Questions:

Question 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 2

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 3

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 4

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 5

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 6

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 7

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 8

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;

Question 9

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;
```

Question 10

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 11

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 12

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 13

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 14

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.

Question 15

A data analyst is processing a complex aggregation on a table with zero null values and their query returns the following result:

group_1	group_2	sum
null	null	100
null	Y	70
null	Z	30
A	null	50
A	Y	30
A	Z	20
B	null	50
B	Y	40
B	Z	10

Which of the following queries did the analyst run to obtain the above result?

- A.

```
SELECT
    group_1,
    group_2,
    count(values) AS count
FROM my_table
GROUP BY group_1, group_2 INCLUDING NULL;
```

- B.

```
SELECT
    group_1,
    group_2,
    count(values) AS count
FROM my_table
GROUP BY group_1, group_2 WITH ROLLUP;
```

- **C.**

```
SELECT
    group_1,
    group_2,
    count(values) AS count
FROM my_table
GROUP BY group_1, group 2;
```

- **D.**

```
SELECT
    group_1,
    group_2,
    count(values) AS count
FROM my_table
GROUP BY group_1, group_2, (group_1, group_2);
```

- **E.**

```
SELECT
    group_1,
    group_2,
    count(values) AS count
FROM my_table
GROUP BY group_1, group_2 WITH CUBE;
```

Question 16:

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 17:

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 18:

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 19:

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 20:

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 21:

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 22:

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 23:

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 24:

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 25:

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 26:

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 27:

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 28:

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 29:

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 30:

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 31:

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 32:

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 33:

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 34:

Which of the following statements about Databricks SQL is true? Select one response.

A. With Databricks SQL, queries deliver up to 2x better price/performance than other cloud data warehouses.

B. Delta Live Tables can be created in Databricks SQL.

C. Databricks SQL automatically configures scaling when creating SQL warehouses.

D. Databricks SQL clusters are powered by Photon

Question 35:

Which of the following features is used by Databricks SQL to ensure your data is secure? Select one response.

A. Built-in data governance

B. Delta sharing

C. Integration with 3rd party tools

D. Automatically scalable cloud infrastructure

Question 36:

What is the primary purpose of Databricks SQL?

A. To provide better price/performance and simplify discovery for BI tools.

B. To manage administration and governance of data warehouses.

C. To support a broad set of BI tools, including Tableau and Power BI.

D. All of the above.

Question 37:

Which of the following statements about the lakehouse medallion architecture is true? Select one response.

A. The data in a single upstream table could be used to generate multiple downstream tables.
B. The silver layer is for reporting and uses more de-normalized and read-optimized data models with fewer joins.
C. The gold layer provides a broad view of all key business entities, concepts and transactions.
D. Only minimal or "just-enough" transformations and data cleansing rules are applied to each layer in the medallion architecture.

Question 38:

What is the primary purpose of the bronze layer in the "bronze-silver-gold medallion" paradigm in Delta Lake?
A. To store data in a format suitable for individual business projects or reports.
B. To perform data cleansing, joining, and enrichment on raw data.
C. To provide a "single source of truth" for the enterprise across various projects.
D. To ingest raw data quickly, keeping it in its original format for both current and future projects.

Question 39:

Which of the following statements describes the relationship between the silver and gold layer of data? Select one response.
A. The gold layer has less clean data than the silver layer.
B. Project-specific business rules are applied from the silver to gold layer.
C. Self-service analytics are enabled for the gold layer for ad-hoc reporting in the silver layer.
D. The gold layer is where we land all the data from external source systems, which are represented by the silver layer.

Question 40:

Which of the following statements describes the purpose of Databricks SQL warehouses? Select one response.
A. SQL warehouses enable data analysts to find and share dashboards.
B. SQL warehouses are a declarative framework for building data processing pipelines.
C. SQL warehouses provide data discovery capabilities across Databricks workspaces.
D. SQL warehouses allow users to run SQL commands on data objects within Databricks SQL

Question 41:

What are the benefits of Delta Lake within the Lakehouse Architecture?

A. Real-time data processing with low latency
B. Exclusive support for batch processing
C. ACID transactions, metadata scalability, and storage improvement
D. Data isolation for multiple software development environments

Question 42:

Which of the following statements about SQL warehouse sizing and scaling is true? Select two responses.
A. Increasing maximum scaling allows for multiple users to use the same warehouse at the same time.
B. Scaling is set to a minimum of 1 and a maximum of 1 by default.
C. The higher the cluster size, the higher the latency in your queries.
D. The auto-stop feature will restart the warehouse if it remains idle during the auto-stop period.

Question 43:

Which feature of the platform provides users with the ability to quickly connect to third-party tools with simple to implement integrations? Select one response.

A. SQL Editor
B. Partner Connect
C. Workflows
D. Features

Question 44:

What types of customized visualizations can be created using Databricks SQL?
A. Pie charts, bar graphs, and line charts
B. Counter and funnels
C. Bar charts, combo charts, and geographical maps
D. All of the above

Question 45:

How can you enable aggregation in a Databricks SQL visualization?

A. Modify the underlying SQL query to add an aggregation column.
B. Select the aggregation type directly in the visualization editor.
C. Use the Aggregation drop-down menu in the Visualization Type options.
D. Aggregation is not supported in Databricks SQL visualizations.

Question 46:

How can you add a query visualization to a Databricks dashboard?

A. Select a query and choose the visualization type.
B. Drag and drop a visualization from the sidebar.
C. Copy and paste the visualization code into the dashboard.
D. Query visualizations cannot be added to a Databricks dashboard.

Question 47:

What is the benefit of setting a refresh schedule for a Databricks dashboard?

A. To change the color palette of visualizations.
B. To organize and label workspace objects.
C. To keep the data underlying visualizations up-to-date.
D. To create query parameters for customization.

Question 48:

Which of the following can be added to a query so that the code can be rerun with different variable inputs? Select one response.

A. User-defined functions
B. Parameters
C. Vectors
D. SQL warehouses

Question 49:

A data analyst needs to create a visualization out of the following query:

SELECT order_date FROM sales WHERE order_date >= to_date('2020-01-01') AND order_date <= to_date('2021-01-01');

Which of the following visualization types is best suited to depict the results of this query? Select one response.
A. Funnel
B. Stacked bar chart
C. Bar chart
D. Boxplot

Question 50:

A team of stakeholders needs to be notified of changes in a dashboard's statistics on a daily basis. Which of the following actions can be taken to ensure they always have the newest information? Select one response.

A. A refresh schedule can be configured and stakeholders can be subscribed to the dashboard's output.
B. A trigger alert can be created for the dashboard and stakeholders can be added to the alert notification list.
C. A webhook can be created and shared with stakeholders.

D. None of the above

Question 51:

Which of the following data visualizations displays a single number by default? Select one response.

A. Bar chart
B. Counter
C. Map - markers
D. Funnel

Question 52:

Which of the following automations are available in Databricks SQL? Select one response.

A. Query refresh schedules
B. Dashboard refresh schedules
C. Alerts
D. All of the above

Question 53:

What is the purpose of Alerts in Databricks SQL?

A. To automatically execute SQL queries.
B. To organize queries within a folder structure.
C. To trigger notifications based on specific conditions in scheduled queries.
D. To share dashboards with other team members.

Question 54:

What is the purpose of configuring a refresh schedule for a query in Databricks SQL?.

A. To automatically pull new data into a table.
B. To create a new table based on specified criteria.
C. To manually execute queries on-demand.
D. To edit existing data in the database.

Question 55:

What level of permissions is the owner of a query granted on their query? Select one response.

A. Can View
B. Can Run
C. Can Edit
D. Can Manage

Question 56:

A company needs to analyze a large amount of data stored in its Hadoop cluster. Which of the following best describes the benefit of using Databricks SQL with a Hadoop cluster?

A. Databricks SQL provides faster query processing than traditional Hadoop tools.
B. Databricks SQL allows users to store and analyze data directly in Hadoop.
C. Databricks SQL provides more advanced security features than Hadoop.
D. Databricks SQL provides better support for unstructured data than Hadoop.

Question 57:

A manufacturing company wants to use data from sensors installed on the machinery to continually monitor the performance of its production line. Which of the following Databricks SQL features would be most beneficial in this situation?

A. Databricks SQL can be used to ingest streaming data in real-time
B. Databricks SQL can be used to design and create visualizations using BI tools
C. Databricks SQL can be used to query data across multiple data sources
D. Databricks SQL can be used to handle unstructured data

Question 58:

A data analyst has been asked to create a Databricks SQL query that will summarize sales data by product category and month. Which SQL function can you use to accomplish this?

A. AVG
B. SUM
C. GROUP BY
D. ORDER BY

Question 59:

A data analyst of a large online retailer wants to integrate Databricks SQL with Partner Connect to obtain real-time data on customer behavior from a social media platform. Which of the following steps would the data analyst take to achieve the desired outcome?

1. Use Databricks SQL to ingest the data from the social media platform and then connect it to Partner Connect.
2. Use Partner Connect to ingest the data from the social media platform and then connect it to Databricks SQL.
3. Use an ETL tool to ingest the data from the social media platform and then connect it to both Partner Connect and Databricks SQL.
4. Use an API to ingest the data from the social media platform and then connect it to both Partner Connect and Databricks SQL.

Question 60:

A Data analyst has been tasked with optimizing a Databricks SQL query for a large dataset. What should you consider when trying to improve query performance?

A. Increasing the size of the cluster to handle the data
B. Partitioning the data into smaller chunks
C. Using a higher level of parallelism for the query
D. Increasing the timeout for the query

Question 61:

Which layer of the Medallion Architecture is responsible for providing a unified view of data from various sources?

A. Bronze layer
B. Silver layer
C. Gold layer
D. None of the above

Question 62:

A data analyst has created a Delta Lake table in Databricks and wants to optimize the performance of queries that filter on a specific column. Which Delta Lake feature should the data analyst use to improve query performance?

A. Indexing
B. Partitioning
C. Caching
D. Z-Ordering

Question 63:

What features does Data Explorer in Databricks offer to simplify the management of data, and how do they improve the data management process?

A. Data Explorer provides a visual interface for creating and managing tables, making it easier to navigate and organize data.
B. Data Explorer allows users to create and edit SQL queries directly within the interface, reducing the need to switch between different tools.
C. Data Explorer offers data profiling and visualization tools that can help users better understand the structure and content of their data.
D. All of the above.

Question 64:

A data analyst at a healthcare company is tasked with managing a Databricks table containing personally identifiable information (PII) data, including patients' names and medical histories. The analyst wants to ensure that only authorized personnel can access the table. Which of the following Databricks tools can the analyst use to enforce table ownership and restrict access to the PII data?

A. Delta Lake
B. Access Control Lists
C. Apache Spark
D. Structured Streaming

Question 65:

A data analyst is working with a Delta Lake table which includes changing the data types of a column. Which SQL statement should the data analyst use to modify the column data type?

A. ALTER TABLE table_name ADD COLUMN column_name datatype
B. ALTER TABLE table_name DROP COLUMN column_name
C. ALTER TABLE table_name ALTER COLUMN column_name datatype
D. ALTER TABLE table_name RENAME COLUMN column_name TO new_column_name

Question 66:

A data analyst has been given a requirement of creating a Delta Lake table in Databricks that can be efficiently queried using a specific column as the partitioning column. Which data format and partitioning strategy should the data analyst choose?

A. Parquet file format and partition by hash
B. Delta file format and partition by range
C. ORC file format and partition by list
D. CSV file format and partition by round-robin

Question 67:

A data analyst needs to find out the top 5 customers based on the total amount they spent on purchases in the last 30 days from the sales table. Which of the following Databricks SQL statements will yield the correct result?

A. SELECT TOP 5 customer_id, SUM(price) as total_spent FROM sales WHERE date >= DATEADD(day, -30, GETDATE()) GROUP BY customer_id ORDER BY total_spent DESC;
B. SELECT customer_id, SUM(price) as total_spent FROM sales WHERE date >= DATEADD(day, -30, GETDATE()) GROUP BY customer_id ORDER BY total_spent DESC LIMIT 5;
C. SELECT customer_id, SUM(price) as total_spent FROM sales WHERE date >= DATEADD(day, -30, GETDATE()) GROUP BY customer_id HAVING total_spent > 0 ORDER BY total_spent DESC LIMIT 5;
D. SELECT customer_id, SUM(price) as total_spent FROM sales WHERE date BETWEEN DATEADD(day, -30, GETDATE()) AND GETDATE() GROUP BY customer_id ORDER BY total_spent DESC LIMIT 5;

Question 68:

A large retail company has a Lakehouse that stores data on purchase table made by their stores. The data analyst needs to find the total revenue generated by each store for January. Which of the following SQL statements will return the correct results?

A. SELECT store_id, SUM(total_sales) as revenue FROM purchase WHERE date >= '2023-01-01' AND date <= '2023-01-31' GROUP BY store_id ORDER BY revenue DESC;
B. SELECT store_id, SUM(total_sales) as revenue FROM purchase WHERE date BETWEEN '2023-01-01' AND '2023-01-31' GROUP BY store_id ORDER BY revenue DESC LIMIT 5;
C. SELECT store_id, SUM(total_sales) as revenue FROM purchase WHERE date >= '2023-01-01' AND date <= '2023-01-31' GROUP BY store_id HAVING revenue > 0 ORDER BY revenue DESC;

D. SELECT store_id, SUM(total_sales) as revenue FROM purchase WHERE date >= '2023-01-01' AND date <= '2023-01-31' GROUP BY store_id HAVING revenue > 0 ORDER BY revenue ASC;

Question 69:

A healthcare organization has a Lakehouse that stores data on patient appointments. The data analyst needs to find the average duration of appointments for each doctor. Which of the following SQL statements will return the correct results?

A. SELECT doctor_id, AVG(duration) as avg_duration FROM appointments GROUP BY doctor_id;
B. SELECT doctor_id, AVG(duration) as avg_duration FROM appointments GROUP BY doctor_id HAVING avg_duration > 0;
C. SELECT doctor_id, SUM(duration)/COUNT() as avg_duration FROM appointments GROUP BY doctor_id;
D. SELECT doctor_id, duration/COUNT() as avg_duration FROM appointments GROUP BY doctor_id;

Question 70:

A data analyst is working on a project to analyze a large dataset using Databricks SQL. The dataset is too large to fit in memory, so the analyst needs to use a distributed computing approach. Which Databricks SQL feature will best suit their needs?

A. Dashboards
B. Medallion architecture
C. Compute
D. Streaming data

Question 71:

Which of the following statements about the silver layer in the medallion architecture is true?

A. The silver layer is where data is transformed and processed for analytics use
B. The silver layer is where raw data is stored in its original format
C. The silver layer is optimized for fast querying
D. The silver layer is the largest of the three layers

Question 72:

Which of the following statements accurately describes the role of Delta Lake in the architecture of Databricks SQL?

A. Delta Lake provides data ingestion capabilities for Databricks SQL.
B. Delta Lake is a data storage layer that provides high-performance querying capabilities for Databricks SQL.
C. Delta Lake is a transactional storage layer that provides ACID compliance for data processing in Databricks SQL.
D. Delta Lake provides integration capabilities for Databricks SQL with other BI tools and platforms.

Question 73:

Delta Lake supports schema evolution, which allows for changes to the schema of a table without requiring a full rewrite of the table. Which of the following is not a supported schema evolution operation?

A. Adding a new column
B. Removing a column
C. Renaming a column
D. Changing the data type of a column

Question 74:

A data analyst wants to create a view in Databricks that displays only the top 10% of customers based on their total spending. Which SQL query would achieve this goal?

A. SELECT * FROM customers ORDER BY total_spend DESC LIMIT 10%
B. SELECT * FROM customers WHERE total_spend > PERCENTILE(total_spend, 90)
C. SELECT * FROM customers WHERE total_spend > (SELECT PERCENTILE(total_spend, 90) FROM customers)
D. SELECT * FROM customers ORDER BY total_spend DESC OFFSET 10%

Question 75:

A healthcare company stores patient information in a table in Databricks. The company needs to ensure that only authorized personnel can access the table. Which of the following actions would best address this security concern?

A. Assigning table ownership to a generic company account
B. Granting access to the table to all employees
C. Implementing role-based access control with specific privileges assigned to individual users
D. Storing the patient information in an unsecured Excel file

Question 76:

Objective: Identify the benefits of using Databricks SQL for business intelligence (BI) analytics projects over using third-party BI tools? A data analyst is trying to determine whether to develop their dashboard in Databricks SQL or a partner business intelligence (BI) tool like Tableau, Power BI, or Looker.

When is it advantageous to use Databricks SQL instead of using third-party BI tools to develop the dashboard?

A. When the data being transformed as part of the visualizations is very large
B. When the visualizations require custom formatting
C. When the visualizations require production-grade, customizable branding
D. When the data being transformed is in table format

Question 77:

Objective: Aggregate data columns using SQL functions to answer defined business questions.

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; What is the mistake in the query?

A. The query is selecting region, but region should only occur in the ORDER BY clause.
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. The query is using count(*), which will count all the customers in the customers table, no matter the region.

Question 78:

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 code block 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 function(price(customer_spend, customer_units)) AS customer_price FROM customer_summary
B. SELECT double(price(customer_spend, customer_units)) AS customer_price FROM customer_summary
C. SELECT price FROM customer_summary
D. SELECT 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 79:

Where in the Databricks SQL workspace can a data analyst configure a refresh schedule for a query when the query is not attached to a dashboard or alert?

A. The Dashboard Editor
B. The Visualization Editor
C. The Query Editor
D. SQL Warehouse
E. Data Explorer

Question 80:

A data analyst is working with gold-layer tables to complete an ad-hoc project. A stakeholder has provided the analyst with an additional dataset that can be used to augment the gold-layer tables already in use. Which term is used to describe this data augmentation?

A. Data testing
B. Last-mile ETL
C. Ad-hoc improvements
D. Data enhancement
E. Last-mile dashboarding

Question 81:

How can a data analyst determine if query results were pulled from the cache?

A.Go to the Query History tab and click on the text of the query. The slideout shows if the results came from the cache.
B.Go to the Alerts tab and check the Cache Status alert.
C.Go to the Queries tab and click on Cache Status. The status will be green if the results from the last run came from the cache.
D.Go to the SQL Warehouse (formerly SQL Endpoints) tab and click on Cache. The Cache file will show the contents of the cache.
E.Go to the Data tab and click Last Query. The details of the query will show if the results came from the cache.

Question 82:

A data analyst has created a Query in Databricks SQL, and now they want to create two data visualizations from that Query and add both of those data visualizations to the same Databricks SQL Dashboard.
Which of the following steps will they need to take when creating and adding both data visualizations to the Databricks SQL Dashboard?

A.They will need to alter the Query to return two separate sets of results.
B.They will need to add two separate visualizations to the dashboard based on the same Query.
C.They will need to create two separate dashboards.
D.They will need to decide on a single data visualization to add to the dashboard.
E.They will need to copy the Query and create one data visualization per query.

Question 83:

Which of the following is a benefit of Databricks SQL using ANSI SQL as its standard SQL dialect?

A.It has increased customization capabilities
B.It is easy to migrate existing SQL queries to Databricks SQL
C.It allows for the use of Photon's computation optimizations
D.It is more performant than other SQL dialects
E.It is more compatible with Spark's interpreters

Question 84:

How can a data analyst determine if query results were pulled from the cache?

A.Go to the Query History tab and click on the text of the query. The slideout shows if the results came from the cache.
B.Go to the Alerts tab and check the Cache Status alert.
C.Go to the Queries tab and click on Cache Status. The status will be green if the results from the last run came from the cache.
D.Go to the SQL Warehouse (formerly SQL Endpoints) tab and click on Cache. The Cache file will show the contents of the cache.
E.Go to the Data tab and click Last Query. The details of the query will show if the results came from the cache.

Answer Key:

Question 1: B
Question 2: B
Question 3: C
Question 4: B
Question 5: E
Question 6: B
Question 7: B
Question 8: A
Question 9: D
Question 10: B
Question 11: C
Question 12: B
Question 13: E
Question 14: B
Question 15: C
Question 16: B

Question 17: E
Question 18: E
Question 19: C
Question 20: C
Question 21: E
Question 22: A
Question 23: E
Question 24: C
Question 25: E
Question 26: C
Question 27: E
Question 28: E (chances are if says all of the above, that's the answer haha)
Question 29: C
Question 30: A
Question 31: B
Question 32: C
Question 33: B
Question 34: D
Question 35: A
Question 36: C (but D makes more sense...it is an all of the above question...)
Question 37: A
Question 38: C (but D makes more sense...in real world application...)
Question 39: B
Question 40: D
Question 41: B
Question 42: A&B
Question 43: B
Question 44: A (but D makes more sense...it is an all of the above question...)
Question 45: C (but B seems to make sense too....)
Question 46: B
Question 47: C
Question 48: B
Question 49: C
Question 50: A
Question 51: B
Question 52: D
Question 53: C
Question 54: A
Question 55: D
Question 56: A
Question 57: A
Question 58: C
Question 59: B
Question 60: B
Question 61: C
Question 62: D
Question 63: D
Question 64: B
Question 65: C
Question 66: A
Question 67: B
Question 68: A
Question 69: A

Question 70: C
Question 71: A
Question 72: C
Question 73: B
Question 74: C
Question 75: C
Question 76: A
Question 77: B
Question 78: E
Question 79: C
Question 80: D
Question 81: A
Question 82: B
Question 83: B
Question 84: A