Databricks.Databricks-Certified-Professional-Data-Engineer.v2023-02-09.q22

Exam Code:	Databricks-Certified-Professional-Data-Engineer
Exam Name:	Databricks Certified Professional Data Engineer Exam
Certification Provider:	Databricks
Free Question Number:	22
Version:	v2023-02-09
# of views:	832
# of Questions views:	220
https://www.freepdfdumps.com/Databricks.Databricks-Certified-Professional-Data-Engineer.v2023-02-09.q22.html	

NEW QUESTION: 1
Which of the following describes a scenario in which a data engineer will want to use a Job cluster instead of
an all-purpose cluster?
A. An ad-hoc analytics report needs to be developed while minimizing compute costs
B. A Databricks SQL query needs to be scheduled for upward reporting
C. A data engineer needs to manually investigate a production error
D. An automated workflow needs to be run every 30 minutes
E. A data team needs to collaborate on the development of a machine learning model
Answer: D (LEAVE A REPLY)

NEW QUESTION: 2
A data architect has determined that a table of the following format is necessary:
Which of the following code blocks uses SQL DDL commands to create an empty Delta table in the above
format regardless of whether a table already exists with this name?
A. 1. CREATE TABLE table_name AS
2. SELECT id STRING, birthDate DATE, avgRating FLOAT
B. 1. CREATE TABLE IF NOT EXISTS table_name (id STRING, birthDate DATE, avgRating FLOAT)
C. 1. CREATE OR REPLACE TABLE table_name AS
2. SELECT id STRING, birthDate DATE, avgRating FLOAT USING DELTA
D. 1. CREATE OR REPLACE TABLE table_name (id STRING, birthDate DATE, avgRating FLOAT)
E. 1. CREATE OR REPLACE TABLE table_name

2. WITH COLUMNS (id STRING, birthDate DATE, avgRating FLOAT) USING DELTA
Answer: D (LEAVE A REPLY)

NEW QUESTION: 3
Which of the following Structured Streaming queries is performing a hop from a Bronze table to a Silver
table?
A. 1. (spark.read.load(rawSalesLocation)
2. .writeStream
3. .option("checkpointLocation", checkpointPath)
4. .outputMode("append")
5. .table("uncleanedSales")
6.)
B. 1. (spark.table("sales")
2. .agg(sum("sales"),
3. sum("units"))
4. .writeStream
5. .option("checkpointLocation", checkpointPath)
6. .outputMode("complete")
7. .table("aggregatedSales")
8.)
C. 1. (spark.readStream.load(rawSalesLocation)
2. .writeStream
3. .option("checkpointLocation", checkpointPath)
4. .outputMode("append")
5. .table("uncleanedSales")
6.)
D. 1. (spark.table("sales")
2. .withColumn("avgPrice", col("sales") / col("units"))
3. .writeStream
4. .option("checkpointLocation", checkpointPath)
5. .outputMode("append")
6. .table("cleanedSales")
7.)
E. 1. (spark.table("sales")
2. .groupBy("store")
3. .agg(sum("sales"))
4. .writeStream
5. .option("checkpointLocation", checkpointPath)
6. .outputMode("complete")
7. .table("aggregatedSales")

8.)
Answer: D (LEAVE A REPLY)

NEW QUESTION: 4
A junior data engineer has ingested a JSON file into a table raw_table with the following schema:
1. cart_id STRING,
2. items ARRAY<item_id:STRING>
The junior data engineer would like to unnest the items column in raw_table to result in a new table with the
following schema:
1.cart_id STRING,
2.item_id STRING
Which of the following commands should the junior data engineer run to complete this task?
A. 1. SELECT cart_id, flatten(items) AS item_id
2. FROM raw_table;
B. 1. SELECT cart_id, filter(items) AS item_id
2. FROM raw_table;
C. 1. SELECT cart_id, slice(items) AS item_id
2. FROM raw_table;
D. 1. SELECT cart_id, explode(items) AS item_id
2. FROM raw_table;
E. 1. SELECT cart_id, reduce(items) AS item_id
2. FROM raw_table;
Answer: D (LEAVE A REPLY)

NEW QUESTION: 5
A data engineer has written the following query:
1. SELECT *
2. FROM json.`/path/to/json/file.json`;
The data engineer asks a colleague for help to convert this query for use in a Delta Live Tables (DLT)
pipeline. The query should create the first table in the DLT pipeline.
Which of the following describes the change the colleague needs to make to the query?
A. They need to add a live. prefix prior to json. in the FROM line
B. They need to add a CREATE LIVE TABLE table_name AS line at the beginning of the query
C. They need to add the cloud_files(...) wrapper to the JSON file path
D. They need to add a CREATE DELTA LIVE TABLE table_name AS line at the beginning of the query

E. They need to add a COMMENT line at the beginning of the query
Answer: (SHOW ANSWER)

NEW QUESTION: 6
A data engineer has ingested data from an external source into a PySpark DataFrame raw_df. They need to
briefly make this data available in SQL for a data analyst to perform a quality assurance check on the data.
Which of the following commands should the data engineer run to make this data available in SQL for only
the remainder of the Spark session?
A. raw_df.write.save("raw_df")
B. raw_df.saveAsTable("raw_df")
C. raw_df.createTable("raw_df")
D. There is no way to share data between PySpark and SQL
E. raw_df.createOrReplaceTempView("raw_df")
Answer: E (LEAVE A REPLY)

NEW QUESTION: 7
A data engineer is overwriting data in a table by deleting the table and recreating the table. Another data
engineer suggests that this is inefficient and the table should simply be overwritten instead.
Which of the following reasons to overwrite the table instead of deleting and recreating the table is incorrect?
A. Overwriting a table allows for concurrent queries to be completed while in progress
B. Overwriting a table is efficient because no files need to be deleted
C. Overwriting a table maintains the old version of the table for Time Travel
D. Overwriting a table results in a clean table history for logging and audit purposes
E. Overwriting a table is an atomic operation and will not leave the table in an unfinished state
Answer: D (LEAVE A REPLY)

NEW QUESTION: 8
You are asked to create a model to predict the total number of monthly subscribers for a specific magazine.
You are provided with 1 year's worth of subscription and payment data, user demographic data, and 10 years
worth of content of the magazine (articles and pictures). Which algorithm is the most appropriate for building
a predictive model for subscribers?
A. Decision trees

B. TF-IDF
C. Logistic regression
D. Linear regression
Answer: D (LEAVE A REPLY)

NEW QUESTION: 9
A data engineering team has been using a Databricks SQL query to monitor the performance of an ELT job.
The ELT job is triggered by a specific number of input records being ready to process. The Databricks SQL
query returns the number of minutes since the job's most recent runtime.
Which of the following approaches can enable the data engineering team to be notified if the ELT job has not
been run in an hour?
A. They can set up an Alert for the query to notify when the ELT job fails
B. They can set up an Alert for the accompanying dashboard to notify them if the returned value is greater
than 60
C. They can set up an Alert for the accompanying dashboard to notify when it has not re-freshed in 60
minutes
D. They can set up an Alert for the query to notify them if the returned value is greater than 60
E. This type of alerting is not possible in Databricks
Answer: (SHOW ANSWER)

NEW QUESTION: 10
Which of the following is a Continuous Probability Distributions?
A. Poisson probability distribution
B. Binomial probability distribution
C. Normal probability distribution
D. Negative binomial distribution
Answer: C (LEAVE A REPLY)

NEW QUESTION: 11
A data engineer wants to horizontally combine two tables as a part of a query. They want to use a shared
column as a key column, and they only want the query result to contain rows whose value in the key column is
present in both tables.
Which of the following SQL commands can they use to accomplish this task?

A. LEFT JOIN
B. OUTER JOIN
C. MERGE
D. UNION
E. INNER JOIN
Answer: E (LEAVE A REPLY)

NEW QUESTION: 12
Which of the following data workloads will utilize a Bronze table as its source?
A. A job that aggregates cleaned data to create standard summary statistics
B. A job that develops a feature set for a machine learning application
C. A job that enriches data by parsing its timestamps into a human-readable format
D. A job that queries aggregated data to publish key insights into a dashboard
E. A job that ingests raw data from a streaming source into the Lakehouse
Answer: (SHOW ANSWER)

NEW QUESTION: 13
Which of the following commands will return records from an existing Delta table my_table where duplicates
have been removed?
A. 1. MERGE INTO my_table a
2. USING new_records b ON a.id = b.id
3. WHEN NOT MATCHED
4. THEN INSERT *;
B. 1. SELECT DISTINCT *
2. FROM my_table;
C. 1. DROP DUPLICATES
2. FROM my_table;
D. 1. MERGE INTO my_table a
2. USING new_records b;
E. 1. SELECT *
2. FROM my_table
3. WHERE duplicate = False;
Answer: (SHOW ANSWER)

NEW QUESTION: 14
Two junior data engineers are authoring separate parts of a single data pipeline notebook. They are working on
separate Git branches so they can pair program on the same notebook simultaneously. A senior data engineer

experienced in Databricks suggests there is a better alternative for this type of collaboration.
Which of the following supports the senior data engineer's claim?
A. Databricks Notebooks support the use of multiple languages in the same notebook
B. Databricks Notebooks support real-time co-authoring on a single notebook
C. Databricks Notebooks support the creation of interactive data visualizations
D. Databricks Notebooks support automatic change-tracking and versioning
E. Databricks Notebooks support commenting and notification comments
Answer: B (LEAVE A REPLY)

NEW QUESTION: 15
Projecting a multi-dimensional dataset onto which vector has the greatest variance?
A. first principal component
B. first eigenvector
C. not enough information given to answer
D. second eigenvector
E. second principal component
Answer: A (LEAVE A REPLY)
Explanation
The method based on principal component analysis (PCA) evaluates the features according to the projection of
the largest eigenvector of the correlation matrix on the initial dimensions, the method based on Fisher's linear
discriminant analysis evaluates. Them according to the magnitude of the components of the discriminant
vector.
The first principal component corresponds to the greatest variance in the data, by definition. If we project the
data onto the first principal component line, the data is more spread out (higher variance) than if projected onto
any other line, including other principal components.

NEW QUESTION: 16
What are the advantages of the Hashing Features?
A. Requires the less memory
B. Less pass through the training data
C. Easily reverse engineer vectors to determine which original feature mapped to a vector location
Answer: A,B (LEAVE A REPLY)
Explanation

SGD-based classifiers avoid the need to predetermine vector size by simply picking a reasonable size and
shoehorning the training data into vectors of that size. This approach is known as feature hashing. The
shoehorning is done by picking one or more locations by using a hash of the name of the variable for
continuous variables or a hash of the variable name and the category name or word for categorical, text*like, or
word-like data.
This hashed feature approach has the distinct advantage of requiring less memory and one less pass through
the training data, but it can make it much harder to reverse engineer vectors to determine which original
feature mapped to a vector location. This is because multiple features may hash to the same location. With
large vectors or with multiple locations per feature, this isn't a problem for accuracy but it can make it hard to
understand what a classifier is doing.
An additional benefit of feature hashing is that the unknown and unbounded vocabularies typical of word-like
variables aren't a problem.

NEW QUESTION: 17
A data engineer needs to dynamically create a table name string using three Python variables: region, store,
and year. An example of a table name is below when region = "nyc", store = "100", and year = "2021":
nyc100_sales_2021

Which of the following commands should the data engineer use to construct the table name in Py-thon?
A. "{region}+{store}+_sales_+{year}"
B. "{region}+{store}+"_sales_"+{year}"
C. f"{region}+{store}+_sales_+{year}"
D. "{region}{store}_sales_{year}"
E. f"{region}{store}_sales_{year}"
Answer: E (LEAVE A REPLY)

NEW QUESTION: 18
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 Triggered 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 down. The compute resources will be
terminated
B. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will
persist after the pipeline is stopped to allow for additional testing
C. All datasets will be updated continuously and the pipeline will not shut down. The compute resources
will persist with the pipeline
D. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will
be deployed for the update and terminated when the pipeline is stopped
E. All datasets will be updated once and the pipeline will shut down. The compute resources will persist to
allow for additional testing
Answer: E (LEAVE A REPLY)

NEW QUESTION: 19
Which of the following describes a benefit of a data lakehouse that is unavailable in a traditional data
warehouse?
A. A data lakehouse captures snapshots of data for version control purposes
B. A data lakehouse provides a relational system of data management

C. A data lakehouse enables both batch and streaming analytics
D. A data lakehouse couples storage and compute for complete control
E. A data lakehouse utilizes proprietary storage formats for data
Answer: C (LEAVE A REPLY)

NEW QUESTION: 20
A denote the event 'student is female' and let B denote the event 'student is French'. In a class of 100 students
suppose 60 are French, and suppose that 10 of the French students are females. Find the probability that if I
pick a French student, it will be a girl, that is, find P(A|B).
A. 1/3
B. 2/3
C. 1/6
D. 2/6
Answer: C (LEAVE A REPLY)
Explanation
Since 10 out of 100 students are both French and female, then
P(AandB)=10100
Also. 60 out of the 100 students are French, so
P(B)=60100
So the required probability is:
P(A|B)=P(AandB)P(B)=10/10060/100=16

NEW QUESTION: 21
A dataset has been defined using Delta Live Tables and includes an expectations clause:
1. CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01')
What is the expected behaviour 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 added to the target dataset and flagged as in-valid in a field added to the target dataset
D. Records that violate the expectation are dropped from the target dataset and loaded into a quarantine table
E. Records that violate the expectation are added to the target dataset and recorded as invalid in the event log
Answer: E (LEAVE A REPLY)

NEW QUESTION: 22
Suppose there are three events then which formula must always be equal to P(E1|E2,E3)?
A. P(E1,E2,E3)P(E1)/P(E2:E3)
B. P(E1,E2;E3)/P(E2,E3)
C. P(E1,E2|E3)P(E2|E3)P(E3)
D. P(E1,E2|E3)P(E3)
E. P(E1,E2,E3)P(E2)P(E3)
Answer: B (LEAVE A REPLY)
Explanation
This is an application of conditional probability: P(E1,E2)=P(E1|E2)P(E2). so
P(E1|E2) = P(E1.E2)/P(E2)
P(E1,E2,E3)/P(E2,E3)
If the events are A and B respectively, this is said to be "the probability of A given B"
It is commonly denoted by P(A|B):or sometimes PB(A). In case that both "A" and "B" are categorical
variables, conditional probability table is typically used to represent the conditional probability.