Databricks

Exam Questions Databricks-Certified-Data-Engineer-Associate

Databricks Certified Data Engineer Associate Exam

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
A data analyst has developed a query that runs against Delta table. They want help from the data engineering team to implement a series of tests to ensure the data returned by the query is clean. However, the data engineering team uses Python for its tests rather than SQL.
Which of the following operations could the data engineering team use to run the query and operate with the results in PySpark?

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

Answer: C

Explanation:
from pyspark.sql import SparkSession spark = SparkSession.builder.getOrCreate()
df = spark.sql("SELECT * FROM sales") print(df.count())

NEW QUESTION 4
A data analysis team has noticed that their Databricks SQL queries are running too slowly when connected to their always-on SQL endpoint. They claim that this issue is present when many members of the team are running small queries simultaneously. They ask the data engineering team for help. The data engineering team notices that each of the team's queries uses the same SQL endpoint.
Which of the following approaches can the data engineering team use to improve the latency of the team's queries?

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

Answer: A

Explanation:
When many users are running small queries simultaneously on a SQL endpoint, the database can become overloaded, causing slow query execution times. By increasing the cluster size of the SQL endpoint, the database can handle more simultaneous queries, resulting in faster query execution times.

NEW QUESTION 5
A data engineer is attempting to drop a Spark SQL table my_table. The data engineer wants to delete all table metadata and data.
They run the following command: DROP TABLE IF EXISTS my_table
While the object no longer appears when they run SHOW TABLES, the data files still exist.
Which of the following describes why the data files still exist and the metadata files were deleted?

A. The table's data was larger than 10 GB
B. The table's data was smaller than 10 GB
C. The table was external
D. The table did not have a location
E. The table was managed

Answer:

C

Explanation:
The reason why the data files still exist while the metadata files were deleted is because the table was external. When a table is external in Spark SQL (or in other database systems), it means that the table metadata (such as schema information and table structure) is managed externally, and Spark SQL assumes that the data is managed and maintained outside of the system. Therefore, when you execute a DROP TABLE statement for an external table, it removes only the table metadata from the catalog, leaving the data files intact. On the other hand, for managed tables (option E), Spark SQL manages both the metadata and the data files. When you drop a managed table, it deletes both the metadata and the associated data files, resulting in a complete removal of the table.

NEW QUESTION 6
Which of the following must be specified when creating a new Delta Live Tables pipeline?

A. A key-value pair configuration
B. The preferred DBU/hour cost
C. A path to cloud storage location for the written data
D. A location of a target database for the written data
E. At least one notebook library to be executed

Answer: E

Explanation:
https://docs.databricks.com/en/delta-live-tables/tutorial-pipelines.html

NEW QUESTION 7
Which of the following is a benefit of the Databricks Lakehouse Platform embracing open source technologies?

A. Cloud-specific integrations
B. Simplified governance
C. Ability to scale storage
D. Ability to scale workloads
E. Avoiding vendor lock-in

Answer: E

Explanation:
https://double.cloud/blog/posts/2023/01/break-free-from-vendor-lock-in-with-open-source-tech/

NEW QUESTION 8
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 9
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 10
A data engineer is using the following code block as part of a batch ingestion pipeline to read from a composable table:

```
transactions_df = (spark.read
      .schema(schema)
      .format("delta")
      .table("transactions")
)
```

Which of the following changes needs to be made so this code block will work when the transactions table is a stream source?

A. Replace predict with a stream-friendly prediction function
B. Replace schema(schema) with option ("maxFilesPerTrigger", 1)
C. Replace "transactions" with the path to the location of the Delta table
D. Replace format("delta") with format("stream")
E. Replace spark.read with spark.readStream

Answer: E

Explanation:
https://docs.databricks.com/en/structured-streaming/delta-lake.html

NEW QUESTION 10
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 " youll find the answer.https://spark.apache.org/docs/latest/structured-streaming- programming-guide.html#:~:text=The%20engine%20uses%20checkpointing%20and,being%20processe d%20in%20each%20trigger.

NEW QUESTION 14
A data engineer has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to a data analytics dashboard for a retail use case. The job has a Databricks SQL query that returns the number of store-level records where sales is equal to zero. The data engineer wants their entire team to be notified via a messaging webhook whenever this value is greater than 0.
Which of the following approaches can the data engineer use to notify their entire team via a messaging webhook whenever the number of stores with $0 in sales is greater than zero?

A. They can set up an Alert with a custom template.
B. They can set up an Alert with a new email alert destination.
C. They can set up an Alert with one-time notifications.
D. They can set up an Alert with a new webhook alert destination.
E. They can set up an Alert without notifications.

Answer: D

NEW QUESTION 17
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 22
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 customersTO 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 24
A data engineer needs to use a Delta table as part of a data pipeline, but they do not know if they have the appropriate permissions.
In which of the following locations can the data engineer review their permissions on the table?

A. Databricks Filesystem
B. Jobs
C. Dashboards
D. Repos
E. Data Explorer

Answer: E

NEW QUESTION 28
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 33
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,
    FILTER (exp_employees, i -> i.years_exp > 5) AS exp_employees
FROM stores;
```

A. Option A
B. Option B
C. Option C
D. Option D
E. Option E

Answer: A

NEW QUESTION 38
A data engineer wants to schedule their Databricks SQL dashboard to refresh once per day, but they only want the associated SQL endpoint to be running when it is necessary.
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 ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.
B. They can set up the dashboard's SQL endpoint to be serverless.
C. They can turn on the Auto Stop feature for the SQL endpoint.
D. They can reduce the cluster size of the SQL endpoint.
E. They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.

Answer: C

NEW QUESTION 41
Which of the following commands can be used to write data into a Delta table while avoiding the writing of duplicate records?

A. DROP
B. IGNORE
C. MERGE
D. APPEND
E. INSERT

Answer: C

Explanation:
To write data into a Delta table while avoiding the writing of duplicate records, you can use the MERGE command. The MERGE command in Delta Lake allows you to combine the ability to insert new records and update existing records in a single atomic operation. The MERGE command compares the data being written with the existing data in the Delta table based on specified matching criteria, typically using a primary key or unique identifier. It then performs conditional actions, such as inserting new records or updating existing records, depending on the comparison results. By using the MERGE command, you can handle the prevention of duplicate records in a more controlled and efficient manner. It allows you to synchronize and reconcile data from different sources while avoiding duplication and ensuring data integrity.

NEW QUESTION 42
A data engineer has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to an ELT job. The ELT job has its Databricks SQL query that returns the number of input records containing unexpected NULL values. The data engineer wants their entire team to be notified via a messaging webhook whenever this value reaches 100.
Which of the following approaches can the data engineer use to notify their entire team via a messaging webhook whenever the number of NULL values reaches

100?

A. They can set up an Alert with a custom template.
B. They can set up an Alert with a new email alert destination.
C. They can set up an Alert with a new webhook alert destination.
D. They can set up an Alert with one-time notifications.
E. They can set up an Alert without notifications.

Answer: C

Explanation:
To achieve this, the data engineer can set up an Alert in the Databricks workspace that triggers when the query results exceed the threshold of 100 NULL values. They can create a new webhook alert destination in the Alert's configuration settings and provide the necessary messaging webhook URL to receive notifications. When the Alert is triggered, it will send a message to the configured webhook URL, which will then notify the entire team of the issue.

NEW QUESTION 43
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 48
A data engineer needs to create a table in Databricks using data from a CSV file at location /path/to/csv.
They run the following command:

```
CREATE TABLE new_table

_____
OPTIONS (
  header = "true",
  delimiter = "|"
)
LOCATION "path/to/csv"
```

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

A. None of these lines of code are needed to successfully complete the task
B. USING CSV
C. FROM CSV
D. USING DELTA
E. FROM "path/to/csv"

Answer: B

NEW QUESTION 50
An engineering manager wants to monitor the performance of a recent project using a Databricks SQL query. For the first week following the project's release, the manager wants the query results to be updated every minute. However, the manager is concerned that the compute resources used for the query will be left running and cost the organization a lot of money beyond the first week of the project's release.
Which of the following approaches can the engineering team use to ensure the query does not cost the organization any money beyond the first week of the project's release?

A. They can set a limit to the number of DBUs that are consumed by the SQL Endpoint.
B. They can set the query's refresh schedule to end after a certain number of refreshes.
C. They cannot ensure the query does not cost the organization money beyond the first week of the project's release.
D. They can set a limit to the number of individuals that are able to manage the query's refresh schedule.
E. They can set the query's refresh schedule to end on a certain date in the query scheduler.

Answer: E

Explanation:
If a dashboard is configured for automatic updates, it has a Scheduled button at the top, rather than a Schedule button. To stop automatically updating the dashboard and remove its subscriptions:
Click Scheduled.
In the Refresh every drop-down, select Never.
Click Save. The Scheduled button label changes to Schedule. Source:https://learn.microsoft.com/en-us/azure/databricks/sql/user/dashboards/

NEW QUESTION 51
A data engineer has a Job that has a complex run schedule, and they want to transfer that schedule to other Jobs.
Rather than manually selecting each value in the scheduling form in Databricks, which of the following tools can the data engineer use to represent and submit the schedule programmatically?

A. pyspark.sql.types.DateType
B. datetime
C. pyspark.sql.types.TimestampType
D. Cron syntax
E. There is no way to represent and submit this information programmatically

Answer: D

NEW QUESTION 53
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 Production 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?

A. All datasets will be updated at set intervals until the pipeline is shut dow
B. The compute resources will persist to allow for additional testing.
C. All datasets will be updated once and the pipeline will persist without any processin
D. The compute resources will persist but go unused.
E. All datasets will be updated at set intervals until the pipeline is shut dow
F. The compute resources will be deployed for the update and terminated when the pipeline is stopped.
G. All datasets will be updated once and the pipeline will shut dow
H. The compute resources will be terminated.
I. All datasets will be updated once and the pipeline will shut dow
J. The compute resources will persist to allow for additional testing.

Answer: C

Explanation:
In a Delta Live Table pipeline running in Continuous Pipeline Mode, when you click Start to update the pipeline, the following outcome is expected: All datasets defined using STREAMING LIVE TABLE and LIVE TABLE against Delta Lake table sources will be updated at set intervals. The compute resources will be deployed for the update process and will be active during the execution of the pipeline. The compute resources will be terminated when the pipeline is stopped or shut down. This mode allows for continuous and periodic updates to the datasets as new data arrives or changes in the
underlying Delta Lake tables occur. The compute resources are provisioned and utilized during the update intervals to process the data and perform the necessary operations.

NEW QUESTION 54
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?

A. All datasets will be updated once and the pipeline will shut dow
B. The compute resources will be terminated.
C. All datasets will be updated at set intervals until the pipeline is shut dow
D. The compute resources will persist until the pipeline is shut down.
E. All datasets will be updated once and the pipeline will persist without any processin
F. The compute resources will persist but go unused.
G. All datasets will be updated once and the pipeline will shut dow
H. The compute resources will persist to allow for additional testing.
I. All datasets will be updated at set intervals until the pipeline is shut dow
J. The compute resources will persist to allow for additional testing.

Answer: E

Explanation:
You can optimize pipeline execution by switching between development and production modes. Use the Delta Live Tables Environment Toggle Icon buttons in the Pipelines UI to switch between these two modes. By default, pipelines run in development mode.
When you run your pipeline in development mode, the Delta Live Tables system does the following:
Reuses a cluster to avoid the overhead of restarts. By default, clusters run for two hours when development mode is enabled. You can change this with the pipelines.clusterShutdown.delay setting in the Configure your compute settings.
Disables pipeline retries so you can immediately detect and fix errors. In production mode, the Delta Live Tables system does the following:
Restarts the cluster for specific recoverable errors, including memory leaks and stale credentials.
Retries execution in the event of specific errors, for example, a failure to start a cluster. https://docs.databricks.com/en/delta-live-tables/updates.html#optimize-execution

NEW QUESTION 56
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 60
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?

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

NEW QUESTION 64
Which of the following Structured Streaming queries is performing a hop from a Silver table to a Gold table?
A.

```
(spark.readStream.load(rawSalesLocation)
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```

B.

```
(spark.read.load(rawSalesLocation)
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```

C.

```
(spark.table("sales")
    .withColumn("avgPrice", col("sales") / col("units"))
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```

D.

```
(spark.table("sales")
    .filter(col("units") > 0)
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```

E.

```
(spark.table("sales")
    .groupBy("store")
    .agg(sum("sales"))
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("complete")
    .table("newSales")
)
```

A.

Answer: E

NEW QUESTION 66
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 70
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%2Doriented%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 73
In which of the following file formats is data from Delta Lake tables primarily stored?

A. Delta
B. CSV
C. Parquet
D. JSON
E. A proprietary, optimized format specific to Databricks

Answer: C

Explanation:
https://docs.delta.io/latest/delta-faq.html

NEW QUESTION 76
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 81
A data engineer is designing a data pipeline. The source system generates files in a shared directory that is also used by other processes. As a result, the files should be kept as is and will accumulate in the directory. The data engineer needs to identify which files are new since the previous run in the pipeline, and set up the pipeline to only ingest those new files with each run.
Which of the following tools can the data engineer use to solve this problem?

A. Unity Catalog
B. Delta Lake
C. Databricks SQL
D. Data Explorer
E. Auto Loader

Answer: E

Explanation:
Auto Loader incrementally and efficiently processes new data files as they arrive in cloud storage without any additional

setup.https://docs.databricks.com/en/ingestion/auto-loader/index.html

NEW QUESTION 85
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 89
A data architect has determined that a table of the following format is necessary:

employeeId	startDate	avgRating
a1	2009-01-06	5.5
a2	2018-11-21	7.1
...	...	...

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
)
```

A. Option A
B. Option B
C. Option C
D. Option D
E. Option E

Answer: E

NEW QUESTION 92
.....

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questons and Answers in PDF Format

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