Databricks

Exam Questions Databricks-Certified-Data-Engineer-Associate

Databricks Certified Data Engineer Associate Exam

NEW QUESTION 1
In which of the following scenarios should a data engineer select a Task in the Depends On field of a new Databricks Job Task?

A. When another task needs to be replaced by the new task
B. When another task needs to fail before the new task begins
C. When another task has the same dependency libraries as the new task
D. When another task needs to use as little compute resources as possible
E. When another task needs to successfully complete before the new task begins

Answer: E

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_transactionsINNER 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_transactionsOUTER JOIN SELECT * FROM april_transactions;
D. CREATE TABLE all_transactions AS SELECT * FROM march_transactionsINTERSECT 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 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 of the following lines of code fills in the above blank to successfully complete the task?

A. org.apache.spark.sql.jdbc
B. autoloader
C. DELTA
D. sqlite

E. org.apache.spark.sql.sqlite

Answer: A

Explanation:
CREATE TABLE new_employees_table USING JDBC
OPTIONS (
url "<jdbc_url>",
dbtable "<table_name>", user '<username>', password '<password>'
) AS
SELECT * FROM employees_table_vw https://docs.databricks.com/external-data/jdbc.html#language-sql

NEW QUESTION 6
A data engineer is working with two tables. Each of these tables is displayed below in its entirety.

sales

customer_id	spend	units
a1	28.94	7
a3	874.12	23
a4	8.99	1

favorite_stores

customer_id	store_id
a1	s1
a2	s1
a4	s2

The data engineer runs the following query to join these tables together:

```
SELECT
    sales.customer_id,
    sales.spend,
    favorite_stores.store_id
FROM sales
LEFT JOIN favorite_stores
ON sales.customer_id = favorite_stores.customer_id;
```

Which of the following will be returned by the above query?

A.

customer_id	spend	store_id
a1	28.94	s1
a4	8.99	s2

B.

customer_id	spend	units	store_id
a1	28.94	7	s1
a4	8.99	1	s2

C.

customer_id	spend	store_id
a1	28.94	s1
a3	874.12	NULL
a4	8.99	s2

D.

customer_id	spend	store_id
a1	28.94	s1
a2	NULL	s1
a3	874.12	NULL
a4	8.99	s2

E.

customer_id	spend	store_id
a1	28.94	s1
a2	NULL	s1
a4	8.99	s2

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

E. Option E

Answer: C

NEW QUESTION 7
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 8
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 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 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 11
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 12
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 17
Which of the following is hosted completely in the control plane of the classic Databricks architecture?

A. Worker node
B. JDBC data source
C. Databricks web application
D. Databricks Filesystem
E. Driver node

Answer: C

Explanation:
In the classic Databricks architecture, the control plane includes components like the Databricks web application, the Databricks REST API, and the Databricks Workspace. These components are responsible for managing and controlling the Databricks environment, including cluster provisioning, notebook management, access control, and job scheduling. The other options, such as worker nodes, JDBC data sources, Databricks Filesystem (DBFS), and driver nodes, are typically part of the data plane or the execution environment, which is separate from the control plane. Worker nodes are responsible for executing tasks and computations,

JDBC data sources are used to connect to external databases, DBFS is a distributed file system for data storage, and driver nodes are responsible for coordinating the execution of Spark jobs.

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 26
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 30
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 31
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 33
A data engineer runs a statement every day to copy the previous day's sales into the table transactions. Each day's sales are in their own file in the location "/transactions/raw".
Today, the data engineer runs the following command to complete this task:

```
COPY INTO transactions
FROM "/transactions/raw"
FILEFORMAT = PARQUET;
```

After running the command today, the data engineer notices that the number of records in table transactions has not changed.
Which of the following describes why the statement might not have copied any new records into the table?

A. The format of the files to be copied were not included with the FORMAT_OPTIONS keyword.
B. The names of the files to be copied were not included with the FILES keyword.
C. The previous day's file has already been copied into the table.
D. The PARQUET file format does not support COPY INTO.
E. The COPY INTO statement requires the table to be refreshed to view the copied rows.

Answer: C

Explanation:
https://docs.databricks.com/en/ingestion/copy-into/index.html The COPY
INTO SQL command lets you load data from a file location into a Delta table. This is a re- triable and idempotent operation; files in the source location that have already been loaded are skipped. if there are no new records, the only consistent choice is C no new files were loaded because already loaded files were skipped.

NEW QUESTION 36
Which of the following describes the relationship between Gold tables and Silver tables?

A. Gold tables are more likely to contain aggregations than Silver tables.
B. Gold tables are more likely to contain valuable data than Silver tables.
C. Gold tables are more likely to contain a less refined view of data than Silver tables.
D. Gold tables are more likely to contain more data than Silver tables.
E. Gold tables are more likely to contain truthful data than Silver tables.

Answer: A

Explanation:
In some data processing pipelines, especially those following a typical "Bronze-Silver-Gold" data lakehouse architecture, Silver tables are often considered a more refined version of the raw or Bronze data. Silver tables may include data cleansing, schema enforcement, and some initial transformations. Gold tables, on the other hand, typically represent a stage where data is further enriched, aggregated, and processed to provide valuable insights for analytical purposes. This could indeed involve more aggregations compared to Silver tables.

NEW QUESTION 40
A data engineer has a Python notebook in Databricks, but they need to use SQL to accomplish a specific task within a cell. They still want all of the other cells to use Python without making any changes to those cells.
Which of the following describes how the data engineer can use SQL within a cell of their Python notebook?

A. It is not possible to use SQL in a Python notebook
B. They can attach the cell to a SQL endpoint rather than a Databricks cluster
C. They can simply write SQL syntax in the cell
D. They can add %sql to the first line of the cell
E. They can change the default language of the notebook to SQL

Answer: D

NEW QUESTION 45
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:
(f"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 46
A single Job runs two notebooks as two separate tasks. A data engineer has noticed that one of the notebooks is running slowly in the Job's current run. The data engineer asks a tech lead for help in identifying why this might be the case.
Which of the following approaches can the tech lead use to identify why the notebook is running slowly as part of the Job?

A. They can navigate to the Runs tab in the Jobs UI to immediately review the processing notebook.
B. They can navigate to the Tasks tab in the Jobs UI and click on the active run to review the processing notebook.
C. They can navigate to the Runs tab in the Jobs UI and click on the active run to review the processing notebook.
D. There is no way to determine why a Job task is running slowly.
E. They can navigate to the Tasks tab in the Jobs UI to immediately review the processing notebook.

Answer: C

Explanation:
The job run details page contains job output and links to logs, including information about the success or failure of each task in the job run. You can access job run details from the Runs tab for the job. To view job run details from the Runs tab, click the link for the run in the Start time column in the runs list view. To return to the Runs tab for the job, click the Job ID value.
If the job contains multiple tasks, click a task to view task run details, including: the cluster that ran the task
the Spark UI for the task logs for the task
metrics for the task
https://docs.databricks.com/en/workflows/jobs/monitor-job-runs.html#job-run-details

NEW QUESTION 47
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 50
Which of the following data workloads will utilize a Gold table as its source?

A. A job that enriches data by parsing its timestamps into a human-readable format
B. A job that aggregates uncleaned data to create standard summary statistics
C. A job that cleans data by removing malformatted records
D. A job that queries aggregated data designed to feed into a dashboard
E. A job that ingests raw data from a streaming source into the Lakehouse

Answer: D

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 55
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 59
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 62
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 63
.....

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