Product Questions: 100
Version: 6.1

Question: 1

A data organization leader is upset about the data analysis team's reports being different from the data engineering team's reports. The leader believes the siloed nature of their organization's data engineering and data analysis architectures is to blame.
Which of the following describes how a data lakehouse could alleviate this issue?

A. Both teams would autoscale their work as data size evolves
B. Both teams would use the same source of truth for their work
C. Both teams would reorganize to report to the same department
D. Both teams would be able to collaborate on projects in real-time
E. Both teams would respond more quickly to ad-hoc requests

Answer: B

Explanation:

A data lakehouse is a data management architecture that combines the flexibility, cost-efficiency, and scale of data lakes with the data management and ACID transactions of data warehouses, enabling business intelligence (BI) and machine learning (ML) on all data12. By using a data lakehouse, both the data analysis and data engineering teams can access the same data sources and formats, ensuring data consistency and quality across their reports. A data lakehouse also supports schema enforcement and evolution, data validation, and time travel to old table versions, which can help resolve data conflicts and errors1. Reference: 1: What is a Data Lakehouse? - Databricks 2: What is a data lakehouse? | IBM

Question: 2

Which of the following describes a scenario in which a data team will want to utilize cluster pools?

A. An automated report needs to be refreshed as quickly as possible.
B. An automated report needs to be made reproducible.
C. An automated report needs to be tested to identify errors.
D. An automated report needs to be version-controlled across multiple collaborators.
E. An automated report needs to be runnable by all stakeholders.

Answer: A

Explanation:

Databricks cluster pools are a set of idle, ready-to-use instances that can reduce cluster start and auto-scaling times. This is useful for scenarios where a data team needs to run an automated report as quickly as possible, without waiting for the cluster to launch or scale up. Cluster pools can also help save costs by reusing idle instances across different clusters and avoiding DBU charges for idle instances in the pool. Reference: Best practices: pools | Databricks on AWS, Best practices: pools - Azure Databricks | Microsoft Learn, Best practices: pools | Databricks on Google Cloud

Question: 3

Which of the following is hosted completely in the control plane of the classic Databricks architecture?

A. Worker node
B. JDBC data source
C. Databricks web application
D. Databricks Filesystem
E. Driver node

Answer: C

Explanation:

: The Databricks web application is the user interface that allows you to create and manage workspaces, clusters, notebooks, jobs, and other resources. It is hosted completely in the control plane of the classic Databricks architecture, which includes the backend services that Databricks manages in your Databricks account. The other options are part of the compute plane, which is where your data is processed by compute resources such as clusters. The compute plane is in your own cloud account and network. Reference: Databricks architecture overview, Security and Trust Center

Question: 4

Which of the following benefits of using the Databricks Lakehouse Platform is provided by Delta Lake?

A. The ability to manipulate the same data using a variety of languages
B. The ability to collaborate in real time on a single notebook
C. The ability to set up alerts for query failures
D. The ability to support batch and streaming workloads
E. The ability to distribute complex data operations

Answer: D

Explanation:

Delta Lake is the optimized storage layer that provides the foundation for storing data and tables in the Databricks lakehouse. Delta Lake is fully compatible with Apache Spark APIs, and was developed for tight integration with Structured Streaming, allowing you to easily use a single copy of data for

both batch and streaming operations and providing incremental processing at scale1. Delta Lake supports upserts using the merge operation, which enables you to efficiently update existing data or insert new data into your Delta tables2. Delta Lake also provides time travel capabilities, which allow you to query previous versions of your data or roll back to a specific point in time3. Reference: 1: What is Delta Lake? | Databricks on AWS 2: Upsert into a table using merge | Databricks on AWS 3: [Query an older snapshot of a table (time travel) | Databricks on AWS]
Learn more
1learn.microsoft.com2medium.com3slideshare.net4docs.databricks.com5github.com6key2consulting.com

Question: 5

Which of the following describes the storage organization of a Delta table?

A. Delta tables are stored in a single file that contains data, history, metadata, and other attributes.
B. Delta tables store their data in a single file and all metadata in a collection of files in a separate location.
C. Delta tables are stored in a collection of files that contain data, history, metadata, and other attributes.
D. Delta tables are stored in a collection of files that contain only the data stored within the table.
E. Delta tables are stored in a single file that contains only the data stored within the table.

Answer: C

Explanation:

Delta Lake is the optimized storage layer that provides the foundation for storing data and tables in the Databricks lakehouse. Delta Lake is open source software that extends Parquet data files with a file-based transaction log for ACID transactions and scalable metadata handling1. Delta Lake stores its data and metadata in a collection of files in a directory on a cloud storage system, such as AWS S3 or Azure Data Lake Storage2. Each Delta table has a transaction log that records the history of operations performed on the table, such as insert, update, delete, merge, etc. The transaction log also stores the schema and partitioning information of the table2. The transaction log enables Delta Lake to provide ACID guarantees, time travel, schema enforcement, and other features1. Reference:
What is Delta Lake? | Databricks on AWS
Quickstart — Delta Lake Documentation

Question: 6

Which of the following code blocks will remove the rows where the value in column age is greater than 25 from the existing Delta table my_table and save the updated table?

A. SELECT * FROM my_table WHERE age > 25;
B. UPDATE my_table WHERE age > 25;
C. DELETE FROM my_table WHERE age > 25;
D. UPDATE my_table WHERE age <= 25;
E. DELETE FROM my_table WHERE age <= 25;

Answer: C

Explanation:

: The DELETE command in Delta Lake allows you to remove data that matches a predicate from a Delta table. This command will delete all the rows where the value in the column age is greater than 25 from the existing Delta table my_table and save the updated table. The other options are either incorrect or do not achieve the desired result. Option A will only select the rows that match the predicate, but not delete them. Option B will update the rows that match the predicate, but not delete them. Option D will update the rows that do not match the predicate, but not delete them. Option E will delete the rows that do not match the predicate, which is the opposite of what we want. Reference: Table deletes, updates, and merges — Delta Lake Documentation

Question: 7

A data engineer has realized that they made a mistake when making a daily update to a table. They need to use Delta time travel to restore the table to a version that is 3 days old. However, when the data engineer attempts to time travel to the older version, they are unable to restore the data because the data files have been deleted.
Which of the following explains why the data files are no longer present?

A. The VACUUM command was run on the table
B. The TIME TRAVEL command was run on the table
C. The DELETE HISTORY command was run on the table
D. The OPTIMIZE command was nun on the table
E. The HISTORY command was run on the table

Answer: A

Explanation:

The VACUUM command is used to remove files that are no longer referenced by a Delta table and are older than the retention threshold1. The default retention period is 7 days2, but it can be changed by setting
the delta.logRetentionDuration and delta.deletedFileRetentionDuration configurations3. If the VACUUM command was run on the table with a retention period shorter than 3 days, then the data files that were needed to restore the table to a 3-day-old version would have been deleted. The other commands do not delete data files from the table. The TIME TRAVEL command is used to query a historical version of the table4. The DELETE HISTORY command is not a valid command in Delta Lake. The OPTIMIZE command is used to improve the performance of the table by compacting small files into larger ones5. The HISTORY command is used to retrieve information about the operations performed on the table. Reference: 1: VACUUM | Databricks on AWS 2: Work with Delta Lake table history | Databricks on AWS 3: [Delta Lake configuration | Databricks on AWS] 4: Work with Delta Lake table history - Azure Databricks 5: [OPTIMIZE | Databricks on AWS] : [HISTORY | Databricks on AWS]

Question: 8

Which of the following Git operations must be performed outside of Databricks Repos?

A. Commit
B. Pull
C. Push
D. Clone
E. Merge

Answer: E

Explanation:

Databricks Repos is a visual Git client and API in Databricks that supports common Git operations such as commit, pull, push, branch management, and visual comparison of diffs when committing1. However, merge is not supported in the Git dialog2. You need to use the Repos UI or your Git provider to merge branches3. Merge is a way to combine the commit history from one branch into another branch1. During a merge, a merge conflict is encountered when Git cannot automatically combine code from one branch into another. Merge conflicts require manual resolution before a merge can be completed1. Reference: 4: Run Git operations on Databricks Repos4, 1: CI/CD techniques with Git and Databricks Repos1, 3: Collaborate in Repos3, 2: Databricks Repos - What it is and how we can use it2.
Databricks Repos is a visual Git client and API in Databricks that supports common Git operations such as commit, pull, push, merge, and branch management. However, to clone a remote Git repository to a Databricks repo, you must use the Databricks UI or API. You cannot clone a Git repo using the CLI through a cluster's web terminal, as the files won't display in the Databricks UI1. Reference: 1: Run Git operations on Databricks Repos | Databricks on AWS2

Question: 9

Which of the following data lakehouse features results in improved data quality over a traditional data lake?

A. A data lakehouse provides storage solutions for structured and unstructured data.
B. A data lakehouse supports ACID-compliant transactions.
C. A data lakehouse allows the use of SQL queries to examine data.
D. A data lakehouse stores data in open formats.
E. A data lakehouse enables machine learning and artificial Intelligence workloads.

Answer: B

Explanation:

: A data lakehouse is a data management architecture that combines the flexibility, cost-efficiency, and scale of data lakes with the data management and ACID transactions of data warehouses, enabling business intelligence (BI) and machine learning (ML) on all data12. One of the key features of a data lakehouse is that it supports ACID-compliant transactions, which means that it ensures data integrity, consistency, and isolation across concurrent read and write operations3. This feature results in improved data quality over a traditional data lake, which does not support transactions and may suffer from data corruption, duplication, or inconsistency due to concurrent or streaming data ingestion and processing . Reference: 1: What is a Data Lakehouse? - Databricks 2: What is a Data Lakehouse? Definition, features & benefits. - Qlik 3: ACID Transactions - Databricks : [Data Lake vs Data Warehouse: Key Differences] : [Data Lakehouse: The Future of Data Engineering]

Question: 10

A data engineer needs to determine whether to use the built-in Databricks Notebooks versioning or version their project using Databricks Repos.
Which of the following is an advantage of using Databricks Repos over the Databricks Notebooks versioning?

A. Databricks Repos automatically saves development progress
B. Databricks Repos supports the use of multiple branches
C. Databricks Repos allows users to revert to previous versions of a notebook
D. Databricks Repos provides the ability to comment on specific changes
E. Databricks Repos is wholly housed within the Databricks Lakehouse Platform

Answer: B

Explanation:

Databricks Repos is a visual Git client and API in Databricks that supports common Git operations such as cloning, committing, pushing, pulling, and branch management. Databricks Notebooks versioning is a legacy feature that allows users to link notebooks to GitHub repositories and perform basic Git operations. However, Databricks Notebooks versioning does not support the use of multiple branches for development work, which is an advantage of using Databricks Repos. With Databricks Repos, users can create and manage branches for different features, experiments, or bug fixes, and merge, rebase, or resolve conflicts between them. Databricks recommends using a separate branch for each notebook and following data science and engineering code development best practices using Git for version control, collaboration, and CI/CD. Reference: Git integration with Databricks Repos - Azure Databricks | Microsoft Learn, Git version control for notebooks (legacy) | Databricks on AWS, Databricks Repos Is Now Generally Available - New 'Files' Feature in ..., Databricks Repos - What it is and how we can use it | Adatis.

Question: 11

A data engineer has left the organization. The data team needs to transfer ownership of the data engineer's Delta tables to a new data engineer. The new data engineer is the lead engineer on the data team.
Assuming the original data engineer no longer has access, which of the following individuals must be the one to transfer ownership of the Delta tables in Data Explorer?

A. Databricks account representative
B. This transfer is not possible
C. Workspace administrator
D. New lead data engineer
E. Original data engineer

Answer: C

Explanation:

The workspace administrator is the only individual who can transfer ownership of the Delta tables in Data Explorer, assuming the original data engineer no longer has access. The workspace administrator has the highest level of permissions in the workspace and can manage all resources, users, and groups. The other options are either not possible or not sufficient to perform the ownership transfer. The Databricks account representative is not involved in the workspace management. The transfer is possible and not dependent on the original data engineer. The new lead data engineer may not have the necessary permissions to access or modify the Delta tables, unless granted by the workspace administrator or the original data engineer before leaving. Reference: Workspace access control, Manage Unity Catalog object ownership.

Question: 12

A data analyst has created a Delta table sales that is used by the entire data analysis team. They want help from the data engineering team to implement a series of tests to ensure the data is clean. However, the data engineering team uses Python for its tests rather than SQL.
Which of the following commands could the data engineering team use to access sales in PySpark?

A. SELECT * FROM sales
B. There is no way to share data between PySpark and SQL.
C. spark.sql("sales")
D. spark.delta.table("sales")
E. spark.table("sales")

Answer: E

Explanation:

The data engineering team can use the spark.table method to access the Delta table sales in PySpark. This method returns a DataFrame representation of the Delta table, which can be used for further processing or testing. The spark.table method works for any table that is registered in the Hive metastore or the Spark catalog, regardless of the file format1. Alternatively, the data engineering team can also use the DeltaTable.forPath method to load the Delta table from its path2. Reference: 1: SparkSession | PySpark 3.2.0 documentation 2: Welcome to Delta Lake's Python documentation page — delta-spark 2.4.0 documentation

Question: 13

Which of the following commands will return the location of database customer360?

A. DESCRIBE LOCATION customer360;
B. DROP DATABASE customer360;
C. DESCRIBE DATABASE customer360;
D. ALTER DATABASE customer360 SET DBPROPERTIES ('location' = '/user'};
E. USE DATABASE customer360;

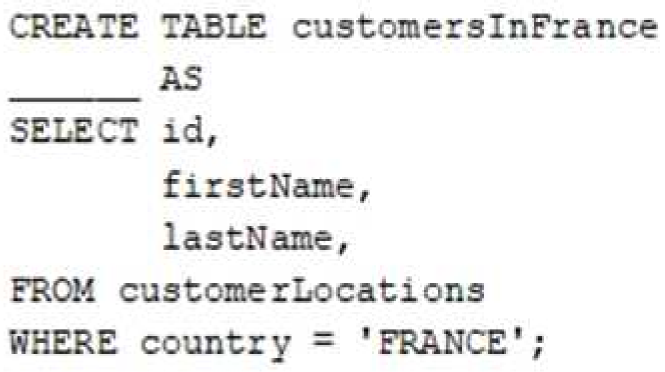
Answer: C

Explanation:

The command DESCRIBE DATABASE customer360; will return the location of the database customer360, along with its comment and properties. This command is an alias for DESCRIBE SCHEMA customer360;, which can also be used to get the same information. The other commands will either drop the database, alter its properties, or use it as the current database, but will not return its location12. Reference:
DESCRIBE DATABASE | Databricks on AWS
DESCRIBE DATABASE - Azure Databricks - Databricks SQL

Question: 14

A data engineer wants to create a new table containing the names of customers that live in France. They have written the following command:



```
CREATE TABLE customersInFrance
______ AS
SELECT id,
       firstName,
       lastName,
FROM customerLocations
WHERE country = 'FRANCE';
```

A senior data engineer mentions that it is organization policy to include a table property indicating that the new table includes personally identifiable information (PII).
Which of the following lines of code fills in the above blank to successfully complete the task?

A. There is no way to indicate whether a table contains PII.
B. "COMMENT PII"
C. TBLPROPERTIES PII
D. COMMENT "Contains PII"
E. PII

Answer: D

Explanation:

In Databricks, when creating a table, you can add a comment to columns or the entire table to provide more information about the data it contains. In this case, since it's organization policy to indicate that the new table includes personally identifiable information (PII), option D is correct. The line of code would be added after defining the table structure and before closing with a semicolon. Reference: Data Engineer Associate Exam Guide, CREATE TABLE USING (Databricks SQL)

Question: 15

Which of the following benefits is provided by the array functions from Spark SQL?

A. An ability to work with data in a variety of types at once
B. An ability to work with data within certain partitions and windows
C. An ability to work with time-related data in specified intervals
D. An ability to work with complex, nested data ingested from JSON files
E. An ability to work with an array of tables for procedural automation

Answer: D

Explanation:

The array functions from Spark SQL are a subset of the collection functions that operate on array columns1. They provide an ability to work with complex, nested data ingested from JSON files or other sources2. For example, the explode function can be used to transform an array column into multiple rows, one for each element in the array3. The array_contains function can be used to check if a value is present in an array column4. The array_join function can be used to concatenate all elements of an array column with a delimiter. These functions can be useful for processing JSON data that may have nested arrays or objects. Reference: 1: Spark SQL, Built-in Functions - Apache Spark 2: Spark SQL Array Functions Complete List - Spark By Examples 3: Spark SQL Array Functions - Syntax and Examples - DWgeek.com 4: Spark SQL, Built-in Functions - Apache Spark : Spark SQL, Built-in Functions - Apache Spark : [Working with Nested Data Using Higher Order Functions in SQL on Databricks - The Databricks Blog]

Question: 16

Which of the following commands can be used to write data into a Delta table while avoiding the writing of duplicate records?

A. DROP
B. IGNORE
C. MERGE
D. APPEND
E. INSERT

Answer: C

Explanation:

The MERGE command can be used to upsert data from a source table, view, or DataFrame into a target Delta table. It allows you to specify conditions for matching and updating existing records, and inserting new records when no match is found. This way, you can avoid writing duplicate records into a Delta table1. The other commands (DROP, IGNORE, APPEND, INSERT) do not have this functionality and may result in duplicate records or data loss234. Reference: 1: Upsert into a Delta Lake table using merge | Databricks on AWS 2: SQL DELETE | Databricks on AWS 3: SQL INSERT INTO | Databricks on AWS 4: SQL UPDATE | Databricks on AWS

Question: 17