Custom View Settings

## Topic 1 - Exam A

Question #1 *Topic 1*

Which of the following layers of the medallion architecture is most commonly used by data analysts?

A. None of these layers are used by data analysts
B. Gold
C. All of these layers are used equally by data analysts
D. Silver
E. Bronze

Question #2 *Topic 1*

A data analyst has recently joined a new team that uses Databricks SQL, but the analyst has never used Databricks before. The analyst wants to know where in Databricks SQL they can write and execute SQL queries.
On which of the following pages can the analyst write and execute SQL queries?

A. Data page
B. Dashboards page
C. Queries page
D. Alerts page
E. SQL Editor page

Question #3 *Topic 1*

Which of the following describes how Databricks SQL should be used in relation to other business intelligence (BI) tools like Tableau, Power BI, and looker?

A. As an exact substitute with the same level of functionality
B. As a substitute with less functionality
C. As a complete replacement with additional functionality
D. As a complementary tool for professional-grade presentations
E. As a complementary tool for quick in-platform BI work

Question #4 *Topic 1*

Which of the following approaches can be used to connect Databricks to Fivetran for data ingestion?

A. Use Workflows to establish a SQL warehouse (formerly known as a SQL endpoint) for Fivetran to interact with

B. Use Delta Live Tables to establish a cluster for Fivetran to interact with

C. Use Partner Connect's automated workflow to establish a cluster for Fivetran to interact with

D. Use Partner Connect's automated workflow to establish a SQL warehouse (formerly known as a SQL endpoint) for Fivetran to interact with

E. Use Workflows to establish a cluster for Fivetran to interact with

Question #5 *Topic 1*

Data professionals with varying titles use the Databricks SQL service as the primary touchpoint with the Databricks Lakehouse Platform. However, some users will use other services like Databricks Machine Learning or Databricks Data Science and Engineering.
Which of the following roles uses Databricks SQL as a secondary service while primarily using one of the other services?

A. Business analyst

B. SQL analyst

C. Data engineer

D. Business intelligence analyst

E. Data analyst

Question #6 *Topic 1*

A data analyst has set up a SQL query to run every four hours on a SQL endpoint, but the SQL endpoint is taking too long to start up with each run.
Which of the following changes can the data analyst make to reduce the start-up time for the endpoint while managing costs?

A. Reduce the SQL endpoint cluster size

B. Increase the SQL endpoint cluster size

C. Turn off the Auto stop feature

D. Increase the minimum scaling value

E. Use a Serverless SQL endpoint

Question #7 *Topic 1*

A data engineering team has created a Structured Streaming pipeline that processes data in micro-batches and populates gold-level tables. The microbatches are triggered every minute.
A data analyst has created a dashboard based on this gold-level data. The project stakeholders want to see the results in the dashboard updated within one minute or less of new data becoming available within the gold-level tables.
Which of the following cautions should the data analyst share prior to setting up the dashboard to complete this task?

A. The required compute resources could be costly

B. The gold-level tables are not appropriately clean for business reporting

C. The streaming data is not an appropriate data source for a dashboard

D. The streaming cluster is not fault tolerant

E. The dashboard cannot be refreshed that quickly

Question #8 *Topic 1*

Which of the following approaches can be used to ingest data directly from cloud-based object storage?

A. Create an external table while specifying the DBFS storage path to FROM

B. Create an external table while specifying the DBFS storage path to PATH

C. It is not possible to directly ingest data from cloud-based object storage

D. Create an external table while specifying the object storage path to FROM

E. Create an external table while specifying the object storage path to LOCATION

Question #9 *Topic 1*

A data analyst wants to create a dashboard with three main sections: Development, Testing, and Production. They want all three sections on the same dashboard, but they want to clearly designate the sections using text on the dashboard.
Which of the following tools can the data analyst use to designate the Development, Testing, and Production sections using text?

A. Separate endpoints for each section

B. Separate queries for each section

C. Markdown-based text boxes

D. Direct text written into the dashboard in editing mode

E. Separate color palettes for each section

Question #10 *Topic 1*

A data analyst needs to use the Databricks Lakehouse Platform to quickly create SQL queries and data visualizations. It is a requirement that the compute resources in the platform can be made serverless, and it is expected that data visualizations can be placed within a dashboard.
Which of the following Databricks Lakehouse Platform services/capabilities meets all of these requirements?

A. Delta Lake

B. Databricks Notebooks

C. Tableau

D. Databricks Machine Learning

E. Databricks SQL

Question #11 *Topic 1*

A data analyst is attempting to drop a table my_table. The analyst wants to delete all table metadata and data.
They run the following command:
DROP TABLE IF EXISTS my_table;
While the object no longer appears when they run SHOW TABLES, the data files still exist.
Which of the following describes why the data files still exist and the metadata files were deleted?

A. The table's data was larger than 10 GB

B. The table did not have a location

C. The table was external

D. The table's data was smaller than 10 GB

E. The table was managed

Question #12 *Topic 1*

After running DESCRIBE EXTENDED accounts.customers;, the following was returned:

```
Name        accounts.customers
Location    dbfs:/stakeholders/customers
Provider    delta
Owner       root
Type        EXTERNAL
```

Now, a data analyst runs the following command:
DROP accounts.customers;
Which of the following describes the result of running this command?

A. Running SELECT * FROM delta. `dbfs:/stakeholders/customers` results in an error.

B. Running SELECT * FROM accounts.customers will return all rows in the table.

C. All files with the .customers extension are deleted.

D. The accounts.customers table is removed from the metastore, and the underlying data files are deleted.

E. The accounts.customers table is removed from the metastore, but the underlying data files are untouched.

Question #13 *Topic 1*

Which of the following should data analysts consider when working with personally identifiable information (PII) data?

A. Organization-specific best practices for PII data

B. Legal requirements for the area in which the data was collected

C. None of these considerations

D. Legal requirements for the area in which the analysis is being performed

E. All of these considerations

Question #14 *Topic 1*

Delta Lake stores table data as a series of data files, but it also stores a lot of other information.
Which of the following is stored alongside data files when using Delta Lake?

A. None of these

B. Table metadata, data summary visualizations, and owner account information

C. Table metadata

D. Data summary visualizations

E. Owner account information

Question #15 *Topic 1*

Which of the following is an advantage of using a Delta Lake-based data lakehouse over common data lake solutions?

A. ACID transactions

B. Flexible schemas

C. Data deletion

D. Scalable storage

E. Open-source formats

Question #16 *Topic 1*

Which of the following benefits of using Databricks SQL is provided by Data Explorer?

A. It can be used to run UPDATE queries to update any tables in a database.

B. It can be used to view metadata and data, as well as view/change permissions.

C. It can be used to produce dashboards that allow data exploration.

D. It can be used to make visualizations that can be shared with stakeholders.

E. It can be used to connect to third party BI cools.

Question #17 *Topic 1*

The stakeholders.customers table has 15 columns and 3,000 rows of data. The following command is run:

```
CREATE TEMP VIEW stakeholders.eur_customers AS
    SELECT * FROM stakeholders.customers
    WHERE continent = 'eur';
```

After running SELECT * FROM stakeholders.eur_customers, 15 rows are returned. After the command executes completely, the user logs out of Databricks.
After logging back in two days later, what is the status of the stakeholders.eur_customers view?

A. The view remains available and SELECT * FROM stakeholders.eur_customers will execute correctly.

B. The view has been dropped.

C. The view is not available in the metastore, but the underlying data can be accessed with SELECT * FROM delta.`stakeholders.eur_customers`.

D. The view remains available but attempting to SELECT from it results in an empty result set because data in views are automatically deleted after logging out.

E. The view has been converted into a table.

Question #18 *Topic 1*

A data analyst created and is the owner of the managed table my_ table. They now want to change ownership of the table to a single other user using Data Explorer.
Which of the following approaches can the analyst use to complete the task?

A. Edit the Owner field in the table page by removing their own account

B. Edit the Owner field in the table page by selecting All Users

C. Edit the Owner field in the table page by selecting the new owner's account

D. Edit the Owner field in the table page by selecting the Admins group

E. Edit the Owner field in the table page by removing all access

Question #19 *Topic 1*

A data analyst has a managed table table_name in database database_name. They would now like to remove the table from the database and all of the data files associated with the table. The rest of the tables in the database must continue to exist.
Which of the following commands can the analyst use to complete the task without producing an error?

A. DROP DATABASE database_name;

B. DROP TABLE database_name.table_name;

C. DELETE TABLE database_name.table_name;

D. DELETE TABLE table_name FROM database_name;

E. DROP TABLE table_name FROM database_name;

Question #20 Topic 1

A data analyst runs the following command:

SELECT age, country -

FROM my_table -
WHERE age >= 75 AND country = 'canada';
Which of the following tables represents the output of the above command?

A.

| age | country |
|---|---|
| 80 | canada |
| *NULL* | canada |
| 90 | *NULL* |

B.

| age | country |
|---|---|
| 80 | *NULL* |
| 75 | *NULL* |
| 90 | *NULL* |

C.

| id | age | country |
|---|---|---|
| 900 | 80 | canada |
| 901 | 75 | canada |
| 902 | 90 | canada |

D.

| age | country |
|---|---|
| 80 | canada |
| 14 | canada |
| 90 | canada |

E.

| age | country |
|---|---|
| 80 | canada |
| 75 | canada |
| 90 | canada |

Question #21 Topic 1

A data analyst runs the following command:
INSERT INTO stakeholders.suppliers TABLE stakeholders.new_suppliers;
What is the result of running this command?

A. The suppliers table now contains both the data it had before the command was run and the data from the new_suppliers table, and any duplicate data is deleted.

B. The command fails because it is written incorrectly.

C. The suppliers table now contains both the data it had before the command was run and the data from the new_suppliers table, including any duplicate data.

D. The suppliers table now contains the data from the new_suppliers table, and the new_suppliers table now contains the data from the suppliers table.

E. The suppliers table now contains only the data from the new_suppliers table.

Question #22 *Topic 1*

A data engineer is working with a nested array column products in table transactions. They want to expand the table so each unique item in products for each row has its own row where the transaction_id column is duplicated as necessary.
They are using the following incomplete command:

```
SELECT
    transaction_id,
    _______ AS product
FROM transactions;
```

Which of the following lines of code can they use to fill in the blank in the above code block so that it successfully completes the task?

A. array distinct(products)

B. explode(products)

C. reduce(products)

D. array(products)

E. flatten(products)

Question #23 *Topic 1*

A data analysis team is working with the table_bronze SQL table as a source for one of its most complex projects. A stakeholder of the project notices that some of the downstream data is duplicative. The analysis team identifies table_bronze as the source of the duplication.
Which of the following queries can be used to deduplicate the data from table_bronze and write it to a new table table_silver?

A. CREATE TABLE table_silver AS -
SELECT DISTINCT *
FROM table_bronze;

B. CREATE TABLE table_silver AS -
INSERT *
FROM table_bronze;

C. CREATE TABLE table_silver AS -
MERGE DEDUPLICATE *
FROM table_bronze;

D. INSERT INTO TABLE table_silver -
SELECT * FROM table_bronze;

E. INSERT OVERWRITE TABLE table_silver
SELECT * FROM table_bronze;

A business analyst has been asked to create a data entity/object called sales_by_employee. It should always stay up-to-date when new data are added to the sales table. The new entity should have the columns sales_person, which will be the name of the employee from the employees table, and sales, which will be all sales for that particular sales person. Both the sales table and the employees table have an employee_id column that is used to identify the sales person.

Which of the following code blocks will accomplish this task?

A.
```
CREATE TEMPORARY TABLE sales_by_employee AS
    SELECT employees.employee_name sales_person,
           sales.sales
    FROM sales
    JOIN employees
    ON employees.employee_id = sales.employee_id;
```

B.
```
CREATE OR REPLACE VIEW sales_by_employee USING
    SELECT employees.employee_name sales_person,
           sales.sales
    FROM sales
    JOIN employees
    ON employees.employee_id = sales.employee_id;
```

C.
```
SELECT employees.employee_name sales_person,
       sales.sales
    FROM sales
    JOIN employees
    ON employees.employee_id = sales.employee_id USING
    CREATE OR REPLACE VIEW sales_by_employee;
```

D.
```
CREATE OR REPLACE VIEW sales_by_employee AS
    SELECT employees.employee_name sales_person,
           sales.sales FROM sales
    JOIN employees
    ON employees.employee_id = sales.employee_id;
```

E.
```
CREATE OR REPLACE TABLE sales_by_employee AS
   SELECT employees.employee_name sales_person,
          sales.sales
    FROM sales
    JOIN employees
    ON employees.employee_id = sales.employee_id;
```

A data analyst has been asked to use the below table sales_table to get the percentage rank of products within region by the sales:

| region | product | sales |
|---|---|---|
| WEST | A | 1880.59 |
| EAST | A | 2045.99 |
| EAST | B | 4583.23 |
| WEST | B | 3391.19 |

The result of the query should look like this:

| region | product | sales |
|---|---|---|
| EAST | B | 0 |
| EAST | A | 1 |
| WEST | B | 0 |
| WEST | A | 1 |

Which of the following queries will accomplish this task?

A.
```
SELECT
    region,
    product,
    RANK() OVER (
        PARTITION BY region
        ORDER BY sales DESC
    ) AS rank
FROM sales_table;
GROUP BY region, product;
```

B.
```
SELECT
    region,
    product,
    PERCENT_RANK () OVER (
        PARTITION BY region
        ORDER BY sales DESC
    ) AS rank
FROM sales_table;
GROUP BY region, product;
```

C.
```
SELECT
    region,
    product,
    PERCENT_RANK () OVER (
        ORDER BY sales DESC
    ) AS rank
FROM sales_table;
```

D.
```
SELECT
    region,
    product,
    PERCENT RANK () OVER (
        PARTITION BY product
        ORDER BY sales DESC
     ) AS rank
FROM sales_table;
GROUP BY region, product;
```

E.
```
SELECT
    region,
    product,
    RANK() OVER (
        PARTITION BY product
    ) AS rank
FROM sales_table;
```

Question #26 *Topic 1*

In which of the following situations should a data analyst use higher-order functions?

A. When custom logic needs to be applied to simple, unnested data

B. When custom logic needs to be converted to Python-native code

C. When custom logic needs to be applied at scale to array data objects

D. When built-in functions are taking too long to perform tasks

E. When built-in functions need to run through the Catalyst Optimizer

Question #27 *Topic 1*

Consider the following two statements:

Statement 1:

```
SELECT *
    FROM customers
    LEFT SEMI JOIN orders
    ON customers.customer_id = orders.customer_id;
```

Statement 2:

```
SELECT *
    FROM customers
    LEFT ANTI JOIN orders
    ON customers.customer_id = orders.customer_id;
```

Which of the following describes how the result sets will differ for each statement when they are run in Databricks SQL?

A. The first statement will return all data from the customers table and matching data from the orders table. The second statement will return all data from the orders table and matching data from the customers table. Any missing data will be filled in with NULL.

B. When the first statement is run, only rows from the customers table that have at least one match with the orders table on customer_id will be returned. When the second statement is run, only those rows in the customers table that do not have at least one match with the orders table on customer_id will be returned.

C. There is no difference between the result sets for both statements.

D. Both statements will fail because Databricks SQL does not support those join types.

E. When the first statement is run, all rows from the customers table will be returned and only the customer_id from the orders table will be returned. When the second statement is run, only those rows in the customers table that do not have at least one match with the orders table on customer_id will be returned.

Question #28 *Topic 1*

A data analyst has created a user-defined function using the following line of code:
CREATE FUNCTION price(spend DOUBLE, units DOUBLE)

RETURNS DOUBLE -
RETURN spend / units;
Which of the following code blocks can be used to apply this function to the customer_spend and customer_units columns of the table customer_summary to create column customer_price?

A. SELECT PRICE customer_spend, customer_units AS customer_price
FROM customer_summary

B. SELECT price -
FROM customer_summary

C. SELECT function(price(customer_spend, customer_units)) AS customer_price
FROM customer_summary

D. SELECT double(price(customer_spend, customer_units)) AS customer_price
FROM customer_summary

E. SELECT price(customer_spend, customer_units) AS customer_price
FROM customer_summary

Question #29 *Topic 1*

A data analyst has been asked to count the number of customers in each region and has written the following query:

```
SELECT region, count(*) AS number_of_customers
    FROM customers
    ORDER BY region;
```

If there is a mistake in the query, which of the following describes the mistake?

A. The query is using count(*), which will count all the customers in the customers table, no matter the region.

B. The query is missing a GROUP BY region clause.

C. The query is using ORDER BY, which is not allowed in an aggregation.

D. There are no mistakes in the query.

E. The query is selecting region, but region should only occur in the ORDER BY clause.