

Exam Questions Databricks-Certified-Data-Engineer-Associate

Databricks Certified Data Engineer Associate Exam

<https://www.2passeasy.com/dumps/Databricks-Certified-Data-Engineer-Associate/>



NEW QUESTION 1

Which of the following commands will return the location of database customer360?

- A. DESCRIBE LOCATION customer360;
- B. DROP DATABASE customer360;
- C. DESCRIBE DATABASE customer360;
- D. ALTER DATABASE customer360 SET DBPROPERTIES ('location' = '/user');
- E. USE DATABASE customer360;

Answer: C

Explanation:

To retrieve the location of a database named "customer360" in a database management system like Hive or Databricks, you can use the DESCRIBE DATABASE command followed by the database name. This command will provide information about the database, including its location.

NEW QUESTION 2

A data engineer has created a new database using the following command: CREATE DATABASE IF NOT EXISTS customer360;
In which of the following locations will the customer360 database be located?

- A. dbfs:/user/hive/database/customer360
- B. dbfs:/user/hive/warehouse
- C. dbfs:/user/hive/customer360
- D. More information is needed to determine the correct response

Answer: B

Explanation:

dbfs:/user/hive/warehouse - which is the default location

NEW QUESTION 3

Which of the following approaches should be used to send the Databricks Job owner an email in the case that the Job fails?

- A. Manually programming in an alert system in each cell of the Notebook
- B. Setting up an Alert in the Job page
- C. Setting up an Alert in the Notebook
- D. There is no way to notify the Job owner in the case of Job failure
- E. MLflow Model Registry Webhooks

Answer: B

Explanation:

<https://docs.databricks.com/en/workflows/jobs/job-notifications.html>

NEW QUESTION 4

A data engineering team has two tables. The first table march_transactions is a collection of all retail transactions in the month of March. The second table april_transactions is a collection of all retail transactions in the month of April. There are no duplicate records between the tables.
Which of the following commands should be run to create a new table all_transactions that contains all records from march_transactions and april_transactions without duplicate records?

- A. CREATE TABLE all_transactions AS SELECT * FROM march_transactions INNER JOIN SELECT * FROM april_transactions;
- B. CREATE TABLE all_transactions AS SELECT * FROM march_transactions UNION SELECT * FROM april_transactions;
- C. CREATE TABLE all_transactions AS SELECT * FROM march_transactions OUTER JOIN SELECT * FROM april_transactions;
- D. CREATE TABLE all_transactions AS SELECT * FROM march_transactions INTERSECT SELECT * FROM april_transactions;
- E. CREATE TABLE all_transactions AS SELECT * FROM march_transactions MERGE SELECT * FROM april_transactions;

Answer: B

Explanation:

To create a new table all_transactions that contains all records from march_transactions and april_transactions without duplicate records, you should use the UNION operator, as shown in option B. This operator combines the result sets of the two tables while automatically removing duplicate records.

NEW QUESTION 5

A data organization leader is upset about the data analysis team's reports being different from the data engineering team's reports. The leader believes the siloed nature of their organization's data engineering and data analysis architectures is to blame.
Which of the following describes how a data lakehouse could alleviate this issue?

- A. Both teams would autoscale their work as data size evolves
- B. Both teams would use the same source of truth for their work
- C. Both teams would reorganize to report to the same department
- D. Both teams would be able to collaborate on projects in real-time
- E. Both teams would respond more quickly to ad-hoc requests

Answer: B

Explanation:

A data lakehouse is designed to unify the data engineering and data analysis architectures by integrating features of both data lakes and data warehouses. One of

the key benefits of a data lakehouse is that it provides a common, centralized data repository (the "lake") that serves as a single source of truth for data storage and analysis. This allows both data engineering and data analysis teams to work with the same consistent data sets, reducing discrepancies and ensuring that the reports generated by both teams are based on the same underlying data.

NEW QUESTION 6

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?

- 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

NEW QUESTION 7

Which of the following data lakehouse features results in improved data quality over a traditional data lake?

- A. A data lakehouse provides storage solutions for structured and unstructured data.
- B. A data lakehouse supports ACID-compliant transactions.
- C. A data lakehouse allows the use of SQL queries to examine data.
- D. A data lakehouse stores data in open formats.
- E. A data lakehouse enables machine learning and artificial Intelligence workloads.

Answer: B

Explanation:

One of the key features of a data lakehouse that results in improved data quality over a traditional data lake is its support for ACID (Atomicity, Consistency, Isolation, Durability) transactions. ACID transactions provide data integrity and consistency guarantees, ensuring that operations on the data are reliable and that data is not left in an inconsistent state due to failures or concurrent access. In a traditional data lake, such transactional guarantees are often lacking, making it challenging to maintain data quality, especially in scenarios involving multiple data writes, updates, or complex transformations. A data lakehouse, by offering ACID compliance, helps maintain data quality by providing strong consistency and reliability, which is crucial for data pipelines and analytics.

NEW QUESTION 8

A data engineering team has noticed that their Databricks SQL queries are running too slowly when they are submitted to a non-running SQL endpoint. The data engineering team wants this issue to be resolved.

Which of the following approaches can the team use to reduce the time it takes to return results in this scenario?

- A. They can turn on the Serverless feature for the SQL endpoint and change the Spot Instance Policy to "Reliability Optimized."
- B. They can turn on the Auto Stop feature for the SQL endpoint.
- C. They can increase the cluster size of the SQL endpoint.
- D. They can turn on the Serverless feature for the SQL endpoint.
- E. They can increase the maximum bound of the SQL endpoint's scaling range

Answer: C

Explanation:

<https://www.databricks.com/blog/2022/03/10/top-5-databricks-performance-tips.html>

NEW QUESTION 9

Which of the following can be used to simplify and unify siloed data architectures that are specialized for specific use cases?

- A. None of these
- B. Data lake
- C. Data warehouse
- D. All of these
- E. Data lakehouse

Answer: E

NEW QUESTION 10

Which of the following describes the storage organization of a Delta table?

- A. Delta tables are stored in a single file that contains data, history, metadata, and other attributes.
- B. Delta tables store their data in a single file and all metadata in a collection of files in a separate location.
- C. Delta tables are stored in a collection of files that contain data, history, metadata, and other attributes.
- D. Delta tables are stored in a collection of files that contain only the data stored within the table.
- E. Delta tables are stored in a single file that contains only the data stored within the table.

Answer: C

Explanation:

Delta tables store data in a structured manner using Parquet files, and they also maintain metadata and transaction logs in separate directories. This organization

allows for versioning, transactional capabilities, and metadata tracking in Delta Lake. Thank you for pointing out the error, and I appreciate your understanding.

NEW QUESTION 10

Which of the following benefits is provided by the array functions from Spark SQL?

- A. An ability to work with data in a variety of types at once
- B. An ability to work with data within certain partitions and windows
- C. An ability to work with time-related data in specified intervals
- D. An ability to work with complex, nested data ingested from JSON files
- E. An ability to work with an array of tables for procedural automation

Answer: D

Explanation:

Array functions in Spark SQL are primarily used for working with arrays and complex, nested data structures, such as those often encountered when ingesting JSON files. These functions allow you to manipulate and query nested arrays and structures within your data, making it easier to extract and work with specific elements or values within complex data formats. While some of the other options (such as option A for working with different data types) are features of Spark SQL or SQL in general, array functions specifically excel at handling complex, nested data structures like those found in JSON files.

NEW QUESTION 14

Which of the following tools is used by Auto Loader process data incrementally?

- A. Checkpointing
- B. Spark Structured Streaming
- C. Data Explorer
- D. Unity Catalog
- E. Databricks SQL

Answer: B

Explanation:

The Auto Loader process in Databricks is typically used in conjunction with Spark Structured Streaming to process data incrementally. Spark Structured Streaming is a real-time data processing framework that allows you to process data streams incrementally as new data arrives. The Auto Loader is a feature in Databricks that works with Structured Streaming to automatically detect and process new data files as they are added to a specified data source location. It allows for incremental data processing without the need for manual intervention.

How does Auto Loader track ingestion progress? As files are discovered, their metadata is persisted in a scalable key-value store (RocksDB) in the checkpoint location of your Auto Loader pipeline. This key-value store ensures that data is processed exactly once. In case of failures, Auto Loader can resume from where it left off by information stored in the checkpoint location and continue to provide exactly-once guarantees when writing data into Delta Lake. You don't need to maintain or manage any state yourself to achieve fault tolerance or exactly-once semantics. <https://docs.databricks.com/ingestion/auto-loader/index.html>

NEW QUESTION 18

A data analyst has created a Delta table sales that is used by the entire data analysis team. They want help from the data engineering team to implement a series of tests to ensure the data is clean. However, the data engineering team uses Python for its tests rather than SQL.

Which of the following commands could the data engineering team use to access sales in PySpark?

- A. `SELECT * FROM sales`
- B. There is no way to share data between PySpark and SQL.
- C. `spark.sql("sales")`
- D. `spark.delta.table("sales")`
- E. `spark.table("sales")`

Answer: E

Explanation:

<https://spark.apache.org/docs/3.2.1/api/python/reference/api/pyspark.sql.Session.html>

NEW QUESTION 21

A new data engineering team has been assigned to an ELT project. The new data engineering team will need full privileges on the table sales to fully manage the project.

Which of the following commands can be used to grant full permissions on the database to the new data engineering team?

- A. `GRANT ALL PRIVILEGES ON TABLE sales TO team;`
- B. `GRANT SELECT CREATE MODIFY ON TABLE sales TO team;`
- C. `GRANT SELECT ON TABLE sales TO team;`
- D. `GRANT USAGE ON TABLE sales TO team;`
- E. `GRANT ALL PRIVILEGES ON TABLE team TO sales;`

Answer: A

NEW QUESTION 23

A new data engineering team has been assigned to work on a project. The team will need access to database customers in order to see what tables already exist. The team has its own group team.

Which of the following commands can be used to grant the necessary permission on the entire database to the new team?

- A. `GRANT VIEW ON CATALOG customers TO team;`
- B. `GRANT CREATE ON DATABASE customers TO team;`
- C. `GRANT USAGE ON CATALOG team TO customers;`

- D. GRANT CREATE ON DATABASE team TO customers;
E. GRANT USAGE ON DATABASE customers TO team;

Answer: E

Explanation:

The GRANT statement is used to grant privileges on a database, table, or view to a user or role. The ALL PRIVILEGES option grants all possible privileges on the specified object, such as CREATE, SELECT, MODIFY, and USAGE. The syntax of the GRANT statement is:

GRANT privilege_type ON object TO user_or_role;

Therefore, to grant full permissions on the database customers to the new data engineering team, the command should be:

GRANT ALL PRIVILEGES ON DATABASE customers TO team;

NEW QUESTION 25

Which of the following describes the relationship between Bronze tables and raw data?

- A. Bronze tables contain less data than raw data files.
B. Bronze tables contain more truthful data than raw data.
C. Bronze tables contain aggregates while raw data is unaggregated.
D. Bronze tables contain a less refined view of data than raw data.
E. Bronze tables contain raw data with a schema applied.

Answer: E

Explanation:

The Bronze layer is where we land all the data from external source systems. The table structures in this layer correspond to the source system table structures "as-is," along with any additional metadata columns that capture the load date/time, process ID, etc. The focus in this layer is quick Change Data Capture and the ability to provide an historical archive of source (cold storage), data lineage, auditability, reprocessing if needed without rereading the data from the source system.
<https://www.databricks.com/glossary/medallion-architecture#:~:text=Bronze%20layer%20%28raw%20data%29>

NEW QUESTION 27

In order for Structured Streaming to reliably track the exact progress of the processing so that it can handle any kind of failure by restarting and/or reprocessing, which of the following two approaches is used by Spark to record the offset range of the data being processed in each trigger?

- A. Checkpointing and Write-ahead Logs
B. Structured Streaming cannot record the offset range of the data being processed in each trigger.
C. Replayable Sources and Idempotent Sinks
D. Write-ahead Logs and Idempotent Sinks
E. Checkpointing and Idempotent Sinks

Answer: A

Explanation:

The engine uses checkpointing and write-ahead logs to record the offset range of the data being processed in each trigger. -- in the link search for "The engine uses " you'll find the answer.
<https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html#:~:text=The%20engine%20uses%20checkpointing%20and,being%20processed%20in%20each%20trigger.>

NEW QUESTION 28

A data engineer wants to create a new table containing the names of customers that 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 (PII).

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

- A. There is no way to indicate whether a table contains PII.
B. "COMMENT PII"
C. TBLPROPERTIES PII
D. COMMENT "Contains PII"
E. PII

Answer: D

Explanation:

Ref:<https://www.databricks.com/discover/pages/data-quality-management>
CREATE TABLE my_table (id INT COMMENT 'Unique Identification Number', name STRING COMMENT 'PII', age INT COMMENT 'PII') TBLPROPERTIES ('contains_pii'=True) COMMENT 'Contains PII';

NEW QUESTION 32

A new data engineering team team. has been assigned to an ELT project. The new data engineering team will need full privileges on the database customers to fully manage the project.

Which of the following commands can be used to grant full permissions on the database to the new data engineering team?

- A. GRANT USAGE ON DATABASE customers TO team;
- B. GRANT ALL PRIVILEGES ON DATABASE team TO customers;
- C. GRANT SELECT PRIVILEGES ON DATABASE customers TO teams;
- D. GRANT SELECT CREATE MODIFY USAGE PRIVILEGES ON DATABASE customers TO team;
- E. GRANT ALL PRIVILEGES ON DATABASE customers TO team;

Answer: E

Explanation:

To grant full privileges on the database "customers" to the new data engineering team, you can use the GRANT ALL PRIVILEGES command as shown in option E. This command provides the team with all possible privileges on the specified database, allowing them to fully manage it.

NEW QUESTION 34

A data engineer is running code in a Databricks Repo that is cloned from a central Git repository. A colleague of the data engineer informs them that changes have been made and synced to the central Git repository. The data engineer now needs to sync their Databricks Repo to get the changes from the central Git repository. Which of the following Git operations does the data engineer need to run to accomplish this task?

- A. Merge
- B. Push
- C. Pull
- D. Commit
- E. Clone

Answer: C

Explanation:

From the docs:

In Databricks Repos, you can use Git functionality to: Clone, push to, and pull from a remote Git repository.

Create and manage branches for development work, including merging, rebasing, and resolving conflicts.

Create notebooks—including IPYNB notebooks—and edit them and other files.

Visually compare differences upon commit and resolve merge conflicts. Source: <https://docs.databricks.com/en/repos/index.html>

NEW QUESTION 39

Which of the following benefits of using the Databricks Lakehouse Platform is provided by Delta Lake?

- A. The ability to manipulate the same data using a variety of languages
- B. The ability to collaborate in real time on a single notebook
- C. The ability to set up alerts for query failures
- D. The ability to support batch and streaming workloads
- E. The ability to distribute complex data operations

Answer: D

Explanation:

Delta Lake is a key component of the Databricks Lakehouse Platform that provides several benefits, and one of the most significant benefits is its ability to support both batch and streaming workloads seamlessly. Delta Lake allows you to process and analyze data in real-time (streaming) as well as in batch, making it a versatile choice for various data processing needs. While the other options may be benefits or capabilities of Databricks or the Lakehouse Platform in general, they are not specifically associated with Delta Lake.

NEW QUESTION 40

Which of the following commands will return the number of null values in the member_id column?

- A. SELECT count(member_id) FROM my_table;
- B. SELECT count(member_id) - count_null(member_id) FROM my_table;
- C. SELECT count_if(member_id IS NULL) FROM my_table;
- D. SELECT null(member_id) FROM my_table;
- E. SELECT count_null(member_id) FROM my_table;

Answer: C

Explanation:

<https://docs.databricks.com/en/sql/language-manual/functions/count.html>

Returns

A BIGINT.

If * is specified also counts row containing NULL values.

If expr are specified counts only rows for which all expr are not NULL. If DISTINCT duplicate rows are not counted.

NEW QUESTION 42

Which of the following describes a scenario in which a data engineer will want to use a single-node cluster?

- A. When they are working interactively with a small amount of data
- B. When they are running automated reports to be refreshed as quickly as possible
- C. When they are working with SQL within Databricks SQL

- D. When they are concerned about the ability to automatically scale with larger data
- E. When they are manually running reports with a large amount of data

Answer: A

Explanation:

A Single Node cluster is a cluster consisting of an Apache Spark driver and no Spark workers. A Single Node cluster supports Spark jobs and all Spark data sources, including Delta Lake. A Standard cluster requires a minimum of one Spark worker to run Spark jobs.

NEW QUESTION 44

A data engineer needs to determine whether to use the built-in Databricks Notebooks versioning or version their project using Databricks Repos. Which of the following is an advantage of using Databricks Repos over the Databricks Notebooks versioning?

- A. Databricks Repos automatically saves development progress
- B. Databricks Repos supports the use of multiple branches
- C. Databricks Repos allows users to revert to previous versions of a notebook
- D. Databricks Repos provides the ability to comment on specific changes
- E. Databricks Repos is wholly housed within the Databricks Lakehouse Platform

Answer: B

Explanation:

An advantage of using Databricks Repos over the built-in Databricks Notebooks versioning is the ability to work with multiple branches. Branching is a fundamental feature of version control systems like Git, which Databricks Repos is built upon. It allows you to create separate branches for different tasks, features, or experiments within your project. This separation helps in parallel development and experimentation without affecting the main branch or the work of other team members. Branching provides a more organized and collaborative development environment, making it easier to merge changes and manage different development efforts. While Databricks Notebooks versioning also allows you to track versions of notebooks, it may not provide the same level of flexibility and collaboration as branching in Databricks Repos.

NEW QUESTION 47

A data engineer has a Python variable `table_name` that they would like to use in a SQL query. They want to construct a Python code block that will run the query using `table_name`.

They have the following incomplete code block:

```
("SELECT customer_id, spend FROM {table_name}")
```

Which of the following can be used to fill in the blank to successfully complete the task?

- A. `spark.delta.sql`
- B. `spark.delta.table`
- C. `spark.table`
- D. `dbutils.sql`
- E. `spark.sql`

Answer: E

NEW QUESTION 49

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

```
id STRING = 'a1'
```

```
rank INTEGER = 6 rating FLOAT = 9.4
```

Which of the following SQL commands 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. `my_table UNION VALUES ('a1', 6, 9.4)`
- C. `INSERT VALUES ('a1', 6, 9.4) INTO my_table`
- D. `UPDATE my_table VALUES ('a1', 6, 9.4)`
- E. `UPDATE VALUES ('a1', 6, 9.4) my_table`

Answer: A

NEW QUESTION 54

A dataset has been defined using Delta Live Tables and includes an expectations clause:

```
CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION FAIL UPDATE
```

What is the expected behavior when a batch of data containing data that violates these constraints is processed?

- A. Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.
- B. Records that violate the expectation cause the job to fail.
- C. Records that violate the expectation are dropped from the target dataset and loaded into a quarantine table.
- D. Records that violate the expectation are added to the target dataset and recorded as invalid in the event log.
- E. Records that violate the expectation are added to the target dataset and flagged as invalid in a field added to the target dataset.

Answer: B

Explanation:

<https://docs.databricks.com/en/delta-live-tables/expectations.html> Action

Result

warn (default)

Invalid records are written to the target; failure is reported as a metric for the dataset. drop

Invalid records are dropped before data is written to the target; failure is reported as a metrics for the dataset.

fail

Invalid records prevent the update from succeeding. Manual intervention is required before re-processing.

NEW QUESTION 56

Which of the following code blocks will remove the rows where the value in column age is greater than 25 from the existing Delta table my_table and save the updated table?

- A. SELECT * FROM my_table WHERE age > 25;
- B. UPDATE my_table WHERE age > 25;
- C. DELETE FROM my_table WHERE age > 25;
- D. UPDATE my_table WHERE age <= 25;
- E. DELETE FROM my_table WHERE age <= 25;

Answer: C

NEW QUESTION 57

A data engineer has developed a data pipeline to ingest data from a JSON source using Auto Loader, but the engineer has not provided any type inference or schema hints in their pipeline. Upon reviewing the data, the data engineer has noticed that all of the columns in the target table are of the string type despite some of the fields only including float or boolean values.

Which of the following describes why Auto Loader inferred all of the columns to be of the string type?

- A. There was a type mismatch between the specific schema and the inferred schema
- B. JSON data is a text-based format
- C. Auto Loader only works with string data
- D. All of the fields had at least one null value
- E. Auto Loader cannot infer the schema of ingested data

Answer: B

Explanation:

JSON data is a text-based format that uses strings to represent all values. When Auto Loader infers the schema of JSON data, it assumes that all values are strings. This is because Auto Loader cannot determine the type of a value based on its string representation. <https://docs.databricks.com/en/ingestion/auto-loader/schema.html> For example, the following JSON string represents a value that is logically a boolean: JSON "true" Use code with caution. Learn more However, Auto Loader would infer that the type of this value is string. This is because Auto Loader cannot determine that the value is a boolean based on its string representation. In order to get Auto Loader to infer the correct types for columns, the data engineer can provide type inference or schema hints. Type inference hints can be used to specify the types of specific columns. Schema hints can be used to provide the entire schema of the data. Therefore, the correct answer is B. JSON data is a text-based format.

NEW QUESTION 59

A data engineer has a single-task Job that runs each morning before they begin working. After identifying an upstream data issue, they need to set up another task to run a new notebook prior to the original task.

Which of the following approaches can the data engineer use to set up the new task?

- A. They can clone the existing task in the existing Job and update it to run the new notebook.
- B. They can create a new task in the existing Job and then add it as a dependency of the original task.
- C. They can create a new task in the existing Job and then add the original task as a dependency of the new task.
- D. They can create a new job from scratch and add both tasks to run concurrently.
- E. They can clone the existing task to a new Job and then edit it to run the new notebook.

Answer: B

Explanation:

To set up the new task to run a new notebook prior to the original task in a single-task Job, the data engineer can use the following approach: In the existing Job, create a new task that corresponds to the new notebook that needs to be run. Set up the new task with the appropriate configuration, specifying the notebook to be executed and any necessary parameters or dependencies. Once the new task is created, designate it as a dependency of the original task in the Job configuration. This ensures that the new task is executed before the original task.

NEW QUESTION 64

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

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

Answer: D

NEW QUESTION 68

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?

- A. Spark SQL Table
- B. View
- C. Database
- D. Temporary view
- E. Delta Table

Answer: D

Explanation:

Temp view : session based Create temp view view_name as query All these are termed as session ended: Opening a new notebook Detaching and reattaching a cluster Installing a python package Restarting a cluster

NEW QUESTION 69

A data engineer has a Job with multiple tasks that runs nightly. Each of the tasks runs slowly because the clusters take a long time to start. Which of the following actions can the data engineer perform to improve the start up time for the clusters used for the Job?

- A. They can use endpoints available in Databricks SQL
- B. They can use jobs clusters instead of all-purpose clusters
- C. They can configure the clusters to be single-node
- D. They can use clusters that are from a cluster pool
- E. They can configure the clusters to autoscale for larger data sizes

Answer: D

Explanation:

Cluster pools are a way to pre-provision clusters that are ready to use. This can reduce the start up time for clusters, as they do not have to be created from scratch. All-purpose clusters are not pre-provisioned, so they will take longer to start up. Jobs clusters are a type of cluster pool, but they are not the best option for this use case. Jobs clusters are designed for long-running jobs, and they can be more expensive than other types of cluster pools. Single-node clusters are the smallest type of cluster, and they will start up the fastest. However, they may not be powerful enough to run the Job's tasks. Autoscaling clusters can scale up or down based on demand. This can help to improve the start up time for clusters, as they will only be created when they are needed. However, autoscaling clusters can also be more expensive than other types of cluster pool <https://docs.databricks.com/en/clusters/pool-best-practices.html>

NEW QUESTION 72

Which of the following describes a benefit of creating an external table from Parquet rather than CSV when using a CREATE TABLE AS SELECT statement?

- A. Parquet files can be partitioned
- B. CREATE TABLE AS SELECT statements cannot be used on files
- C. Parquet files have a well-defined schema
- D. Parquet files have the ability to be optimized
- E. Parquet files will become Delta tables

Answer: C

Explanation:

<https://www.databricks.com/glossary/what-is-parquet#:~:text=Columnar%20storage%20like%20Apache%20Parquet,compared%20to%20Row%20oriented%20databases.> Columnar storage like Apache Parquet is designed to bring efficiency compared to row-based files like CSV. When querying, columnar storage you can skip over the non-relevant data very quickly. As a result, aggregation queries are less time-consuming compared to row-oriented databases.

NEW QUESTION 74

A data engineer has realized that the data files associated with a Delta table are incredibly small. They want to compact the small files to form larger files to improve performance.

Which of the following keywords can be used to compact the small files?

- A. REDUCE
- B. OPTIMIZE
- C. COMPACTION
- D. REPARTITION
- E. VACUUM

Answer: B

Explanation:

OPTIMIZE can be used to club small files into 1 and improve performance.

NEW QUESTION 75

A data engineer has joined an existing project and they see the following query in the project repository:

```
CREATE STREAMING LIVE TABLE loyal_customers AS SELECT customer_id -  
FROM STREAM(LIVE.customers) WHERE loyalty_level = 'high';
```

Which of the following describes why the STREAM function is included in the query?

- A. The STREAM function is not needed and will cause an error.
- B. The table being created is a live table.
- C. The customers table is a streaming live table.
- D. The customers table is a reference to a Structured Streaming query on a PySpark DataFrame.
- E. The data in the customers table has been updated since its last run.

Answer: C

Explanation:

<https://docs.databricks.com/en/sql/load-data-streaming-table.html> Load data into a streaming table

To create a streaming table from data in cloud object storage, paste the following into the query editor, and then click Run:

SQL

Copy to clipboardCopy

```
/* Load data from a volume */
```

```
CREATE OR REFRESH STREAMING TABLE <table-name> AS SELECT * FROM STREAM  
read_files('/Volumes/<catalog>/<schema>/<volume>/<path>/<folder>')
```

```
/* Load data from an external location */  
CREATE OR REFRESH STREAMING TABLE <table-name> AS  
SELECT * FROM STREAM read_files('s3://<bucket>/<path>/<folder>')
```

NEW QUESTION 77

.....

THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual Databricks-Certified-Data-Engineer-Associate Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the Databricks-Certified-Data-Engineer-Associate Product From:

<https://www.2passeasy.com/dumps/Databricks-Certified-Data-Engineer-Associate/>

Money Back Guarantee

Databricks-Certified-Data-Engineer-Associate Practice Exam Features:

- * Databricks-Certified-Data-Engineer-Associate Questions and Answers Updated Frequently
- * Databricks-Certified-Data-Engineer-Associate Practice Questions Verified by Expert Senior Certified Staff
- * Databricks-Certified-Data-Engineer-Associate Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- * Databricks-Certified-Data-Engineer-Associate Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year