DUMPSBOSS.

Databricks Certified Data Engineer Professional Exam

Databricks Databricks-Certified-Professional-Data-Engineer

Version Demo

Total Demo Questions: 15

Total Premium Questions: 220

Buy Premium PDF

https://dumpsboss.co

support@dumpsboss.co

support@dumpsboss.co
dumpsboss.co

QUESTION NO: 1

What is the main difference between the below two commands?

1. INSERT OVERWRITE table_name

2. SELECT * FROM table

1. CREATE OR REPLACE TABLE table_name

2. AS SELECT * FROM table

A. INSERT OVERWRITE replaces data by default, CREATE OR REPLACE replaces data and Schema by default

B. INSERT OVERWRITE replaces data and schema by default, CREATE OR REPLACEreplaces data by default

C. INSERT OVERWRITE maintains historical data versions by de-fault, CREATE OR REPLACEclears the historical data versions by default

D. INSERT OVERWRITE clears historical data versions by de-fault, CREATE OR REPLACE maintains the historical data versions by default

E. Both are same and results in identical outcomes

ANSWER: A

Explanation:

The main difference between INSERT OVERWRITE and CREATE OR REPLACE TABLE(CRAS) is that CRAS can modify the schema of the table, i.e it can add new columns or change data types of existing columns. By default INSERT OVERWRITE only overwrites the data.

INSERT OVERWRITE can also be used to overwrite schema, only when spark.databricks.delta.schema.autoMerge.enabled is set true if this option is not enabled and if there is a schema mismatch command will fail.

QUESTION NO: 2

What is the underlying technology that makes the Auto Loader work?

A. Loader

B. Delta Live Tables

C. Structured Streaming

D. DataFrames

E. Live DataFames

ANSWER: C

QUESTION NO: 3

How does Lakehouse replace the dependency on using Data lakes and Data warehouses in a Data and Analytics solution?

A. Open, direct access to data stored in standard data formats.

B. Supports ACID transactions.

C. Supports BI and Machine learning workloads

D. Support for end-to-end streaming and batch workloads

E. All the above

ANSWER: E

Explanation:

Explanation

Lakehouse combines the benefits of a data warehouse and data lakes,

Lakehouse = Data Lake + DataWarehouse

Here are some of the major benefits of a lakehouse

A lakehouse has the following key features:

- **Transaction support:** In an enterprise lakehouse many data pipelines will often be reading and writing data concurrently. Support for ACID transactions ensures consistency as multiple parties concurrently read or write data, typically using SQL.
- **Schema enforcement and governance:** The Lakehouse should have a way to support schema enforcement and evolution, supporting DW schema architectures such as star/snowflake-schemas. The system should be able to reason about data integrity, and it should have robust governance and auditing mechanisms.
- **BI support:** Lakehouses enable using BI tools directly on the source data. This reduces staleness and improves recency, reduces latency, and lowers the cost of having to operationalize two copies of the data in both a data lake and a warehouse.
- **Storage is decoupled from compute:** In practice this means storage and compute use separate clusters, thus these systems are able to scale to many more concurrent users and larger data sizes. Some modern data warehouses also have this property.
- **Openness:** The storage formats they use are open and standardized, such as Parquet, and they provide an API so a variety of tools and engines, including machine learning and Python/R libraries, can efficiently access the data **directly**.
- **Support for diverse data types ranging from unstructured to structured data**: The lakehouse can be used to store, refine, analyze, and access data types needed for many new data applications, including images, video, audio, semi-structured data, and text.
- **Support for diverse workloads:** including data science, machine learning, and SQL and analytics. Multiple tools might be needed to support all these workloads but they all rely on the same data repository.
- **End-to-end streaming:** Real-time reports are the norm in many enterprises. Support for streaming eliminates the need for separate systems dedicated to serving real-time data applications.

Lakehouse = Data Lake + DataWarehouse

QUESTION NO: 4

Which of the following commands can be used to run one notebook from another notebook?

A. notebook.utils.run("full notebook path")

B. execute.utils.run("full notebook path")

C. dbutils.notebook.run("full notebook path")

D. only job clusters can run notebook

E. spark.notebook.run("full notebook path")

ANSWER: C

Explanation:

Explanation

The answer is dbutils.notebook.run(" full notebook path ")

Here is the full command with additional options.

run(path: String, timeout_seconds: int, arguments: Map): String

1. dbutils.notebook.run("ful-notebook-name", 60, {"argument": "data", "argument2": "data2", ...})

QUESTION NO: 5

You are currently looking at a table that contains data from an e-commerce platform, each row contains a list of items(Item number) that were present in the cart, when the customer makes a change to the cart the entire information is saved as a separate list and appended to an existing list for the duration of the customer session, to identify all the items customer bought you have to make a unique list of items, you were asked to create a unique item's list that was added to the cart by the user, fill in the blanks of below query by choosing the appropriate higher-order function?

Note: See below sample data and expected output.

Schema: cartId INT, items Array

Sample data:

cartId	items
1	[[1,100,200,300], [1,250,300]]
2	[[10,150,200,300], [1,210,300],[350]]

Expected output

cartId	items
1	[1,100,200,300,250]
2	[10,150,200,300,210,350]

Fill in the blanks:

Fill in the blanks:

SELECT cartId, _(_(items)) FROM carts

A. ARRAY_UNION, ARRAY_DISCINT

B. ARRAY_DISTINCT, ARRAY_UNION

C. ARRAY_DISTINCT, FLATTEN

D. FLATTEN, ARRAY_DISTINCT

E. ARRAY_DISTINCT, ARRAY_FLATTEN

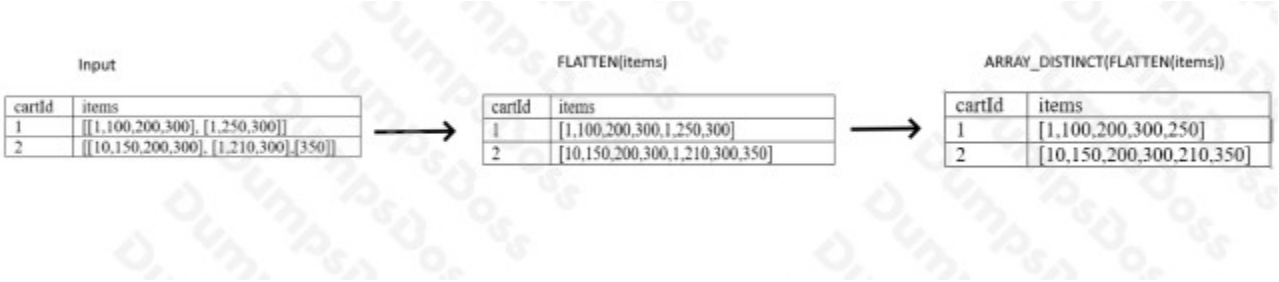
ANSWER: C

Explanation:

Explanation

FLATTEN -> Transforms an array of arrays into a single array.

ARRAY_DISTINCT -> The function returns an array of the same type as the input argument where all duplicate values have been removed.



QUESTION NO: 6

Which of the below commands can be used to drop a DELTA table?

A. DROP DELTA table_name

B. DROP TABLE table_name

C. DROP TABLE table_name FORMAT DELTA

D. DROP table_name

ANSWER: B

QUESTION NO: 7

Which of the following type of tasks cannot setup through a job?

A. Notebook

B. DELTA LIVE PIPELINE

C. Spark Submit

D. Python

E. Databricks SQL Dashboard refresh

ANSWER: E

QUESTION NO: 8

The marketing team is launching a new campaign to monitor the performance of the new campaign for the first two weeks, they would like to set up a dashboard with a refresh schedule to run every 5 minutes, which of the below steps can be taken to reduce of the cost of this refresh over time?

A. Reduce the size of the SQL Cluster size

B. Reduce the max size of auto scaling from 10 to 5

C. Setup the dashboard refresh schedule to end in two weeks

D. Change the spot instance policy from reliability optimized to cost optimized

E. Always use X-small cluster

ANSWER: C

Explanation:

Explanation

The answer is Setup the dashboard refresh schedule to end in two weeks

QUESTION NO: 9

You noticed a colleague is manually copying the data to the backup folder prior to running an up-date command, incase if the update command did not provide the expected outcome so he can use the backup copy to replace table, which Delta Lake feature would you recommend simplifying the process?

A. Use time travel feature to refer old data instead of manually copying

B. Use DEEP CLONE to clone the table prior to update to make a backup copy

C. Use SHADOW copy of the table as preferred backup choice

D. Cloud object storage retains previous version of the file

E. Cloud object storage automatically backups the data

ANSWER: A

Explanation:

Explanation

The answer is, Use time travel feature to refer old data instead of manually copying.

https://databricks.com/blog/2019/02/04/introducing-delta-time-travel-for-large-scale-data-lakes.html

1. SELECT count(*) FROM my_table TIMESTAMP AS OF "2019-01-01"

2. SELECT count(*) FROM my_table TIMESTAMP AS OF date_sub(current_date(), 1)

3. SELECT count(*) FROM my_table TIMESTAMP AS OF "2019-01-01 01:30:00.000"

QUESTION NO: 10

Which of the following techniques structured streaming uses to ensure recovery of failures during stream processing?

A. Checkpointing and Watermarking

B. Write ahead logging and watermarking

C. Checkpointing and write-ahead logging

D. Delta time travel

E. The stream will failover to available nodes in the cluster
Checkpointing and Idempotent sinks

ANSWER: C

Explanation:

The answer is Checkpointing and write-ahead logging.

Structured Streaming uses checkpointing and write-ahead logs to record the offset range of data being processed during each trigger interval.

QUESTION NO: 11

You are noticing job cluster is taking 6 to 8 mins to start which is delaying your job to finish on time, what steps you can take to reduce the amount of time cluster startup time

A. Setup a second job ahead of first job to start the cluster, so the cluster is ready with re-sources when the job starts

B. Use All purpose cluster instead to reduce cluster start up time

C. Reduce the size of the cluster, smaller the cluster size shorter it takes to start the clus-ter

D. Use cluster pools to reduce the startup time of the jobs

E. Use SQL endpoints to reduce the startup time

ANSWER: D

Explanation:

Explanation

The answer is, Use cluster pools to reduce the startup time of the jobs.

Cluster pools allow us to reserve VM's ahead of time, when a new job cluster is created VM are grabbed from the pool. Note: when the VM's are waiting to be used by the cluster only cost incurred is Azure. Databricks run time cost is only billed once VM is allocated to a cluster.

Here is a demo of how to setup and follow some best practices,

https://www.youtube.com/watch?v=FVtITxOabxg&ab_channel=DatabricksAcademy

QUESTION NO: 12

The data engineering team is using a bunch of SQL queries to review data quality and monitor the ETL job every day, which of the following approaches can be used to set up a schedule and auto-mate this process?

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

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

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

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

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

ANSWER: B

Explanation:

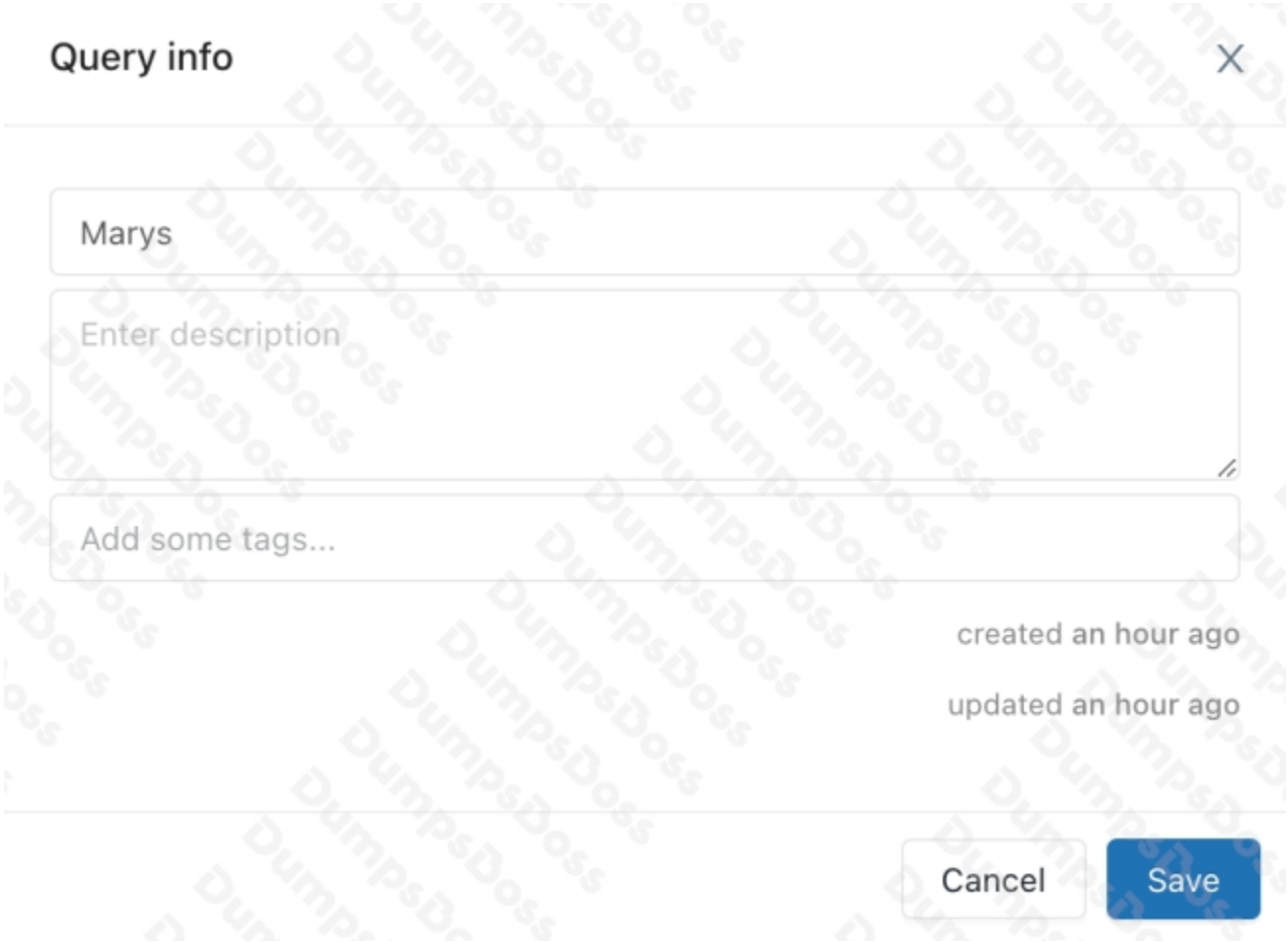
Explanation

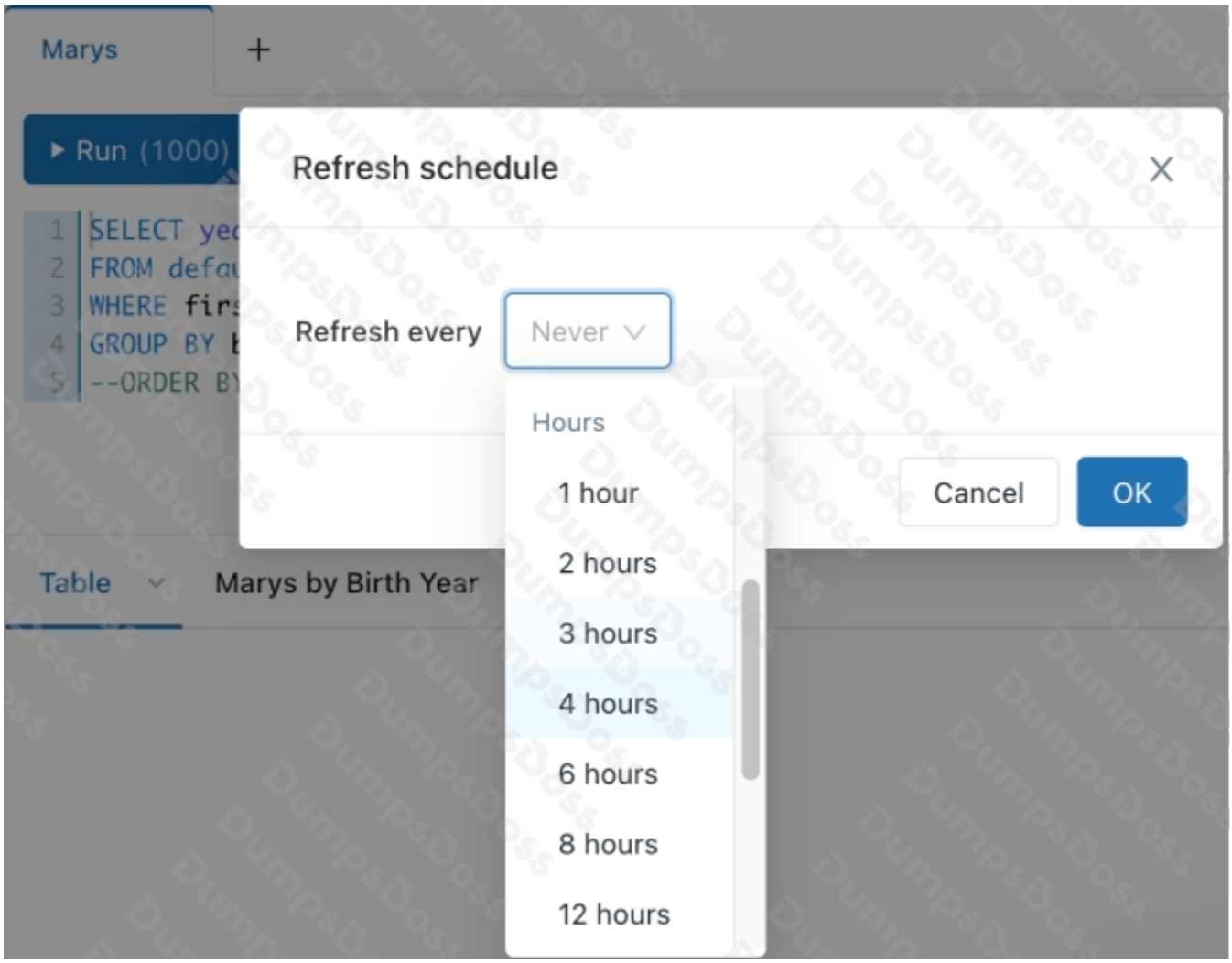
Explanation

Individual queries can be refreshed on a schedule basis,

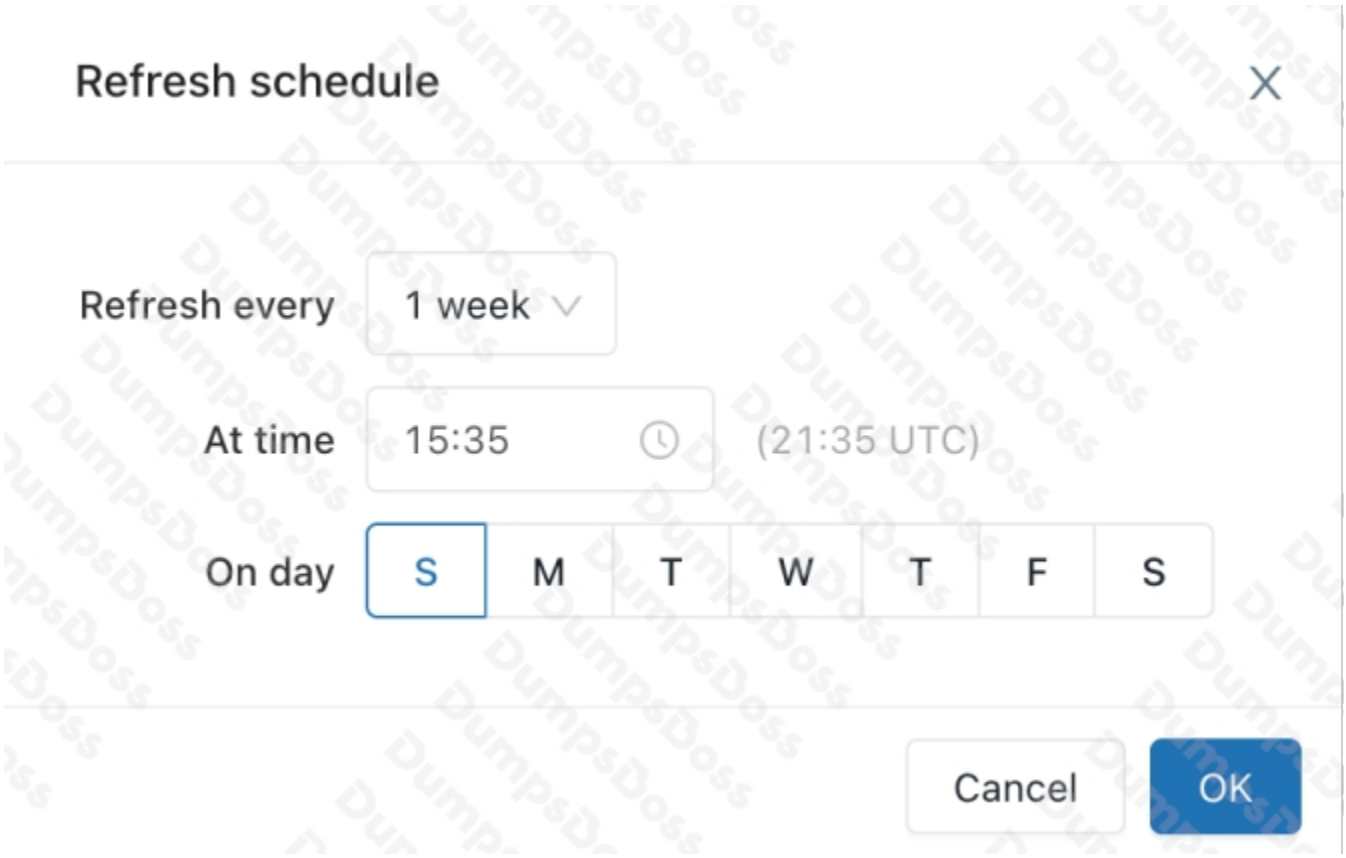
To set the schedule:

1. Click the query info tab.





The picker scrolls and allows you to choose:



Your query will run automatically.

If you experience a scheduled query not executing according to its schedule, you should manually trigger the query to make sure it doesn't fail. However, you should be aware of the following:

If a query execution fails, Databricks SQL retries with a back-off algorithm. The more failures the further away the next retry will be (and it might be beyond the refresh interval).

Refer documentation for additional info,

https://docs.microsoft.com/en-us/azure/databricks/sql/user/queries/schedule-query

QUESTION NO: 13

When using the complete mode to write stream data, how does it impact the target table?

A. Entire stream waits for complete data to write

B. Stream must complete to write the data

C. Target table cannot be updated while stream is pending

D. Target table is overwritten for each batch

E. Delta commits transaction once the stream is stopped

ANSWER: D

Explanation:

Explanation

The answer is Target table is overwritten for each batch

Complete mode - The whole Result Table will be outputted to the sink after every trigger. This is supported for aggregation queries

QUESTION NO: 14

Which of the statements is correct when choosing between lakehouse and Datawarehouse?

A. Traditional Data warehouses have special indexes which are optimized for Machine learning

B. Traditional Data warehouses can serve low query latency with high reliability for BI workloads

C. SQL support is only available for Traditional Datawarehouse's, Lakehouses support Python and Scala

D. Traditional Data warehouses are the preferred choice if we need to support ACID, Lakehouse does not support ACID.

E. Lakehouse replaces the current dependency on data lakes and data warehouses uses an open standard storage format and supports low latency BI workloads.

ANSWER: E

Explanation:

Explanation

The lakehouse replaces the current dependency on data lakes and data warehouses for modern data companies that desire:

• Open, direct access to data stored in standard data formats.

• Indexing protocols optimized for machine learning and data science.

• Low query latency and high reliability for BI and advanced analytics.

QUESTION NO: 15

A team member is leaving the team and he/she is currently the owner of the few tables, instead of transfering the ownership to a user you have decided to transfer the ownership to a group so in the future anyone in the group can manage the permissions rather than a single individual, which of the following commands help you accomplish this?

A. ALTER TABLE table_name OWNER to 'group'

B. TRANSFER OWNER table_name to 'group'

C. GRANT OWNER table_name to 'group'•

D. ALTER OWNER ON table_name to 'group'

E. GRANT OWNER On table_name to 'group'

ANSWER: A

Explanation:

Explanation

The answer is ALTER TABLE table_name OWNER to ‘group’

Assign owner to object