables.

Which of the following changes will need to be made to the pipeline when migrating to Delta Live Tables?

Options:

A- None of these changes will need to be made

B- The pipeline will need to stop using the medallion-based multi-hop architecture

C- The pipeline will need to be written entirely in SQL

D- The pipeline will need to use a batch source in place of a streaming source

E- The pipeline will need to be written entirely in Python

Answer:

A

Explanation:

Delta Live Tables is a declarative framework for building reliable, maintainable, and testable data processing pipelines. You define the transformations to perform on your data and Delta Live Tables manages task orchestration, cluster management, monitoring, data quality, and error handling. Delta Live Tables supports both SQL and Python as the languages for defining your datasets and expectations. Delta Live Tables also supports both streaming and batch sources, and can handle both append-only and upsert data patterns. Delta Live Tables follows the medallion lakehouse architecture, which consists of three layers of data: bronze, silver, and gold. Therefore, migrating to Delta Live Tables does not require any of the changes listed in the options B, C, D, or E. The data engineer and data analyst can use the same languages, sources, and architecture as before, and simply declare their datasets and expectations using Delta Live Tables syntax.Reference:

What is Delta Live Tables?

Transform data with Delta Live Tables

What is the medallion lakehouse architecture?

Question 8

Question Type: MultipleChoice

Which of the following is stored in the Databricks customer's cloud account?

Options:

A- Databricks web application

B- Cluster management metadata

C- Repos

D- Data

E- Notebooks

Answer:

D

Explanation:

The only option that is stored in the Databricks customer's cloud account is data. Data is stored in the customer's cloud storage service, such as AWS S3 or Azure Data Lake Storage. The customer has full control and ownership of their data and can access it directly from their cloud account.

Option A is not correct, as the Databricks web application is hosted and managed by Databricks on their own cloud infrastructure. The customer does not need to install or maintain the web application, but only needs to access it through a web browser.

Option B is not correct, as the cluster management metadata is stored and managed by Databricks on their own cloud infrastructure. The cluster management metadata includes information such as cluster configuration, status, logs, and metrics. The customer can view and manage their clusters through the Databricks web application, but does not have direct access to the cluster management metadata.

Option C is not correct, as the repos are stored and managed by Databricks on their own cloud infrastructure. Repos are version-controlled repositories that store code and data files for Databricks projects. The customer can create and manage their repos through the Databricks web application, but does not have direct access to the repos.

Option E is not correct, as the notebooks are stored and managed by Databricks on their own cloud infrastructure. Notebooks are interactive documents that contain code, text, and visualizations for Databricks workflows. The customer can create and manage their notebooks through the Databricks web application, but does not have direct access to the notebooks.

Databricks Architecture

Databricks Data Sources

Databricks Repos

[Databricks Notebooks]

[Databricks Data Engineer Professional Exam Guide]

Question 9

Question Type: **MultipleChoice**

A data engineer wants to schedule their Databricks SQL dashboard to refresh every hour, but they only want the associated SQL endpoint to be running when it is necessary. The dashboard has multiple queries on multiple datasets associated with it. The data that

feeds the dashboard is automatically processed using a Databricks Job.

Which of the following approaches can the data engineer use to minimize the total running time of the SQL endpoint used in the refresh schedule of their dashboard?

Options:

A- They can turn on the Auto Stop feature for the SQL endpoint.

B- They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.

C- They can reduce the cluster size of the SQL endpoint.

D- They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.

E- They can set up the dashboard's SQL endpoint to be serverless.

Answer:

A

Explanation:

The Auto Stop feature allows the SQL endpoint to automatically stop after a specified period of inactivity. This can help reduce the cost and resource consumption of the SQL endpoint, as it will only run when it is needed to refresh the dashboard or execute queries. The data engineer can configure the Auto Stop setting for the SQL endpoint from the SQL Endpoints UI, by selecting the desired idle time from the Auto Stop dropdown menu. The default idle time is 120 minutes, but it can be set to as low as 15 minutes or as high as 240

minutes. Alternatively, the data engineer can also use the SQL Endpoints REST API to set the Auto Stop setting programmatically.Reference:SQL Endpoints UI,SQL Endpoints REST API,Refreshing SQL Dashboard

Question 10

Question Type: MultipleChoice

An engineering manager uses a Databricks SQL query to monitor ingestion latency for each data source. The manager checks the results of the query every day, but they are manually rerunning the query each day and waiting for the results.

Which of the following approaches can the manager use to ensure the results of the query are updated each day?

Options:

A- They can schedule the query to refresh every 1 day from the SQL endpoint's page in Databricks SQL.

B- They can schedule the query to refresh every 12 hours from the SQL endpoint's page in Databricks SQL.

C- They can schedule the query to refresh every 1 day from the query's page in Databricks SQL.

D- They can schedule the query to run every 1 day from the Jobs UI.

E- They can schedule the query to run every 12 hours from the Jobs UI.

Answer:

C

Explanation:

Databricks SQL allows users to schedule queries to run automatically at a specified frequency and time zone. This can help users to keep their dashboards or alerts updated with the latest data. To schedule a query, users need to do the following steps:

In the Query Editor, click Schedule > Add schedule to open a menu with schedule settings.

Choose when to run the query. Use the dropdown pickers to specify the frequency, period, starting time, and time zone. Optionally, select the Show cron syntax checkbox to edit the schedule in Quartz Cron Syntax.

Choose More options to show optional settings. Users can also choose a name for the schedule, and a SQL warehouse to power the query.

Click Create. The query will run automatically according to the schedule.

The other options are incorrect because they do not refer to the correct location or frequency to schedule the query. The query's page in Databricks SQL is the place where users can edit, run, or schedule the query. The SQL endpoint's page in Databricks SQL is the place where users can manage the SQL warehouses and SQL endpoints. The Jobs UI is the place where users can create, run, or schedule jobs that execute notebooks, JARs, or Python scripts.Reference:Schedule a query,What are Databricks SQL alerts?,Jobs.

Question 11

Question Type: MultipleChoice

A data engineer that is new to using Python needs to create a Python function to add two integers together and return the sum?

Which of the following code blocks can the data engineer use to complete this task?

A)

```
function add_integers(x, y):
    return x + y
```

B)

```
function add_integers(x, y):
    x + y
```

C)

```
def add_integers(x, y):
    print(x + y)
```

D)

```
def add_integers(x, y):
    return x + y
```

E)

```
def add_integers(x, y):
    x + y
```

Options:

A- Option A

B- Option B

C- Option C

D- Option D

E- Option E

Answer:

D

Explanation:

https://www.w3schools.com/python/python_functions.asp

https://www.geeksforgeeks.org/python-functions/

Question 12

Question Type: MultipleChoice

A data engineer wants to create a relational object by pulling data from two tables. The relational object does not need to be used by other data engineers in other sessions. In order to save on storage costs, the data engineer wants to avoid copying and storing physical data.

Which of the following relational objects should the data engineer create?

Options:

A- Spark SQL Table

B- View

C- Database

D- Temporary view

E- Delta Table

Answer:

D

Explanation:

A temporary view is a relational object that is defined in the metastore and points to an existing DataFrame. It does not copy or store any physical data, but only saves the query that defines the view. The lifetime of a temporary view is tied to the SparkSession that was used to create it, so it does not persist across different sessions or applications. A temporary view is useful for accessing the same data multiple times within the same notebook or session, without incurring additional storage costs. The other options are either materialized (A, E), persistent (B, C), or not relational objects .Reference:Databricks Documentation - Temporary View,Databricks Community - How do temp views actually work?,Databricks Community - What's the difference between a Global view and a Temp view?,Big Data Programmers - Temporary View in Databricks.

To Get Premium Files for Databricks-Certified-Data-Engineer-Associate Visit

https://www.p2pexams.com/products/databricks-certified-data-engineer-associate

For More Free Questions Visit

https://www.p2pexams.com/databricks/pdf/databricks-certified-data-engineer-associate