DUMPSARENA

Databricks Certified Data Engineer Professional Exam

Databricks Databricks-Certified-Professional-Data-Engineer

Version Demo

Total Demo Questions: 15

Total Premium Questions: 220

Buy Premium PDF

https://dumpsarena.co

sales@dumpsarena.co

sales@dumpsarena.co
dumpsarena.co

QUESTION NO: 1

You would like to build a spark streaming process to read from a Kafka queue and write to a Delta table every 15 minutes, what is the correct trigger option

A. trigger("15 minutes")

B. trigger(process "15 minutes")

C. trigger(processingTime = 15)

D. trigger(processingTime = "15 Minutes")

E. trigger(15)

ANSWER: D

Explanation:

Explanation

The answer is trigger(processingTime = "15 Minutes")

Triggers:

• Unspecified

This is the default. This is equivalent to using processingTime="500ms"

• Fixed interval micro-batches .trigger(processingTime="2 minutes")

The query will be executed in micro-batches and kicked off at the user-specified intervals

• One-time micro-batch .trigger(once=True)

The query will execute a single micro-batch to process all the available data and then stop on its own

• One-time micro-batch.trigger .trigger(availableNow=True) -- New feature a better version of (once=True)

Databricks supports trigger(availableNow=True) in Databricks Runtime 10.2 and above for Delta Lake and Auto Loader sources. This functionality combines the batch processing approach of trigger once with the ability to configure batch size, resulting in multiple parallelized batches that give greater control for right-sizing batches and the resultant files.

QUESTION NO: 2

Your colleague was walking you through how a job was setup, but you noticed a warning message that said, “Jobs running on all-purpose cluster are considered all purpose compute", the colleague was not sure why he was getting the warning message, how do you best explain this warning mes-sage?

A. All-purpose clusters cannot be used for Job clusters, due to performance issues.

B. All-purpose clusters take longer to start the cluster vs a job cluster

C. All-purpose clusters are less expensive than the job clusters

D. All-purpose clusters are more expensive than the job clusters

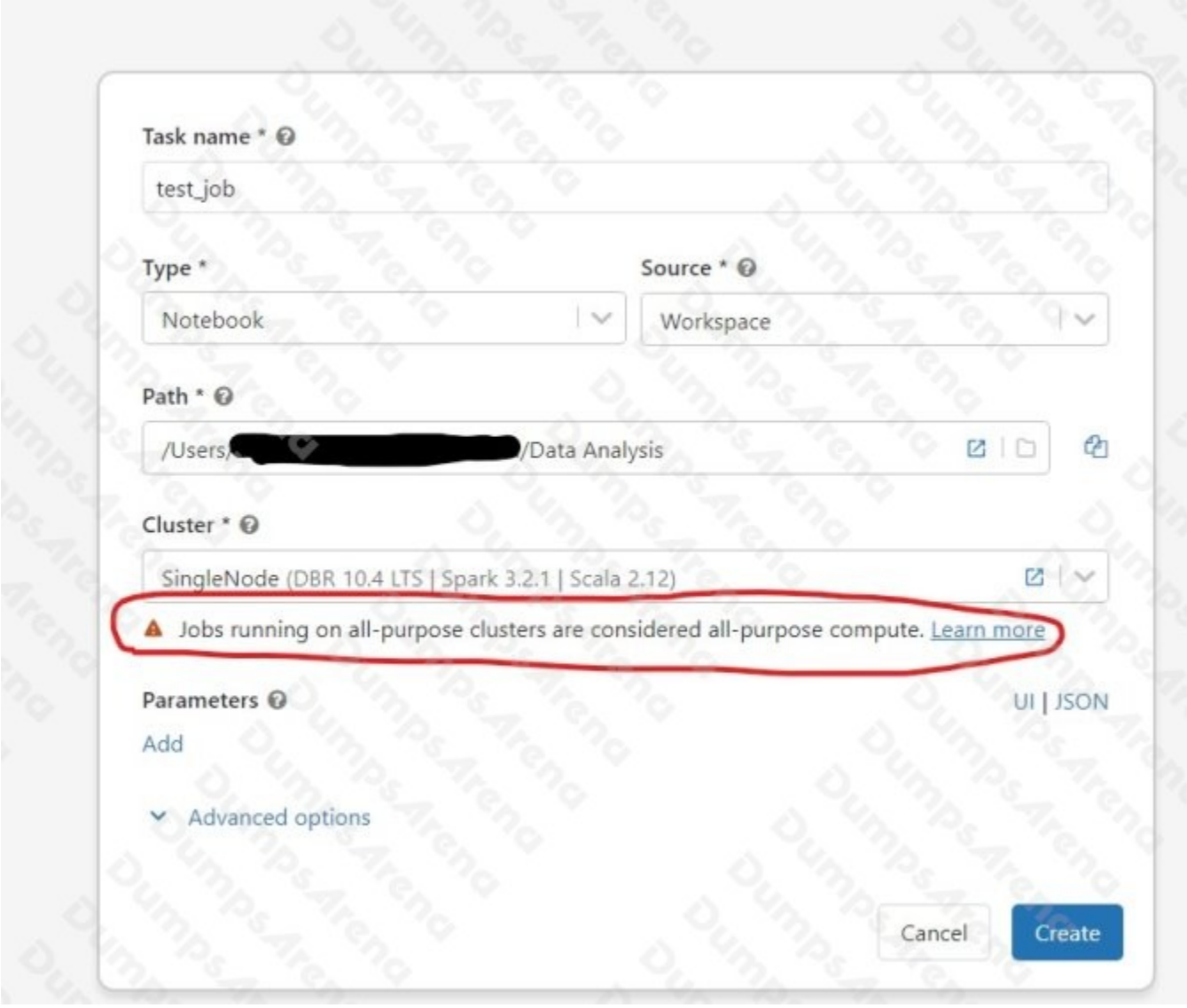
E. All-purpose cluster provide interactive messages that can not be viewed in a job

ANSWER: D

Explanation:

Explanation

Warning message:



Pricing for All-purpose clusters are more expensive than the job clusters

AWS pricing(Aug 15th 2022)

aws	Standard	Premium	Enterprise
	One platform for your data analytics and ML workloads	Data analytics and ML at scale across your business	Data analytics and ML for your mission critical workloads
CLASSIC COMPUTE — Jobs Light Compute Run data engineering pipelines to build data lakes.	$0.07 / DBU	$0.10 / DBU	$0.13 / DBU
Jobs Compute Jobs Compute Photon Run data engineering pipelines to build data lakes and manage data at scale.	$0.10 / DBU	$0.15 / DBU	$0.20 / DBU
Delta Live Tables Delta Live Tables Photon Easily build high quality streaming or batch ETL pipelines using Python or SQL with the DLT Edition that is best for your workload. Learn more	$0.20 - $0.36 / DBU	$0.20 - $0.36 / DBU	$0.20 - $0.36 / DBU
SQL Compute Run SQL queries for BI reporting, analytics and visualization to get timely insights from data lakes.	-	$0.22 / DBU	$0.22 / DBU
All-Purpose Compute All-Purpose Compute Photon Run interactive data science and machine learning workloads. Also good for data engineering, BI and data analytics.	$0.40 / DBU	$0.55 / DBU	$0.65 / DBU

US East (N. Virginia)

Bottom of FormTop of Form

QUESTION NO: 3

The operations team is interested in monitoring the recently launched product, team wants to set up an email alert when the number of units sold increases by more than 10,000 units. They want to monitor this every 5 mins.

Fill in the below blanks to finish the steps we need to take

• Create ___ query that calculates total units sold

• Setup ____ with query on trigger condition Units Sold > 10,000

• Setup ____ to run every 5 mins

• Add destination _____

A. Python, Job, SQL Cluster, email address

B. SQL, Alert, Refresh, email address

C. SQL, Job, SQL Cluster, email address

D. SQL, Job, Refresh, email address

E. Python, Job, Refresh, email address

ANSWER: B

Explanation:

Explanation

The answer is SQL, Alert, Refresh, email address

Here the steps from Databricks documentation,

Create an alert

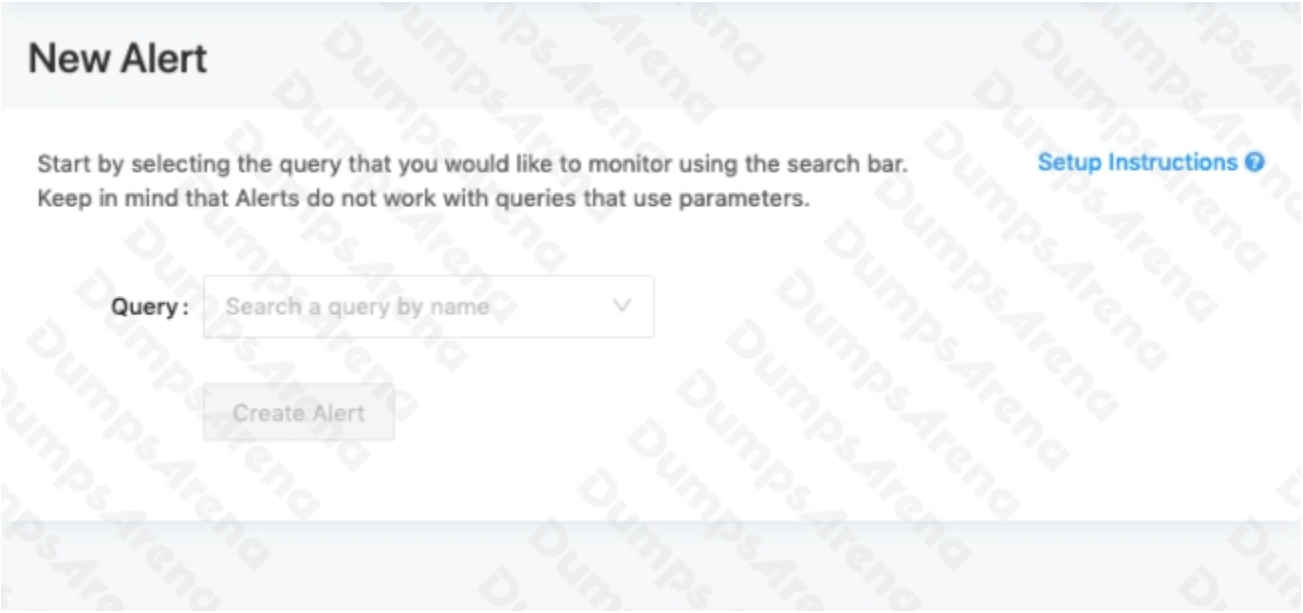
Follow these steps to create an alert on a single column of a query.

1. Do one of the following:

• Click Create in the sidebar and select Alert.

• Click Alerts in the sidebar and click the + New Alert button.

2. Search for a target query.



To alert on multiple columns, you need to modify your query. See Alert on multiple col-umns.

3. In the Trigger when field, configure the alert.

• The Value column drop-down controls which field of your query result is evaluated.

• The Condition drop-down controls the logical operation to be applied.

• The Threshold text input is compared against the Value column using the Condition you specify.

Note

If a target query returns multiple records, Databricks SQL alerts act on the first one. As you change the Value column setting, the current value of that field in the top row is shown beneath it.

4. In the When triggered, send notification field, select how many notifications are sent when your alert is triggered:

• Just once: Send a notification when the alert status changes from OK to TRIGGERED.

• Each time alert is evaluated: Send a notification whenever the alert status is TRIGGERED regardless of its status at the previous evaluation.

• At most every: Send a notification whenever the alert status is TRIGGERED at a spe-cific interval. This choice lets you avoid notification spam for alerts that trigger of-ten.

Regardless of which notification setting you choose, you receive a notification whenever the status goes from OK to TRIGGERED or from TRIGGERED to OK. The schedule settings affect how many notifications you will receive if the status remains TRIGGERED from one execution to the next. For details, see Notification frequency.

5. In the Template drop-down, choose a template:

• Use default template: Alert notification is a message with links to the Alert configuration screen and the Query screen.

• Use custom template: Alert notification includes more specific information about the alert.

a. A box displays, consisting of input fields for subject and body. Any static content is valid, and you can incorporate built-in template variables:

• ALERT_STATUS: The evaluated alert status (string).

• ALERT_CONDITION: The alert condition operator (string).

• ALERT_THRESHOLD: The alert threshold (string or number).

• ALERT_NAME: The alert name (string).

• ALERT_URL: The alert page URL (string).

• QUERY_NAME: The associated query name (string).

• QUERY_URL: The associated query page URL (string).

• QUERY_RESULT_VALUE: The query result value (string or number).

• QUERY_RESULT_ROWS: The query result rows (value array).

• QUERY_RESULT_COLS: The query result columns (string array).

An example subject, for instance, could be: Alert "{{ALERT_NAME}}" changed status to {{ALERT_STATUS}}.

b. Click the Preview toggle button to preview the rendered result.

Important

The preview is useful for verifying that template variables are rendered cor-rectly. It is not an accurate representation of the eventual notification content, as each alert destination can display notifications differently.

c. Click the Save Changes button.

6. In Refresh, set a refresh schedule. An alert's refresh schedule is independent of the query's refresh schedule.

• If the query is a Run as owner query, the query runs using the query owner's cre-dential on the alert's refresh schedule.

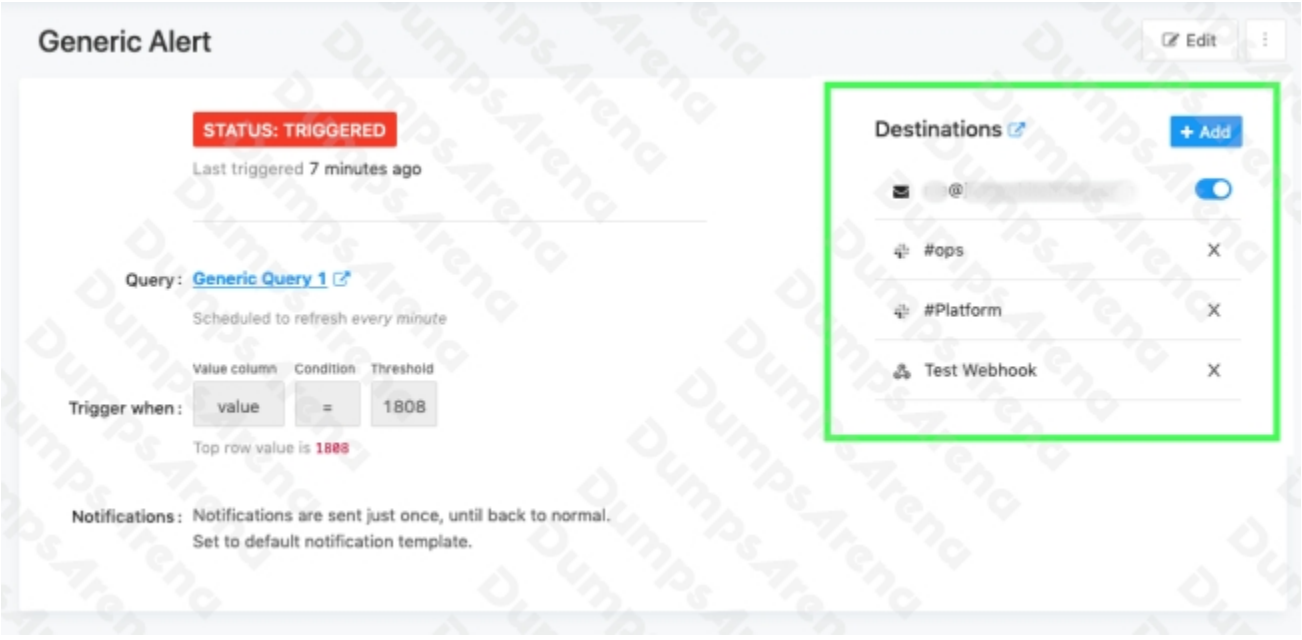
• If the query is a Run as viewer query, the query runs using the alert creator's cre-dential on the alert's refresh schedule.

7. Click Create Alert.

8. Choose an alert destination.

Important

If you skip this step you will not be notified when the alert is triggered.



QUESTION NO: 4

What is the main difference between the below two commands?

1. INSERT OVERWRITE table_name

2. SELECT * FROM table

1. CREATE OR REPLACE TABLE table_name

2. AS SELECT * FROM table

A. INSERT OVERWRITE replaces data by default, CREATE OR REPLACE replaces data and Schema by default

B. INSERT OVERWRITE replaces data and schema by default, CREATE OR REPLACEreplaces data by default

C. INSERT OVERWRITE maintains historical data versions by de-fault, CREATE OR REPLACEclears the historical data versions by default

D. INSERT OVERWRITE clears historical data versions by de-fault, CREATE OR REPLACE maintains the historical data versions by default

E. Both are same and results in identical outcomes

ANSWER: A

Explanation:

The main difference between INSERT OVERWRITE and CREATE OR REPLACE TABLE(CRAS) is that CRAS can modify the schema of the table, i.e it can add new columns or change data types of existing columns. By default INSERT OVERWRITE only overwrites the data.

INSERT OVERWRITE can also be used to overwrite schema, only when spark.databricks.delta.schema.autoMerge.enabled is set true if this option is not enabled and if there is a schema mismatch command will fail.

QUESTION NO: 5

What is the main difference between AUTO LOADER and COPY INTO?

A. COPY INTO supports schema evolution.

B. AUTO LOADER supports schema evolution.

C. COPY INTO supports file notification when performing incremental loads.

D. AUTO LOADER supports reading data from Apache Kafka

E. AUTO LOADER Supports file notification when performing incremental loads.

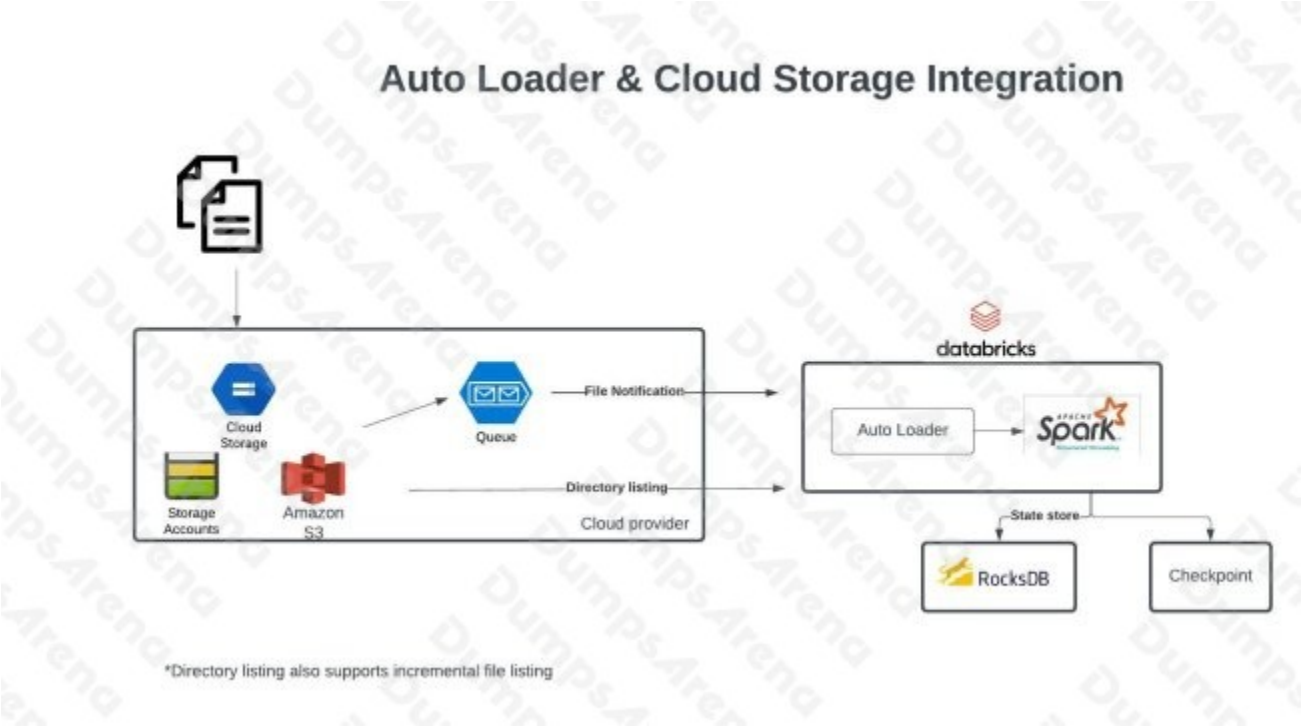
ANSWER: E

Explanation:

Explanation

Auto loader supports both directory listing and file notification but COPY INTO only supports di-rectory listing.

Auto loader file notification will automatically set up a notification service and queue service that subscribe to file events from the input directory in cloud object storage like Azure blob storage or S3. File notification mode is more performant and scalable for large input directories or a high volume of files.



Auto Loader and Cloud Storage Integration

Auto Loader supports a couple of ways to ingest data incrementally

1. Directory listing - List Directory and maintain the state in RocksDB, supports incremental file listing

2. File notification - Uses a trigger+queue to store the file notification which can be later used to retrieve the file, unlike Directory listing File notification can scale up to millions of files per day.

[OPTIONAL]

Auto Loader vs COPY INTO?

Auto Loader

Auto Loader incrementally and efficiently processes new data files as they arrive in cloud storage without any additional setup. Auto Loader provides a new Structured Streaming source called cloudFiles. Given an input directory path on the cloud file storage, the cloudFiles source automatically processes new files as they arrive, with the option of also processing existing files in that directory.

When to use Auto Loader instead of the COPY INTO?

• You want to load data from a file location that contains files in the order of millions or higher. Auto Loader can discover files more efficiently than the COPY INTO SQL command and can split file processing into multiple batches.

• You do not plan to load subsets of previously uploaded files. With Auto Loader, it can be more difficult to reprocess subsets of files. However, you can use the COPY INTO SQL command to reload subsets of files while an Auto Loader stream is simultaneously running.

Auto loader file notification will automatically set up a notification service and queue service that subscribe to file events from the input directory in cloud object storage like Azure blob storage or S3. File notification mode is more performant and scalable for large input directories or a high volume of files.

Here are some additional notes on when to use COPY INTO vs Auto Loader

When to use COPY INTO

https://docs.databricks.com/delta/delta-ingest.html#copy-into-sql-command

When to use Auto Loader

https://docs.databricks.com/delta/delta-ingest.html#auto-loader

QUESTION NO: 6

Which of the following developer operations in CI/CD flow can be implemented in Databricks Re-pos?

A. Delete branch

B. Trigger Databricks CICD pipeline

C. Commit and push code

D. Create a pull request

E. Approve the pull request

ANSWER: C

Explanation:

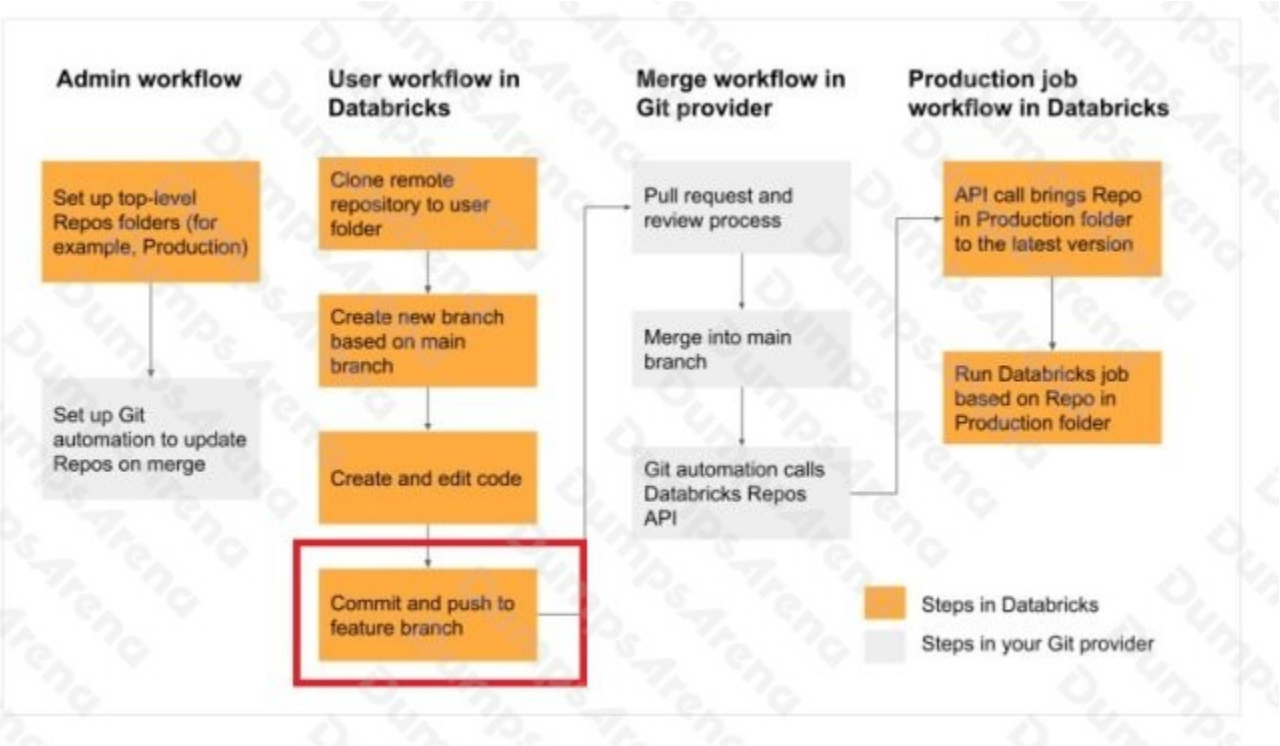
Explanation

The answer is Commit and push code.

See the below diagram to understand the role Databricks Repos and Git provider plays when building a CI/CD workflow.

All the steps highlighted in yellow can be done Databricks Repo, all the steps highlighted in Gray are done in a git provider like Github or Azure Devops.

Exam focus: Please study the below image carefully to understand all of the steps in the CI/CD flow to understand the tasks that are implemented in Databricks Repo vs Git Provider, exam may ask a different type of questions based on this flow.



QUESTION NO: 7

Which of the following Structured Streaming queries is performing a hop from a bronze table to a Silver table?

A. 1. (spark.table("sales").groupBy("store")
2. .agg(sum("sales")).writeStream
3. .option("checkpointLocation",checkpointPath)
4. .outputMode("complete")
5. .table("aggregatedSales"))

B. 1. (spark.table("sales").agg(sum("sales"),sum("units"))
2. .writeStream
3. .option("checkpointLocation",checkpointPath)
4. .outputMode("complete")
5. .table("aggregatedSales"))

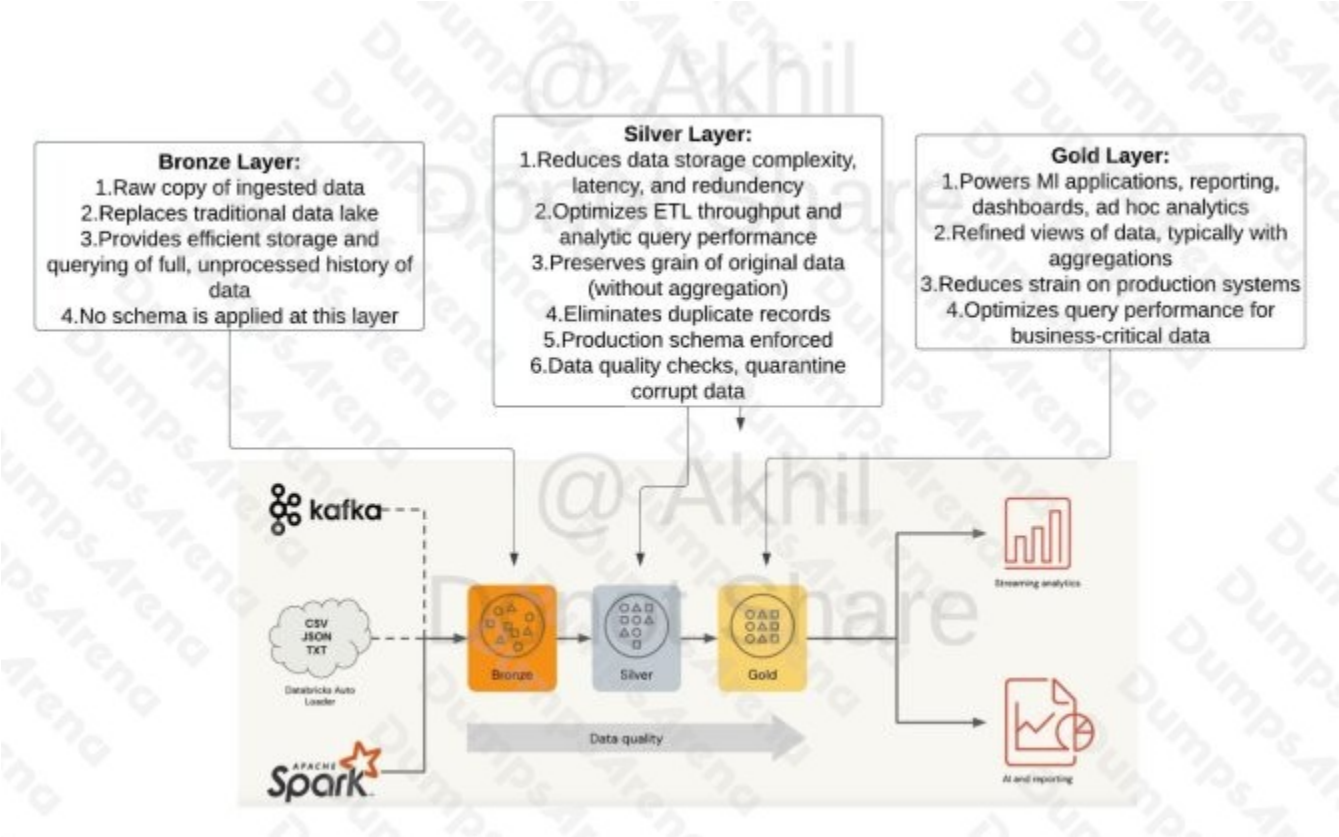
C. 1. (spark.table("sales")
2. .withColumn("avgPrice", col("sales") / col("units"))
3. .writeStream
4. .option("checkpointLocation", checkpointPath)
5. .outputMode("append")
6. .table("cleanedSales"))

D. 1. (spark.readStream.load(rawSalesLocation)
2. .writeStream
3. .option("checkpointLocation", checkpointPath)
4. .outputMode("append")
5. .table("uncleanedSales"))

E. 1. (spark.read.load(rawSalesLocation)
2. .writeStream
3. .option("checkpointLocation", checkpointPath)
4. .outputMode("append")
5. .table("uncleanedSales"))

ANSWER: C

Explanation:



QUESTION NO: 8

identifies if its needs to be converted to Fahrenheit or Celcius with a one-word letter F or C?

select udf_convert(60,'C') will result in 15.5

select udf_convert(10,'F') will result in 50

A. 1. CREATE UDF FUNCTION udf_convert(temp DOUBLE, measure STRING)
2. RETURNS DOUBLE
3. RETURN CASE WHEN measure == 'F' then (temp * 9/5) + 32
4. ELSE (temp – 33) * 5/9
5. END

B. 1. CREATE UDF FUNCTION udf_convert(temp DOUBLE, measure STRING)
2. RETURN CASE WHEN measure == 'F' then (temp * 9/5) + 32
3. ELSE (temp – 33) * 5/9
4. END

C. 1. CREATE FUNCTION udf_convert(temp DOUBLE, measure STRING)
2. RETURN CASE WHEN measure == 'F' then (temp * 9/5) + 32
3. ELSE (temp – 33) * 5/9
4. END

D. 1. CREATE FUNCTION udf_convert(temp DOUBLE, measure STRING)
2. RETURNS DOUBLE
3. RETURN CASE WHEN measure == 'F' then (temp * 9/5) + 32
4. ELSE (temp – 33) * 5/9
5. END

E. 1. CREATE USER FUNCTION udf_convert(temp DOUBLE, measure STRING)
2. RETURNS DOUBLE
3. RETURN CASE WHEN measure == 'F' then (temp * 9/5) + 32
4. ELSE (temp – 33) * 5/9
5. END

ANSWER: D

Explanation:

Explanation

The answer is

1. CREATE FUNCTION udf_convert(temp DOUBLE, measure STRING)

2. RETURNS DOUBLE

3. RETURN CASE WHEN measure == ‘F’ then (temp * 9/5) + 32

4. ELSE (temp – 33) * 5/9

5. END

QUESTION NO: 9

Which of the following table constraints that can be enforced on Delta lake tables are supported?

A. Primary key, foreign key, Not Null, Check Constraints

B. Primary key, Not Null, Check Constraints

C. Default, Not Null, Check Constraints

D. Not Null, Check Constraints

E. Unique, Not Null, Check Constraints

ANSWER: D

Explanation:

Explanation

The answer is Not Null, Check Constraints

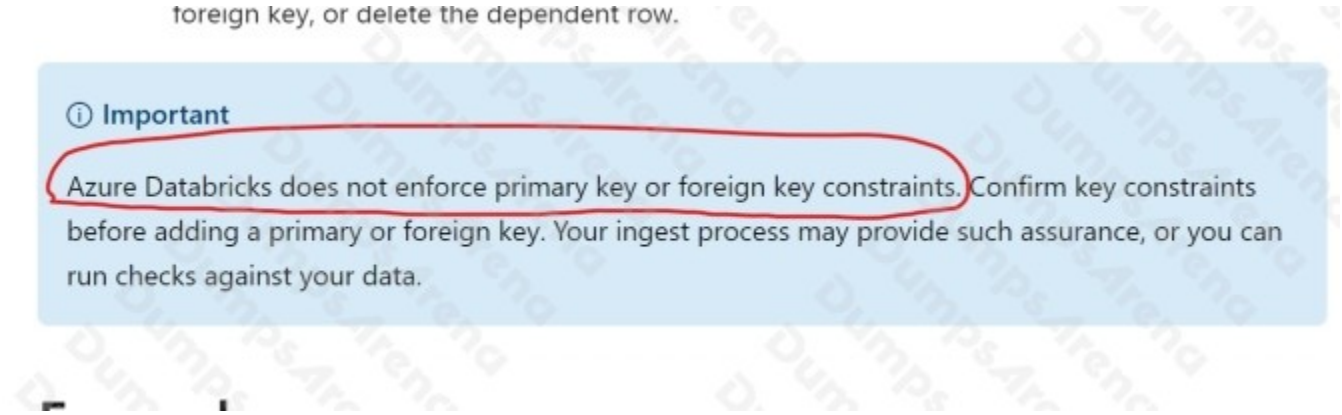
https://docs.microsoft.com/en-us/azure/databricks/delta/delta-constraints

ALTER TABLE events CHANGE COLUMN id SET NOT NULL;

ALTER TABLE events ADD CONSTRAINT dateWithinRange CHECK (date > '1900-01-01');

Note: Databricks as of DBR 11.1 added support for Primary Key and Foreign Key when Unity Catalog is enabled but this is for information purposes only these are not actually enforced. You may ask then why are we defining these if they are not enforced, so especially these information constraints are very helpful if you have a BI tool that can benefit from knowing the relationship between the tables, so it will be easy when creating reports/dashboards or understanding the data model when using any Data modeling tool.

Primary and Foreign Key



QUESTION NO: 10

What is the main difference between the silver layer and gold layer in medallion architecture?

A. Silver optimized to perform ETL, Gold is optimized query performance

B. Gold is optimized go perform ETL, Silver is optimized for query performance

C. Silver is copy of Bronze, Gold is a copy of Silver

D. Silver is stored in Delta Lake, Gold is stored in memory

E. Silver may contain aggregated data, gold may preserve the granularity of original data

ANSWER: A

Explanation:

Explanation

Medallion Architecture – Databricks

Gold Layer:

1. Powers MI applications, reporting, dashboards, ad hoc analytics

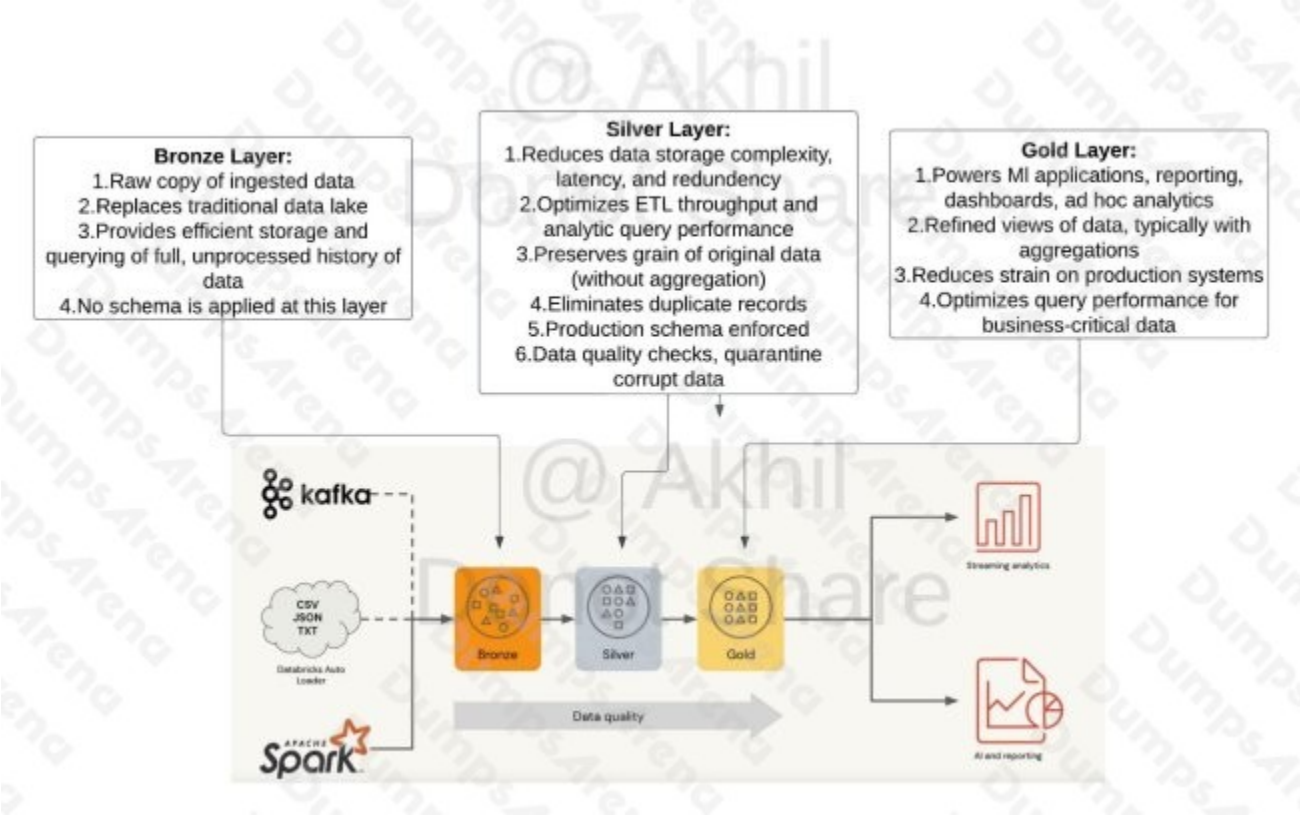
2. Refined views of data, typically with aggregations

3. Reduces strain on production systems

4. Optimizes query performance for business-critical data

Exam focus: Please review the below image and understand the role of each layer(bronze, silver, gold) in medallion architecture, you will see varying questions targeting each layer and its purpose.

Sorry I had to add the watermark some people in Udemy are copying my content.



QUESTION NO: 11

Where are Interactive notebook results stored in Databricks product architecture?

A. Data plane

B. Control plane

C. Data and Control plane

D. JDBC data source

E. Databricks web application

ANSWER: C

Explanation:

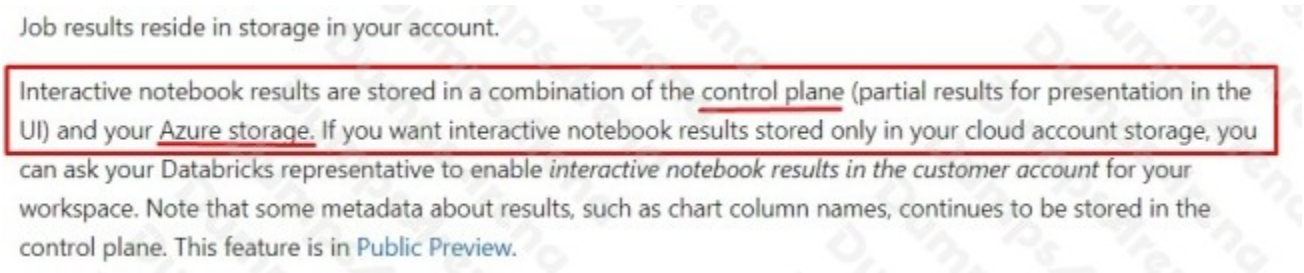
Explanation

The answer is Data and Control plane,

Only Job results are stored in Data Plane(your storage), Interactive notebook results are stored in a combination of the control plane (partial results for presentation in the UI) and customer storage.

https://docs.microsoft.com/en-us/azure/databricks/getting-started/overview#--high-level-architecture

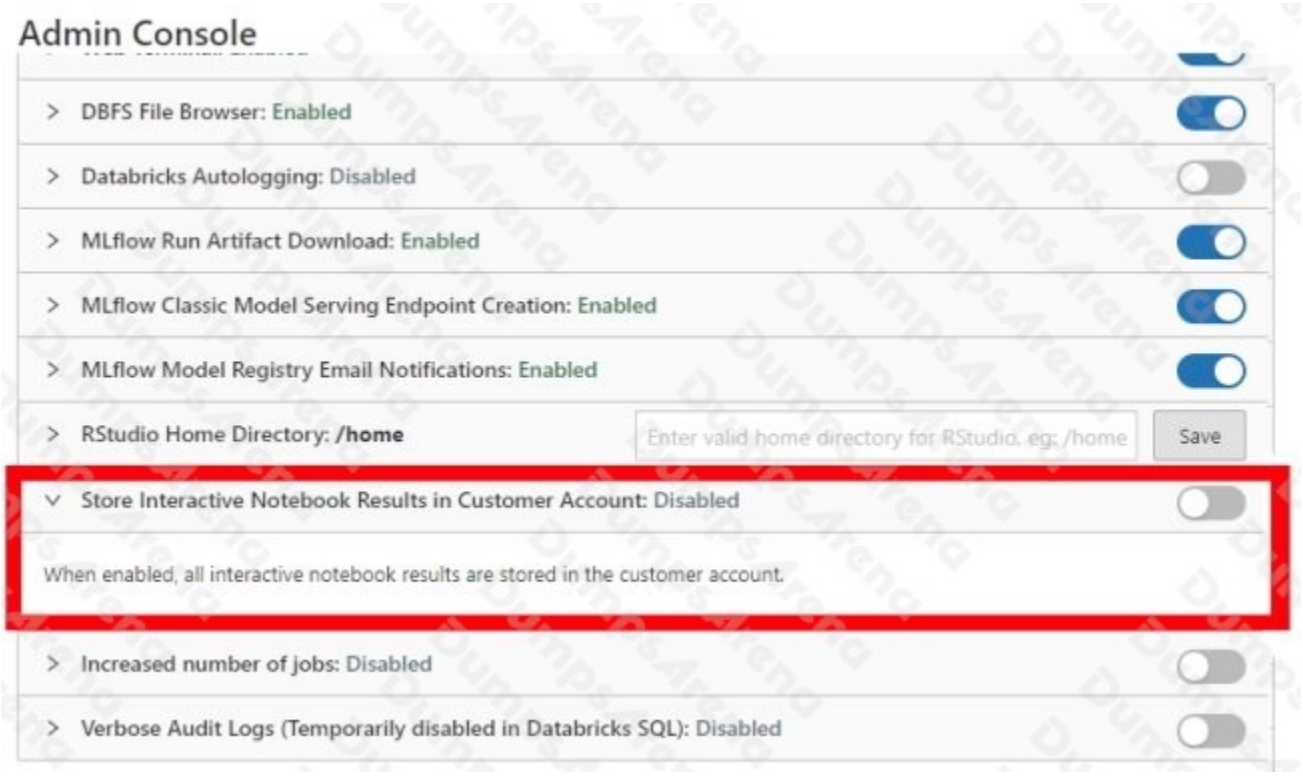
Snippet from the above documentation,


Job results reside in storage in your account.

Interactive notebook results are stored in a combination of the control plane (partial results for presentation in the UI) and your Azure storage. If you want interactive notebook results stored only in your cloud account storage, you can ask your Databricks representative to enable *interactive notebook results in the customer account* for your workspace. Note that some metadata about results, such as chart column names, continues to be stored in the control plane. This feature is in Public Preview.

How to change this behavior?

You can change this behavior using Workspace/Admin Console settings for that workspace, once enabled all of the interactive results are stored in the customer account(data plane) except the new notebook visualization feature Databricks has recently introduced, this still stores some metadata in the control pane irrespective of the below settings. please refer to the documentation for more details.



Why is this important to know?

I recently worked on a project where we had to deal with sensitive information of customers and we had a security requirement that all of the data need to be stored in the data plane including notebook results.

QUESTION NO: 12

Which of the following programming languages can be used to build a Databricks SQL dashboard?

A. Python

B. Scala

C. SQL

D. R

E. All of the above

ANSWER: C

QUESTION NO: 13

Your team member is trying to set up a delta pipeline and build a second gold table to the same pipeline with aggregated metrics based on an existing Delta Live table called sales_orders_cleaned but he is facing a problem in starting the pipeline, the pipeline is failing to state it cannot find the table sales_orders_cleaned, you are asked to identify and fix the problem.

1. CREATE LIVE TABLE sales_order_in_chicago

2. AS

3. SELECT order_date, city, sum(price) as sales,

4. FROM sales_orders_cleaned

5. WHERE city = 'Chicago')

6. GROUP BY order_date, city

A. Use STREAMING LIVE instead of LIVE table

B. Delta live table can be used in a group by clause

C. Delta live tables pipeline can only have one table

D. Sales_orders_cleaned table is missing schema name LIVE

E. The pipeline needs to be deployed so the first table is created before we add a second table

ANSWER: D

Explanation:

Explanation

The answer is, Sales_orders_cleaned table is missing schema name LIVE

Every Delta live table should have schema LIVE

Here is the correct syntax,

1. CREATE LIVE TABLE sales_order_in_chicago

2. AS

3. SELECT order_date, city, sum(price) as sales,

4. FROM LIVE.sales_orders_cleaned

5. WHERE city = 'Chicago')

6. GROUP BY order_date, city

QUESTION NO: 14

Newly joined data analyst requested read-only access to tables, assuming you are owner/admin which section of Databricks platform is going to facilitate granting select access to the user

A. Admin console

B. User settings

C. Data explorer

D. Azure Databricks control pane IAM

E. Azure RBAC

ANSWER: C

Explanation:

Explanation

Anser is Data Explorer

https://docs.databricks.com/sql/user/data/index.html

Data explorer lets you easily explore and manage permissions on databases and tables. Users can view schema details, preview sample data, and see table details and properties. Administrators can view and change owners, and admins and data object owners can grant and revoke permissions.

To open data explorer, click Data in the sidebar.

QUESTION NO: 15

What could be the expected output of query SELECT COUNT (DISTINCT *) FROM user on this table

userId	username	email
1	john.smith	john.smith@com
2	NULL	david.clear@com
3	kevin.smith	kevin.smith@com

A. 3

B. 2
(Correct)

C. 1

D. 0

E. NULL

ANSWER: B

Explanation:

The answer is 2,

Count(DISTINCT *) removes rows with any column with a NULL value