DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-ENGINEER Q&As

Databricks Certified Professional Data Engineer Exam

Pass Databricks DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-ENGINEER Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.passapply.com/databricks-certified-professional-data-engineer.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Databricks Official Exam Center

QUESTION 1

When evaluating the Ganglia Metrics for a given cluster with 3 executor nodes, which indicator would signal proper utilization of the VM\\'s resources?

A. The five Minute Load Average remains consistent/flat

B. Bytes Received never exceeds 80 million bytes per second

C. Network I/O never spikes

D. Total Disk Space remains constant

E. CPU Utilization is around 75%

Correct Answer: E

In the context of cluster performance and resource utilization, a CPU utilization rate of around 75% is generally considered a good indicator of efficient resource usage. This level of CPU utilization suggests that the cluster is being effectively

used without being overburdened or underutilized.

A consistent 75% CPU utilization indicates that the cluster\\'s processing power is being effectively employed while leaving some headroom to handle spikes in workload or additional tasks without maxing out the CPU, which could lead to

performance degradation.

A five Minute Load Average that remains consistent/flat (Option A) might indicate underutilization or a bottleneck elsewhere.

Monitoring network I/O (Options B and C) is important, but these metrics alone don\\'t provide a complete picture of resource utilization efficiency. Total Disk Space (Option D) remaining constant is not necessarily an indicator of proper

resource utilization, as it\\'s more related to storage rather than computational efficiency.

References:

Ganglia Monitoring System: Ganglia Documentation

Databricks Documentation on Monitoring: Databricks Cluster Monitoring

QUESTION 2

A data engineer is testing a collection of mathematical functions, one of which calculates the area under a curve as described by another function.

Which kind of the test does the above line exemplify?

A. Integration

B. Unit

C. Manual

D. functional

Correct Answer: B

A unit test is designed to verify the correctness of a small, isolated piece of code, typically a single function. Testing a mathematical function that calculates the area under a curve is an example of a unit test because it is testing a specific,

individual function to ensure it operates as expected.

References:

Software Testing Fundamentals: Unit Testing

QUESTION 3

The data engineer team is configuring environment for development testing, and production before beginning migration on a new data pipeline. The team requires extensive testing on both the code and data resulting from code execution, and the team want to develop and test against similar production data as possible.

A junior data engineer suggests that production data can be mounted to the development testing environments, allowing pre production code to execute against production data. Because all users have

Admin privileges in the development environment, the junior data engineer has offered to configure permissions and mount this data for the team.

Which statement captures best practices for this situation?

A. Because access to production data will always be verified using passthrough credentials it is safe to mount data to any Databricks development environment.

B. All developer, testing and production code and data should exist in a single unified workspace; creating separate environments for testing and development further reduces risks.

C. In environments where interactive code will be executed, production data should only be accessible with read permissions; creating isolated databases for each environment further reduces risks.

D. Because delta Lake versions all data and supports time travel, it is not possible for user error or malicious actors to permanently delete production data, as such it is generally safe to mount production data anywhere.

Correct Answer: C

The best practice in such scenarios is to ensure that production data is handled securely and with proper access controls. By granting only read access to production data in development and testing environments, it mitigates the risk of

unintended data modification. Additionally, maintaining isolated databases for different environments helps to avoid accidental impacts on production data and systems.

References:

Databricks best practices for securing data:

https://docs.databricks.com/security/index.html

QUESTION 4

The data engineering team maintains a table of aggregate statistics through batch nightly updates. This includes total sales for the previous day alongside totals and averages for a variety of time periods including the 7 previous days, year-todate, and quarter-to-date. This table is named store_saies_summary and the schema is as follows:

```
store_id INT, total_sales_qtd FLOAT, avg_daily_sales_qtd FLOAT,  total_sales_ytd FLOAT,
avg_daily_sales_ytd FLOAT, previous_day_sales FLOAT, total_sales_7d FLOAT, avg_daily_sales_7d
FLOAT, updated TIMESTAMP
```

The table daily_store_sales contains all the information needed to update store_sales_summary. The schema for this table is:

store_id INT, sales_date DATE, total_sales FLOAT

If daily_store_sales is implemented as a Type 1 table and the total_sales column might be adjusted after manual data auditing, which approach is the safest to generate accurate reports in the store_sales_summary table?

A. Implement the appropriate aggregate logic as a batch read against the daily_store_sales table and overwrite the store_sales_summary table with each Update.

B. Implement the appropriate aggregate logic as a batch read against the daily_store_sales table and append new rows nightly to the store_sales_summary table.

C. Implement the appropriate aggregate logic as a batch read against the daily_store_sales table and use upsert logic to update results in the store_sales_summary table.

D. Implement the appropriate aggregate logic as a Structured Streaming read against the daily_store_sales table and use upsert logic to update results in the store_sales_summary table.

E. Use Structured Streaming to subscribe to the change data feed for daily_store_sales and apply changes to the aggregates in the store_sales_summary table with each update.

Correct Answer: E

The daily_store_sales table contains all the information needed to update store_sales_summary. The schema of the table is: store_id INT, sales_date DATE, total_sales FLOAT The daily_store_sales table is implemented as a Type 1 table, which means that old values are overwritten by new values and no history is maintained. The total_sales column might be adjusted after manual data auditing, which means that the data in the table may change over time. The safest approach to generate accurate reports in the store_sales_summary table is to use Structured Streaming to subscribe to the change data feed for daily_store_sales and apply changes to the aggregates in the store_sales_summary table with each update. Structured Streaming is a scalable and fault-tolerant stream processing engine built on Spark SQL. Structured Streaming allows processing data streams as if they were tables or DataFrames, using familiar operations such as select, filter, groupBy, or join. Structured Streaming also supports output modes that specify how to write the results of a streaming query to a sink, such as append, update, or complete. Structured Streaming can handle both streaming and batch data sources in a unified manner. The change data feed is a feature of Delta Lake that provides structured streaming sources that can subscribe to changes made to a Delta Lake table. The change data feed captures both data changes and schema changes as ordered events that can be processed by downstream applications or services. The change data feed can be configured with different options, such as starting from a specific version or timestamp, filtering by operation type or partition values, or excluding no-op changes. By using Structured Streaming to subscribe to the change data feed for daily_store_sales, one can capture and process any changes made to the total_sales column due to manual data auditing. By applying these changes to the aggregates in the store_sales_summary table with each update, one can ensure that the reports are always consistent and accurate with the latest data. Verified References: [Databricks Certified Data Engineer Professional], under "Spark Core" section;

Databricks Documentation, under "Structured Streaming" section; Databricks Documentation, under "Delta Change Data Feed" section.

QUESTION 5

The data science team has created and logged a production using MLFlow. The model accepts a list of column names and returns a new column of type DOUBLE.

The following code correctly imports the production model, load the customer table containing the customer_id key column into a Dataframe, and defines the feature columns needed for the model.

```
model = mlflow.pyfunc.spark_udf (spark,
model_uri="models:/churn/prod")

df = spark.table("customers")

columns = ["account_age", "time_since_last_seen", "app_rating"]
```

Which code block will output DataFrame with the schema\\'\\' customer_id LONG, predictions DOUBLE\\'\\'?

A. Model, predict (df, columns)

B. Df, map (lambda k:midel (x [columns]) ,select (\\'\\'customer_id predictions\\'\\')

C. Df. Select (\\'\\'customer_id\\'\\'. Model (\\'\\'columns) alias (\\'\\'predictions\\'\\')

D. Df.apply(model, columns). Select (\\'\\'customer_id, prediction\\'\\'

Correct Answer: A

Given the information that the model is registered with MLflow and assuming predict is the method used to apply the model to a set of columns, we use the model.predict() function to apply the model to the DataFrame df using the specified

columns. The model.predict() function is designed to take in a DataFrame and a list of column names as arguments, applying the trained model to these features to produce a predictions column. When working with PySpark, this predictions

column needs to be selected alongside the customer_id to create a new DataFrame with the schema customer_id LONG, predictions DOUBLE.

References:

MLflow documentation on using Python function models:

https://www.mlflow.org/docs/latest/models.html#python-function-python PySpark MLlib documentation on model prediction:

https://spark.apache.org/docs/latest/ml-pipeline.html#pipeline

Latest DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-ENGINEER Dumps
DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-ENGINEER VCE Dumps
DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-ENGINEER Practice Test