

Hot Databricks-Certified-Professional-Data-Engineer Actual Test Answers & Leading Provider in Qualification Exams & Practical Updated Databricks-Certified-Professional-Data-Engineer Dumps

Databricks Certification
Data Engineering Certification | Question - 1

Which of the following line of code returns approximately 1000 rows, some of them potentially being duplicates, from 2000-row Dataframe *transactionDF* that only has unique rows ?

www.learntospark.com

<code>transactionDF.first(1000)</code>	6%
<code>transactionDF.take(1000).distinct()</code>	72%
<code>transactionDF.sample(True, 0.5)</code>	17%
<code>transactionDF.sample(False, 0.5)</code>	4%
<code>transactionDF.sample(True, 0.5, force=True)</code>	2%

APACHE Spark
www.learntospark.com
Apache Spark Tutorial

Our Databricks Certified Professional Data Engineer Exam (Databricks-Certified-Professional-Data-Engineer) exam questions are being offered in three easy-to-use and compatible formats. These Databricks Certified Professional Data Engineer Exam (Databricks-Certified-Professional-Data-Engineer) exam dumps formats offer a user-friendly interface and are compatible with all devices, operating systems, and browsers. The ExamDumps VCE Databricks Certified Professional Data Engineer Exam (Databricks-Certified-Professional-Data-Engineer) PDF questions file contains real and valid Databricks Databricks-Certified-Professional-Data-Engineer exam questions that assist you in Databricks-Certified-Professional-Data-Engineer exam dumps preparation and boost the candidate's confidence to pass the challenging Databricks Certified Professional Data Engineer Exam (Databricks-Certified-Professional-Data-Engineer) exam easily.

Databricks Certified Professional Data Engineer certification is designed for data engineers who work with the Databricks platform and have a deep understanding of data engineering concepts. Databricks Certified Professional Data Engineer Exam certification exam tests the candidate's ability to design, build, and maintain data pipelines using Databricks, as well as their knowledge of data modeling, data warehousing, and data governance. Databricks Certified Professional Data Engineer Exam certification is recognized globally and indicates that the candidate has the skills and expertise needed to work with Databricks.

>> **Databricks-Certified-Professional-Data-Engineer Actual Test Answers** <<

Updated Databricks-Certified-Professional-Data-Engineer Dumps - Pass Databricks-Certified-Professional-Data-Engineer Rate

Once you have practiced on our Databricks Certified Professional Data Engineer Exam test questions, the system will automatically memorize and analyze all your practice. You must finish the model test in limited time. There have a timer on the right of the interface. Once you begin to do the exercises of the Databricks-Certified-Professional-Data-Engineer test guide, the timer will start to work and count down. If you don't finish doing the exercises, all your exercises of the Databricks-Certified-Professional-Data-Engineer Exam Questions will be delivered automatically. Then the system will generate a report according to your performance. You will clearly know where you are good at or not. Then you can make your own learning plans based on the report of the Databricks-Certified-Professional-Data-Engineer test guide. Also, you will do more practices that you are not good at until you completely have no problem.

Databricks, a unified analytics platform provider that helps organizations process and analyze large data sets, offers a certification exam for data engineers called the Databricks Certified Professional Data Engineer. Databricks-Certified-Professional-Data-Engineer Exam is designed to test the skills and knowledge of data engineers in building and managing data pipelines, ETL processes, and data architectures within the Databricks platform. Databricks Certified Professional Data Engineer Exam certification is intended to validate the expertise of data engineers in implementing and managing data projects, and to demonstrate their competency in using the Databricks platform.

Databricks Certified Professional Data Engineer Exam Sample Questions (Q179-Q184):

NEW QUESTION # 179

You were asked to create a table that can store the below data, orderTime is a timestamp but the finance team when they query this data normally prefer the orderTime in date format, you would like to create a calculated column that can convert the orderTime column timestamp datatype to date and store it, fill in the blank to complete the DDL.

- A. GENERATED DEFAULT AS (CAST(orderTime as DATE))
- B. AS (CAST(orderTime as DATE))
- C. Delta lake does not support calculated columns, value should be inserted into the table as part of the ingestion process
- D. AS DEFAULT (CAST(orderTime as DATE))
- E. **GENERATED ALWAYS AS (CAST(orderTime as DATE))**
Correct)

Answer: E

Explanation:

Explanation

The answer is, GENERATED ALWAYS AS (CAST(orderTime as DATE))

<https://docs.microsoft.com/en-us/azure/databricks/delta/delta-batch#--use-generated-columns> Delta Lake supports generated columns which are a special type of columns whose values are automatically generated based on a user-specified function over other columns in the Delta table. When you write to a table with generated columns and you do not explicitly provide values for them, Delta Lake automatically computes the values.

Note: Databricks also supports partitioning using generated column

NEW QUESTION # 180

A data engineer is using Lakeflow Spark Declarative Pipelines Expectations to track the data quality of incoming sensor data. Periodically, sensors send bad readings that are out of range, and the team is currently flagging those rows with a warning and writing them to the silver table along with the good data. They have been given a new requirement: the bad rows need to be quarantined in a separate quarantine table and no longer included in the silver table.

This is the existing code for the silver table:

```
@dlt.table
```

```
@dlt.expect( "valid_sensor_reading ", "reading < 120 ")
```

```
def silver_sensor_readings():
```

```
return spark.readStream.table( "bronze_sensor_readings ")
```

Which code will satisfy the requirements?

- A. `@dlt.table`
`@dlt.expect("valid_sensor_reading ", "reading < 120 ")`
`def silver_sensor_readings():`
`return spark.readStream.table("bronze_sensor_readings ")`
`@dlt.table`
`@dlt.expect("invalid_sensor_reading ", "reading >= 120 ")`
`def quarantine_sensor_readings():`
`return spark.readStream.table("bronze_sensor_readings ")`
- B. `@dlt.table`
`@dlt.expect_or_drop("valid_sensor_reading ", "reading < 120 ")`
`def silver_sensor_readings():`
`return spark.readStream.table("bronze_sensor_readings ")`
`@dlt.table`
`@dlt.expect("invalid_sensor_reading ", "reading < 120 ")`
`def quarantine_sensor_readings():`
`return spark.readStream.table("bronze_sensor_readings ")`
- C. `@dlt.table`
`@dlt.expect_or_drop("valid_sensor_reading ", "reading < 120 ")`
`def silver_sensor_readings():`
`return spark.readStream.table("bronze_sensor_readings ")`
- D. **`@dlt.table`**

```
@dlt.expect_or_drop("valid_sensor_reading", "reading < 120 ")
def silver_sensor_readings():
    return spark.readStream.table("bronze_sensor_readings ")
@dlt.table
@dlt.expect("invalid_sensor_reading", "reading >= 120 ")
def quarantine_sensor_readings():
    return spark.readStream.table("bronze_sensor_readings ")
```

Answer: D

Explanation:

Databricks documents that expect retains invalid records in the target dataset, while expect_or_drop drops invalid records before writing to the target. Therefore, the silver table must use expect_or_drop so bad records are excluded from silver. (Databricks Documentation) Databricks also documents a quarantine pattern in which invalid records are separated for downstream processing, but the fully documented pattern uses an intermediate quarantine dataset with an is_quarantined flag and then derives valid and invalid paths from it. None of the listed options exactly matches the official quarantine pattern. As written, option B is the closest intended answer because it at least creates a separate quarantine table and removes invalid rows from silver, but strictly speaking, the documented quarantine implementation is more explicit than any option shown here. (Databricks Documentation)

NEW QUESTION # 181

The Databricks CLI is used to trigger a run of an existing job by passing the job_id parameter. The response indicating the job run request was submitted successfully includes a field run_id. Which statement describes what the number alongside this field represents?

- A. The job_id and number of times the job has been run are concatenated and returned.
- B. The job_id is returned in this field.
- C. The number of times the job definition has been run in this workspace.
- **D. The globally unique ID of the newly triggered run.**

Answer: D

Explanation:

* Exact extract: "run_id: The canonical identifier of a run."

References: Databricks Jobs API/CLI response fields.

NEW QUESTION # 182

A data ingestion task requires a one-TB JSON dataset to be written out to Parquet with a target part-file size of 512 MB. Because Parquet is being used instead of Delta Lake, built-in file-sizing features such as Auto-Optimize & Auto-Compaction cannot be used. Which strategy will yield the best performance without shuffling data?

- A. Set spark.sql.adaptive.advisoryPartitionSizeInBytes to 512 MB bytes, ingest the data, execute the narrow transformations, coalesce to 2,048 partitions (1TB*1024*1024/512), and then write to parquet.
- **B. Set spark.sql.shuffle.partitions to 2,048 partitions (1TB*1024*1024/512), ingest the data, execute the narrow transformations, optimize the data by sorting it (which automatically repartitions the data), and then write to parquet.**
- C. Ingest the data, execute the narrow transformations, repartition to 2,048 partitions (1TB* 1024*1024/512), and then write to parquet.
- D. Set spark.sql.shuffle.partitions to 512, ingest the data, execute the narrow transformations, and then write to parquet.
- E. Set spark.sql.files.maxPartitionBytes to 512 MB, ingest the data, execute the narrow transformations, and then write to parquet.

Answer: B

NEW QUESTION # 183

A data engineering team is setting up deployment automation. To deploy workspace assets remotely using the Databricks CLI command, they must configure it with proper authentication.

Which authentication approach will provide the highest level of security ?

