DataBricks

Question 1

- **Question:** Which SQL command should a data analyst use to create a new table in Databricks SQL that includes only unique values from an existing table?
- **Correct Answer:** CREATE TABLE AS SELECT DISTINCT * FROM existing_table;
- **Explanation:** The SQL command `CREATE TABLE AS SELECT DISTINCT * FROM existing_table` creates a new table that includes only unique values from the existing table. The `DISTINCT` keyword ensures that duplicate values are not included in the new table.

Question 2

- **Question:** To quickly create SQL queries and visualizations with serverless compute capabilities, which Databricks Lakehouse Platform service should a data analyst use?
- **Correct Answer:** Delta Lake
- **Explanation:** Delta Lake is the correct choice for quickly creating SQL queries and visualizations with serverless compute capabilities on the Databricks Lakehouse Platform. Delta Lake provides ACID transactions, scalable metadata handling, and data versioning.

Question 3

- **Question:** When using Databricks SQL, what is a key consideration when configuring a dashboard to automatically refresh based on a live data source?
- **Correct Answer:** Monitoring the impact on performance and costs
- **Explanation:** Monitoring the impact on performance and costs is a key consideration when configuring a dashboard to automatically refresh based

on a live data source because live data sources can strain system resources and increase costs.

Question 4

- **Question:** Which of the following layers in the medallion architecture is primarily used for storing raw data?
- **Correct Answer:** Bronze
- **Explanation:** The Bronze layer in the medallion architecture is primarily used for storing raw data before any processing or transformation occurs. It serves as the initial storage layer where data is ingested and stored in its original form.

Question 5

- **Question:** When is it most appropriate for a data analyst to use a heatmap in Databricks SQL?
- **Correct Answer:** To visualize the relationship and correlation between two variables
- **Explanation:** Heatmaps are most appropriate for visualizing the relationship and correlation between two variables in Databricks SQL. The color intensity in the heatmap represents the strength of the relationship.

Question 6

- **Question:** In Databricks SQL, when should a data analyst use higher-order functions like map or filter?
- **Correct Answer:** When custom logic needs to be applied at scale to complex data types like arrays
- **Explanation:** Higher-order functions like map or filter are used when custom logic needs to be applied at scale to complex data types like arrays in Databricks SQL. These functions allow for efficient processing and manipulation of data within arrays.

Question 7

- **Question:** What is the best practice for a data analyst to follow when performing aggregations on large datasets to ensure optimal performance in

Databricks SQL?

- **Correct Answer:** Utilize window functions for partitioned aggregations
- **Explanation:** Utilizing window functions for partitioned aggregations is a best practice for optimizing performance in Databricks SQL. Window functions allow for efficient processing of data within partitions, reducing the need for costly joins.

Question 8

- **Question:** If a data analyst wants to visualize the cumulative total of a metric over time in Databricks SQL, which chart type should they use?
- **Correct Answer:** Cumulative Line Chart
- **Explanation:** A cumulative line chart is the most appropriate chart type for visualizing the cumulative total of a metric over time in Databricks SQL. It shows the running total of the metric as it accumulates over time.

Question 9

- **Question:** In Databricks SQL, when should a data analyst use the JOIN clause with an ON condition?
- **Correct Answer:** To combine rows from two or more tables based on a related column
- **Explanation:** The JOIN clause with an ON condition is used to combine rows from two or more tables based on a related column. This allows the data analyst to retrieve related information from multiple tables in a single query.

Question 10

- **Question:** What feature in Databricks SQL enables data analysts to maintain version control and audit trails for their datasets?
- **Correct Answer:** Delta Lake transaction logs
- **Explanation:** Delta Lake transaction logs in Databricks SQL enable data analysts to maintain version control and audit trails for their datasets. These transaction logs track all changes made to the data, allowing for easy rollback to previous versions and ensuring data integrity and traceability.

Question 11

- **Question:** To ensure data persistence and reuse across sessions in Databricks, which type of table should a data analyst create?
- **Correct Answer:** Managed table
- **Explanation:** Managed tables in Databricks store data within the Databricks environment and manage the data's lifecycle. They ensure data persistence and reuse across sessions.

Question 12

- **Question:** If a data analyst wants to ensure that all data and metadata associated with a table are removed when the table is dropped, which type of table should they avoid using?
- **Correct Answer:** External table
- **Explanation:** External tables in Databricks are not fully managed by the platform, and the data and metadata associated with an external table are stored outside of the Databricks environment. When an external table is dropped, Databricks does not automatically remove the associated data and metadata.

Question 13

- **Question:** When should a data analyst use the PERCENT_RANK() window function in Databricks SQL?
- **Correct Answer:** To calculate the percentage rank of rows within a partition of a dataset
- **Explanation:** The PERCENT_RANK() window function is specifically designed to calculate the percentage rank of rows within a partition of a dataset. It assigns a percentile rank value to each row based on its position within the partition.

Question 14

- **Question:** In Databricks SQL, which visualization type is best suited for showing the distribution of a single numeric variable?
- **Correct Answer:** Histogram

- **Explanation:** Histograms are specifically designed to show the distribution of a single numeric variable by dividing the data into bins or intervals and displaying the frequency of values within each bin.

Question 15

- **Question:** Which Databricks SQL feature allows a data analyst to configure a visualization to dynamically update based on user input or parameter changes?
- **Correct Answer:** Parameterized queries
- **Explanation:** Parameterized queries in Databricks SQL allow a data analyst to configure a visualization to dynamically update based on user input or parameter changes. By defining parameters in the query, users can interact with the visualization and see real-time updates based on their selections.

Question 16

- **Question:** If a data analyst wants to ensure that a specific view always reflects the latest changes from its underlying tables in Databricks, which SQL statement should they use?
- **Correct Answer:** CREATE OR REPLACE VIEW
- **Explanation:** The SQL statement CREATE OR REPLACE VIEW is used to create a new view or replace an existing view with the same name. This ensures that the view always reflects the latest changes from its underlying tables in Databricks by updating the view definition.

Question 17

- **Question:** When should a data analyst use a Sankey diagram in Databricks SQL for data visualization?
- **Correct Answer:** To show the relationship between hierarchical categories and flow between them
- **Explanation:** Sankey diagrams are ideal for visualizing the flow and relationships between hierarchical categories. They show how values are distributed or transferred from one category to another.

Question 18

- **Question:** What is the primary benefit of using Delta Lake’s ACID transactions for data integrity in a data lakehouse architecture?
- **Correct Answer:** Guaranteed data consistency and reliability during concurrent operations
- **Explanation:** The primary benefit of using Delta Lake's ACID transactions is guaranteed data consistency and reliability during concurrent operations. ACID transactions help maintain the integrity of the data by ensuring that operations are completed successfully and reliably, even when multiple operations are being performed simultaneously.

Question 19

- **Question:** What is the main advantage of using Partner Connect in Databricks for integrating with external tools like Fivetran?
- **Correct Answer:** It allows automated setup of clusters and connections
- **Explanation:** Partner Connect in Databricks simplifies the integration process with external tools like Fivetran by automating the setup of clusters and connections. This reduces the need for manual configuration and allows for quicker, more efficient integration.

Question 20

- **Question:** In Databricks SQL, which operation would a data analyst use to aggregate data while also preserving all the individual rows?
- **Correct Answer:** WINDOW FUNCTION
- **Explanation:** WINDOW FUNCTION is the correct choice as it allows a data analyst to perform aggregate functions while still preserving all individual rows in the result set. Window functions operate on a set of rows related to the current row, allowing for calculations across the dataset without collapsing the rows.

Question 21

- **Question:** Which visualization type in Databricks SQL allows data analysts to clearly designate different sections such as "Development," "Testing," and "Production" using formatted text?

- **Correct Answer:** Markdown-based text boxes
- **Explanation:** Markdown-based text boxes in Databricks SQL allow data analysts to add formatted text directly into the visualization, making it easy to designate different sections with clear formatting and styling.

Question 22

- **Question:** Which SQL window function is most appropriate for calculating a running total in Databricks SQL?
- **Correct Answer:** SUM()
- **Explanation:** The SUM() function is the most appropriate SQL window function for calculating a running total in Databricks SQL. It calculates the sum of a specified column for a group of rows and is commonly used for running totals.

Question 23

- **Question:** What is a critical consideration for a data analyst when setting up alerts in Databricks SQL that are based on queries with parameters?
- **Correct Answer:** Alerts only work with queries that have static results
- **Explanation:** Alerts in Databricks SQL are designed to work with queries that have static results. This means that the alert conditions are based on the specific data returned by the query at the time of execution.

Question 24

- **Question:** How should Databricks SQL be used to complement other data engineering tools such as Apache Spark?
- **Correct Answer:** As a tool to simplify SQL-based analytics after data is processed in Apache Spark
- **Explanation:** Databricks SQL is commonly used to simplify SQL-based analytics after data has been processed in Apache Spark. It allows data analysts to easily query and analyze data stored in Apache Spark using familiar SQL syntax.

Question 25

- **Question:** Which of the following best describes a situation where a LEFT SEMI JOIN would be used in Databricks SQL?

- **Correct Answer:** When you want to return only rows from the left table that have matching rows in the right table
- **Explanation:** A LEFT SEMI JOIN in Databricks SQL is specifically used to return only the rows from the left table that have matching rows in the right table. It filters out rows from the left table that do not have corresponding matches in the right table.

Question 26

- **Question:** If a data analyst wants to set up a scheduled query in Databricks SQL to run every day at midnight, which option should they select?
- **Correct Answer:** Job scheduling with CRON expressions
- **Explanation:** Job scheduling with CRON expressions allows the data analyst to set up a scheduled query to run every day at midnight. CRON expressions provide a flexible and precise way to define the schedule for running jobs or queries at specific times.

Question 27

- **Question:** In Databricks SQL, which type of join should be used to return all rows from the left table and the matched rows from the right table?
- **Correct Answer:** LEFT JOIN
- **Explanation:** A LEFT JOIN returns all rows from the left table and the matched rows from the right table. If there is no match, it returns NULL values for the columns from the right table.

  1

  2

Question 28

- **Question:** What is the purpose of using a temporary view in Databricks SQL?
- **Correct Answer:** To store the result of a query for the current session
- **Explanation:** A temporary view is used to store the result of a query for the current session. It is not stored permanently and is only available for the duration of the session in which it was created.

Question 29

- **Question:** When should a data analyst use the RANK() window function in Databricks SQL?
- **Correct Answer:** To assign a unique rank to each row within a partition based on a specified ordering
- **Explanation:** The RANK() window function assigns a unique rank to each row within a partition based on a specified ordering. It assigns the same rank to rows with equal values and skips the next ranks accordingly.

Question 30

- **Question:** Which chart type is most appropriate for visualizing the composition of different categories as parts of a whole in Databricks SQL?
- **Correct Answer:** Pie chart
- **Explanation:** A pie chart is most appropriate for visualizing the composition of different categories as parts of a whole. Each slice represents a category, and the size of the slice corresponds to its proportion of the total.

Question 31

- **Question:** What does the `nvl()` function do in Databricks SQL?
- **Correct Answer:** Replaces null values with a specified replacement value
- **Explanation:** The `nvl()` function replaces null values with a specified replacement value. If the first argument is null, it returns the second argument; otherwise, it returns the first argument.

Question 32

- **Question:** Which SQL clause is used to filter rows based on a range of values in Databricks SQL?
- **Correct Answer:** BETWEEN
- **Explanation:** The `BETWEEN` clause is used to filter rows based on a range of values. It selects rows where a specified value is within a given range (inclusive).

Question 33

- **Question:** In Databricks SQL, what is the purpose of the `explode()` function?
- **Correct Answer:** To transform elements from an array or map into separate rows
- **Explanation:** The `explode()` function transforms elements from an array or map into separate rows. It is used to denormalize nested data structures, making it easier to analyze individual elements.

Question 34

- **Question:** Which SQL function is used to calculate the median value of a column in Databricks SQL?
- **Correct Answer:** APPROX_MEDIAN()
- **Explanation:** The `APPROX_MEDIAN()` function is used to calculate the approximate median value of a column in Databricks SQL. It provides an efficient way to estimate the median, especially for large datasets.

Question 35

- **Question:** What is the benefit of using the `cache table` command in Databricks SQL?
- **Correct Answer:** It speeds up query execution by storing the table data in memory
- **Explanation:** The `CACHE TABLE` command speeds up query execution by storing the table data in memory. This reduces the need to read data from disk repeatedly, resulting in faster query performance.

Question 36

- **Question:** Which of the following is an accurate description of a "Data Lakehouse"?
- **Correct Answer:** A data management architecture that combines the flexibility of data lakes with the data management capabilities of data warehouses
- **Explanation:** A Data Lakehouse is a data management architecture that combines the flexibility of data lakes (for storing diverse data types) with the data management capabilities of data warehouses (for structured querying and ACID transactions).

Question 37

- **Question:** What is a key benefit of using Delta Lake over traditional Apache Spark?
- **Correct Answer:** Support for ACID transactions
- **Explanation:** A key benefit of using Delta Lake over traditional Apache Spark is its support for ACID transactions. This ensures data consistency and reliability, which is critical for data warehousing and business intelligence applications.

Question 38

- **Question:** In Databricks SQL, how can a data analyst create a user-defined function (UDF)?
- **Correct Answer:** By using the CREATE TEMPORARY FUNCTION or CREATE FUNCTION statement
- **Explanation:** A data analyst can create a user-defined function (UDF) in Databricks SQL by using the `CREATE TEMPORARY FUNCTION` (for session-scoped UDFs) or `CREATE FUNCTION` (for catalog-scoped UDFs) statement.

Question 39

- **Question:** What is the purpose of the `lag()` function in Databricks SQL?
- **Correct Answer:** To access data from a previous row in the same result set
- **Explanation:** The `lag()` function is used to access data from a previous row in the same result set. It allows data analysts to compare current row values with values from preceding rows.

Question 40

- **Question:** Which type of chart is best for comparing categorical data against a numerical value in Databricks SQL?
- **Correct Answer:** Bar chart
- **Explanation:** A bar chart is best for comparing categorical data against a numerical value. Each bar represents a category, and the length of the bar corresponds to the numerical value.

Question 41

- **Question:** What is the function of the `collect_list()` function in Databricks SQL?
- **Correct Answer:** To return a list of values from a column in a group
- **Explanation:** The `collect_list()` function returns a list of values from a column in a group. It aggregates all values from a column into an array within each group.

Question 42

- **Question:** How does Databricks SQL handle nested data formats like JSON?
- **Correct Answer:** It provides built-in functions to parse and query JSON data
- **Explanation:** Databricks SQL provides built-in functions to parse and query JSON data. This allows data analysts to efficiently extract and analyze information from nested JSON structures.

Question 43

- **Question:** In Databricks SQL, what does the `first()` function return?
- **Correct Answer:** The first value in a group
- **Explanation:** The `first()` function returns the first value in a group. It is an aggregate function that retrieves the first value of a specified column within each group of rows.

Question 44

- **Question:** Which SQL statement is used to modify the structure of an existing table in Databricks SQL?
- **Correct Answer:** ALTER TABLE
- **Explanation:** The `ALTER TABLE` statement is used to modify the structure of an existing table. It can be used to add, modify, or delete columns, constraints, or other table properties.

Question 45

- **Question:** What is the primary purpose of the Databricks SQL Query History?
- **Correct Answer:** To provide a record of all executed SQL queries for auditing and debugging

- **Explanation:** The primary purpose of the Databricks SQL Query History is to provide a record of all executed SQL queries. This history is valuable for auditing purposes, tracking query execution, and debugging issues.

Question 1

- **Question:** Which of the following layers of the medallion architecture is most commonly used by data analysts?
- **Correct Answer:** Gold
- **Explanation:** Gold is the correct choice as it is the layer of the medallion architecture that is most commonly used by data analysts. The Gold layer typically contains curated and transformed data that is ready for analysis and reporting.

Question 2

- **Question:** 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?
- **Correct Answer:** SQL Editor page
- **Explanation:** The SQL Editor page in Databricks SQL is specifically designed for writing and executing SQL queries. It provides a dedicated environment for data analysts to interact with the data using SQL syntax, making it the correct choice for where the analyst can write and execute SQL queries.

Question 3

- **Question:** 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?
- **Correct Answer:** As a complementary tool for quick in-platform BI work
- **Explanation:** Using Databricks SQL as a complementary tool for quick in-platform BI work is a common and effective approach. Databricks SQL can be used to quickly analyze and visualize data within the Databricks platform,

while other BI tools like Tableau, Power BI, and Looker can be used for more advanced analytics, visualizations, and professional presentations.

Question 4

- **Question:** Which of the following approaches can be used to connect Databricks to Fivetran for data ingestion?
- **Correct Answer:** Use Partner Connect's automated workflow to establish a SQL warehouse (formerly known as a SQL endpoint) for Fivetran to interact with
- **Explanation:** Using Partner Connect's automated workflow to establish a SQL warehouse (formerly known as a SQL endpoint) for Fivetran to interact with is the correct approach for connecting Databricks to Fivetran for data ingestion. This method allows for seamless integration between Databricks and Fivetran, enabling efficient data ingestion processes.

Question 5

- **Question:** 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?
- **Correct Answer:** Data engineer
- **Explanation:** Data engineers primarily use other Databricks services like Databricks Machine Learning or Data Science and Engineering for tasks related to data processing, modeling, and pipeline development. They may use Databricks SQL as a secondary service for querying and analyzing data.

Question 6

- **Question:** 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?
- **Correct Answer:** Turn off the Auto stop feature

- **Explanation:** Turning off the Auto stop feature will prevent the SQL endpoint from shutting down and starting up with each run, reducing the start-up time significantly. This change can help optimize performance and reduce delays while managing costs effectively.

Question 7

- **Question:** A data engineering team has created a Structured Streaming pipeline that processes data in micro-batches and populates gold-level tables. The micro-batches 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?
- **Correct Answer:** The required compute resources could be costly
- **Explanation:** Updating a dashboard in near real-time (within one minute) requires significant compute resources to constantly refresh and process new data. This can lead to high costs, especially if the dashboard relies on continuously streaming data from the gold-level tables. The data analyst should caution stakeholders about the potential increase in costs due to the frequent updates.

Question 8

- **Question:** Which of the following approaches can be used to ingest data directly from cloud-based object storage?
- **Correct Answer:** Create an external table while specifying the object storage path to LOCATION
- **Explanation:** Creating an external table while specifying the object storage path to LOCATION is the correct approach for ingesting data directly from cloud-based object storage. By specifying the location parameter, you can directly access and ingest data from cloud-based object storage like Amazon S3 or Azure Blob Storage in Databricks.

Question 9

- **Question:** 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?
- **Correct Answer:** Markdown-based text boxes
- **Explanation:** Markdown-based text boxes are the correct choice for designating the Development, Testing, and Production sections using text on the dashboard. Markdown allows for easy formatting of text, including headers, lists, and emphasis, making it ideal for creating clear section labels on a dashboard.

Question 10

- **Question:** 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?
- **Correct Answer:** Databricks SQL
- **Explanation:** Databricks SQL is a service within the Databricks Lakehouse Platform that allows users to run SQL queries on data lakes and data warehouses. It provides serverless compute resources, supports data visualization capabilities, and allows users to create dashboards for visualizing query results, making it the correct choice that meets all the specified requirements.

Question 11

- **Question:** 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?
- **Correct Answer:** The table was external

- **Explanation:** If the table was external, the DROP TABLE command only removes the metadata associated with the table, not the actual data files. External tables in Databricks are stored outside the default location and dropping them does not delete the data files.

Question 12

- **Question:** After running DESCRIBE EXTENDED accounts.customers; the following was returned: Now a data analyst runs the following command: DROP accounts.customers; Which of the following describes the result of running this command?
- **Correct Answer:** The accounts.customers table is removed from the metastore but the underlying data files are untouched.
- **Explanation:** The correct explanation is that the DROP command removes the accounts.customers table from the metastore, but it does not delete the underlying data files. This means that the data files associated with the table will still exist even after dropping the table.

Question 13

- **Question:** Which of the following should data analysts consider when working with personally identifiable information (PII) data?
- **Correct Answer:** All of these considerations
- **Explanation:** All of these considerations are essential for data analysts when working with PII data to ensure compliance with organizational policies, legal regulations, and data protection standards. By considering all these factors, data analysts can effectively manage and protect sensitive information.

Question 14

- **Question:** 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?
- **Correct Answer:** Table metadata
- **Explanation:** Table metadata is stored alongside data files when using Delta Lake. This metadata includes information about the table schema, data statistics, and other properties that help manage and optimize data operations.

Question 15

- **Question:** Which of the following is an advantage of using a Delta Lake-based data lakehouse over common data lake solutions?
- **Correct Answer:** ACID transactions
- **Explanation:** ACID transactions are a key advantage of using a Delta Lake-based data lakehouse. Delta Lake provides support for ACID (Atomicity, Consistency, Isolation, Durability) transactions, ensuring data integrity and consistency, which is crucial for data reliability and quality in data lake environments.

Question 16

- **Question:** Which of the following benefits of using Databricks SQL is provided by Data Explorer?
- **Correct Answer:** It can be used to view metadata and data as well as view/change permissions.
- **Explanation:** Data Explorer in Databricks SQL allows users to view metadata and data, as well as view and change permissions. This functionality is essential for data exploration and management within the platform.

Question 17

- **Question:** The stakeholders.customers table has 15 columns and 3000 rows of data. The following command is run: 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?
- **Correct Answer:** The view has been dropped.
- **Explanation:** The view has been dropped. - This statement is correct. Views in Databricks are temporary and are dropped when the user logs out. Therefore, the stakeholders.eur_customers view will no longer be available after logging back in.

Question 18

- **Question:** 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?
- **Correct Answer:** Edit the Owner field on the table page by selecting the new owner's account
- **Explanation:** Editing the Owner field on the table page by selecting the new owner's account is the correct approach to change ownership of the table to a single other user. By selecting the new owner's account, the data analyst can transfer ownership of the table to the specified user while retaining proper access control.

Question 19

- **Question:** A data analyst wants to delete the table database_name.table_name. Which of the following commands will successfully complete the task?
- **Correct Answer:** DROP TABLE database_name.table_name;
- **Explanation:** The DROP TABLE command is used to remove a specific table from the database. By specifying both the database name and table name, the analyst can successfully delete the table and its associated data files without affecting the other tables in the database.

Question 20

- **Question:** A data analyst runs the following SQL query against a table with the following schema: | id | age | country | | :--- | :-- | :------ | | 900 | 80 | canada | | 901 | 75 | canada | | 902 | 90 | canada | | 903 | NULL | usa | | 904 | 60 | usa | | 905 | NULL | usa | SELECT * FROM table1 WHERE country = 'canada' AND age >= 75; Which of the following tables will be returned?
- **Correct Answer:** | id | age | country | |-----|-----|---------| | 900 | 80 | canada | | 901 | 75 | canada | | 902 | 90 | canada |
- **Explanation:** This table includes rows where the country is 'canada', and the age values are 80, 75, and 90, which meet the condition of age being greater than or equal to 75.

Question 21

- **Question:** The suppliers table has the columns supplier_id, supplier_name, and city. The new_suppliers table has the columns supplier_id, supplier_name, city, and state. A data analyst runs the following command: INSERT INTO suppliers SELECT supplier_id, supplier_name, city FROM new_suppliers; Which of the following statements is correct?
- **Correct Answer:** 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.
- **Explanation:** This choice is correct because the INSERT INTO command adds new data from one table to another. In this case, it adds the supplier_id, supplier_name, and city from the new_suppliers table to the suppliers table. The suppliers table will retain its original data, and the data from new_suppliers will be appended to it. No data is deleted during this process, so any duplicate rows between the two tables will also be inserted. Thus, the suppliers table will end up containing both its original data and all data from new_suppliers, including duplicates.

Question 22

- **Question:** A data engineer is working with a nested array column products in the 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: 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?
- **Correct Answer:** explode(products)
- **Explanation:** The "explode(products)" command is the correct choice for expanding a nested array column into individual rows. It takes each item within the products array and creates a new row for it, while duplicating the associated transaction_id column for each expanded row. This approach is commonly used to normalize data structures that contain nested arrays for more straightforward analysis and querying.

Question 23

- **Question:** 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?
- **Correct Answer:** CREATE TABLE table_silver AS SELECT DISTINCT * FROM table_bronze;
- **Explanation:** The query "CREATE TABLE table_silver AS SELECT DISTINCT * FROM table_bronze;" is the correct choice for deduplicating data from table_bronze and writing it to a new table table_silver. The SELECT DISTINCT statement retrieves only the unique rows from table_bronze, effectively deduplicating the data. This method ensures that any duplicative entries present in table_bronze are removed when the data is inserted into table_silver. The other options do not correctly implement deduplication or contain incorrect syntax for the task.

Question 24

- **Question:** 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?
- **Correct Answer:** 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;
- **Explanation:** This code block creates or replaces a table named sales_by_employee by joining the sales and employees tables on the employee_id column. By using a table instead of a view, the data entity will be persistent and always stay up-to-date when new data are added to the sales table.

Question 25

- **Question:** A data analyst is working with the sales_table table. They want to calculate the percentage rank of products within each region based on sales. Which of the following SQL queries will correctly calculate the percentage rank as rank?
- **Correct Answer:** SELECT region, product, PERCENT_RANK() OVER (PARTITION BY region ORDER BY sales DESC) AS rank FROM sales_table;
- **Explanation:** This query is the correct choice. It uses the PERCENT_RANK() function with the OVER clause to calculate the percentage rank of products within each region based on sales. The PARTITION BY clause ensures that the calculation is done separately for each region, meeting the requirements of the task.

Question 26

- **Question:** A data analyst is working with a database that contains customer information. They need to apply custom logic at scale to array data objects within the database. Which of the following should the data analyst use?
- **Correct Answer:** When custom logic needs to be applied at scale to array data objects
- **Explanation:** Using higher-order functions is beneficial when custom logic needs to be applied at scale to array data objects. Higher-order functions allow for operations to be performed on each element of an array efficiently, making them a suitable choice for scenarios where array data objects are involved.

Question 27

- **Question:** A data analyst runs the following two SQL statements against two tables named customers and orders, both of which have a customer_id column: SELECT * FROM customers WHERE customer_id IN (SELECT customer_id FROM orders); SELECT * FROM customers WHERE customer_id NOT IN (SELECT customer_id FROM orders); Which of the following is the most accurate description of the result sets from the two statements?
- **Correct Answer:** 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.

- **Explanation:** This explanation is correct because it accurately describes how the result sets will differ based on the use of IN and NOT IN subqueries. The first statement uses IN, which returns only rows from the customers table where the customer_id exists in the orders table. The second statement uses NOT IN, which returns only rows from the customers table where the customer_id does not exist in the orders table.

Question 28

- **Question:** 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; They now want to run this function on the columns customer_spend and customer_units from the table customer_summary. Which of the following SQL queries is correctly written to accomplish this?

  1

- **Correct Answer:** SELECT price(customer_spend, customer_units) AS customer_price FROM customer_summary
- **Explanation:** Choice E correctly applies the user-defined function price() to the columns customer_spend and customer_units in the SELECT statement to create the new column customer_price. The syntax is valid and will generate the desired output.

Question 29

- **Question:** A data analyst wants to find the total number of customers for each region in the customers table. Which of the following SQL queries will return the correct result?
- **Correct Answer:** SELECT region, count(*) AS num_customers FROM customers GROUP BY region
- **Explanation:** The query uses the `GROUP BY` clause to group the customers by region and the `count(*)` function to count the number of customers in each group. This will provide the total number of customers for each region.

Question 30

- **Question:** A data analyst is working with a table named my_table that has the following schema: | group_1 | group_2 | values |. They want to create a query that will return the count of values for each combination of group_1 and group_2, as well as the overall count. Which of the following queries can be used to obtain the result shown?
- **Correct Answer:** SELECT group_1, group_2, count(values) AS count FROM my_table GROUP BY group_1, group_2 WITH CUBE;
- **Explanation:** This query includes the "WITH CUBE" clause, which is used to generate subtotals and grand totals for all possible combinations of the specified grouping columns. This will provide the count of values for each combination of group_1 and group_2, as well as the overall count.

Question 31

- **Question:** A data analyst is tasked with visualizing the flow of users through a website. Which of the following visualization types is most appropriate for this task?
- **Correct Answer:** Sankey
- **Explanation:** Sankey diagrams are specifically designed to visualize the flow and relationships between different entities or categories. They use interconnected nodes and links to show the movement or transition of users, making them the ideal choice for visualizing the flow of users through a website.

Question 32

- **Question:** A data analyst is working with a table that has the schema | region | number_of_customer |. They want to create a visualization to analyze the data. Which of the following is the default visualization type that Databricks SQL will select?
- **Correct Answer:** Bar chart
- **Explanation:** Bar charts are ideal for comparing discrete categories or groups, making them a suitable choice for visualizing the data with the provided schema.

Question 33

- **Question:** A data analyst adds an area chart to a Databricks SQL Dashboard. The query used by the area chart has a Dashboard Parameter applied to it. Which of the following statements is correct?
- **Correct Answer:** The area chart will use the value selected in the Dashboard Parameter, and this value will also be applied to all other visualizations in the dashboard that use the same parameter.
- **Explanation:** The area chart will use the value selected in the Dashboard Parameter, and this value will also be applied to all other visualizations in the dashboard that use the same parameter.

Question 34

- **Question:** 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?
- **Correct Answer:** Table metadata
- **Explanation:** Table metadata is stored alongside data files when using Delta Lake. This metadata includes information about the table schema, data types, partitions, and other properties that are essential for managing and querying the data effectively.

Question 35

- **Question:** A data analyst runs a SQL query in Databricks SQL. They want to confirm if the results from the last run came from the cache. Which of the following is the correct way to determine this?
- **Correct Answer:** Go to the Query History tab and inspect the Full Duration
- **Explanation:** The Full Duration in the Query History tab will indicate how long the query took to run. If the query results were pulled from the cache, the Full Duration will be significantly shorter compared to when the query had to read the data from the source.

Question 36

- **Question:** Which of the following is a benefit of using ANSI SQL as the standard SQL dialect in Databricks SQL?
- **Correct Answer:** It makes it easy to migrate existing SQL queries to Databricks SQL
- **Explanation:** Using ANSI SQL as the standard SQL dialect in Databricks SQL makes it easy to migrate existing SQL queries to the platform. This is because ANSI SQL is a common standard that is widely supported across different database systems, making the transition smoother and more efficient.

Question 37

- **Question:** A data analyst has created a Databricks SQL Query that produces two different data visualizations. They now want to add both of these data visualizations to the same Databricks SQL Dashboard. Which of the following is the correct approach?
- **Correct Answer:** They will need to add two separate visualizations to the dashboard based on the same Query.
- **Explanation:** Adding two separate visualizations to the dashboard based on the same Query is the correct approach to create and add both data visualizations to the Databricks SQL Dashboard. This allows the data analyst to display different perspectives or insights from the same dataset on the same dashboard.

Question 38

- **Question:** 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?
- **Correct Answer:** The required compute resources could be costly
- **Explanation:** The required compute resources for updating the dashboard within one minute of new data becoming available in the gold-level tables

could be costly. Processing data in micro-batches every minute and updating a dashboard in near real-time can consume significant compute resources.

Question 39

- **Question:** 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?
- **Correct Answer:** The table was external
- **Explanation:** When dropping an external table in Databricks, the DROP TABLE command removes the table's metadata from the metastore but does not delete the underlying data files stored in external storage locations like AWS S3, Azure Blob Storage, or Google Cloud Storage. This behavior allows users to manage their data files independently of the table definitions in Databricks.

Question 40

- **Question:** A data analyst has set up a Databricks SQL Alert on a Query with a configurable Date Range query parameter. However, the alert is not working as expected. Which of the following is the most likely reason?
- **Correct Answer:** Query parameters cannot be used with Alerts.
- **Explanation:** Alerts in Databricks do not support queries that use query parameters. This limitation is because alerts require a consistent basis for evaluation, and parameterized queries can produce varying results, making it difficult to establish reliable alert conditions.

Question 41

- **Question:** Which of the following best describes descriptive statistics?
- **Correct Answer:** A branch of statistics that uses summary statistics to quantitatively describe and summarize data.
- **Explanation:** Descriptive statistics uses summary statistics to quantitatively describe and summarize data. It provides numerical summaries that convey information about the dataset as a whole.

Question 42

- **Question:** Under which of the following conditions will the mean and median values of a variable be meaningfully different?
- **Correct Answer:** When the variable contains a lot of extreme outliers
- **Explanation:** When the variable contains a lot of extreme outliers, the mean and median values will be significantly different. Outliers can skew the mean value, pulling it towards the extreme values, while the median remains unaffected by the presence of outliers.

Question 43

- **Question:** Which of the following is true about Databricks SQL scheduled queries?
- **Correct Answer:** They can be set up in the Query Editor, can range from 1 minute to 2 weeks, differ from alerts, and require appropriate permissions to configure.
- **Explanation:** Databricks SQL scheduled queries can be set up in the Query Editor, can range from 1 minute to 2 weeks, differ from alerts, and require appropriate permissions to configure.

Question 44

- **Question:** A data team has been given a series of projects by a consultant that need to be implemented in the Databricks Lakehouse Platform. Which of the following projects should be completed in Databricks SQL?
- **Correct Answer:** Create a series of ad-hoc queries to explore a table
- **Explanation:** Databricks SQL is designed for SQL-based analytics and is suitable for creating ad-hoc queries to explore and analyze data in tables.

Question 45

- **Question:** Which of the following is NOT true about Databricks SQL alerts?
- **Correct Answer:** They are set up in the SQL Warehouse tab
- **Explanation:** Databricks SQL alerts are not set up in the SQL Warehouse tab. They are typically configured within the query editor or a dedicated alerts section of the Databricks SQL interface.

SQL Commands:

- Use `CREATE TABLE AS SELECT DISTINCT * FROM existing_table;` to create a new table with unique values.
- Use `CREATE OR REPLACE VIEW` to ensure a view reflects the latest changes from underlying tables.
- Use `ALTER TABLE` to modify the structure of an existing table.
- Use `DROP TABLE database_name.table_name;` to delete a table.
- Use `INSERT INTO` to add data from one table to another, including duplicates.
- Use `CACHE TABLE` to speed up query execution by storing data in memory.

Databricks SQL Features & Services:

- Databricks SQL is used for SQL-based analytics and ad-hoc queries.
- Delta Lake enables quick SQL queries and visualizations with serverless compute.
- Partner Connect automates the setup of clusters and connections for integrating with external tools like Fivetran.
- Data Explorer allows viewing metadata and data, and managing permissions.
- The SQL Editor page is used to write and execute SQL queries.

Medallion Architecture:

- The Bronze layer is for storing raw data.
- Silver Layer is the heavy load, meant for data cleaning and preparation.
- The Gold layer is commonly used by data analysts for curated / aggregated and transformed data.

Visualizations:

- Use heatmaps to visualize relationships and correlations between two variables.
- Use cumulative line charts to visualize the cumulative total of a metric over time.

- Use Sankey diagrams to show relationships between hierarchical categories and flow.
- Use markdown-based text boxes to designate sections with formatted text on dashboards.
- Histograms are best for showing the distribution of a single numeric variable.
- Bar charts are best for comparing categorical data against a numerical value and are the default visualization type for region and customer number data.
- Pie charts are appropriate for visualizing the composition of different categories as parts of a whole.
- Parameterized queries allow visualizations to update dynamically based on user input.
- Dashboard Parameters applied to a query affect all visualizations using the same parameter.
- Multiple visualizations can be added to a dashboard based on the same query.

Delta Lake:

- Delta Lake provides ACID transactions for data consistency and reliability.
- Delta Lake transaction logs enable version control and audit trails.
- Delta Lake stores table metadata alongside data files.

Table Types:

- Managed tables ensure data persistence and reuse across sessions.
- Avoid external tables if you need all data and metadata removed when dropped.
- Dropping an external table only removes metadata, not the underlying data files.
- Temporary views store query results for the current session only and are dropped upon logout.

Window Functions:

- Utilize window functions for partitioned aggregations to optimize performance on large datasets.
- `PERCENT_RANK()` calculates the percentage rank of rows within a partition.
- `SUM()` is used for calculating a running total.
- `RANK()` assigns a unique rank to each row within a partition based on ordering.
- Window functions allow aggregation while preserving individual rows.

Functions:

- Higher-order functions (map, filter) are used for applying custom logic at scale to complex data types like arrays.
- `nvl()` replaces null values with a specified replacement value.
- `explode()` transforms elements from an array or map into separate rows.
- `APPROX_MEDIAN()` calculates the approximate median value of a column.
- `lag()` accesses data from a previous row in the same result set.
- `collect_list()` returns a list of values from a column in a group.
- `first()` returns the first value in a group.

Joins:

- Use the `JOIN` clause with `ON` condition to combine rows from tables based on related columns.
- A `LEFT JOIN` returns all rows from the left table and matched rows from the right table.
- A `LEFT SEMI JOIN` returns only rows from the left table that have matching rows in the right table.

Alerts and Scheduling:

- Alerts in Databricks SQL work only with queries that have static results and do not support query parameters.
- Scheduled queries can be set up in the Query Editor, ranging from 1 minute to 2 weeks, and require appropriate permissions.

- Job scheduling with CRON expressions allows setting up queries to run at specific times.
- Alerts are not set up in the SQL Warehouse tab.

Data Lakehouse:

- A Data Lakehouse combines the flexibility of data lakes with the management capabilities of data warehouses.

Query Performance:

- Monitoring the impact on performance and costs is crucial for automatically refreshing dashboards with live data.
- Turning off the Auto stop feature for SQL endpoints can reduce start-up time.
- Near real-time dashboard updates based on streaming data can be costly due to compute resources.
- Inspect the Full Duration in the Query History tab to confirm if query results came from the cache.

Descriptive Statistics:

- Descriptive statistics uses summary statistics to quantitatively describe and summarize data.
- Mean and median values differ significantly when a variable contains extreme outliers.

ANSI SQL:

- Using ANSI SQL as the standard dialect in Databricks SQL makes it easier to migrate existing SQL queries.

Working with PII:

- Consider data minimization, purpose limitation, and security when working with personally identifiable information (PII).

Data Engineering Roles:

- Data engineers primarily use other Databricks services besides Databricks SQL.