

Top of Form

Question 1

What is the best way to describe a data lakehouse compared to a data warehouse?

- ☐ A data lakehouse provides a relational system of data management
- ☐ A data lakehouse captures snapshots of data for version control purposes.
- ☐ A data lakehouse couples storage and compute for complete control.
- ☐ A data lakehouse utilizes proprietary storage formats for data.
- ☐ A data lakehouse enables both batch and streaming analytics.

Explanation

Answer is A data lakehouse enables both batch and streaming analytics.

A lakehouse has the following key features:

Transaction support: In an enterprise lakehouse many data pipelines will often be reading and writing data concurrently. Support for ACID transactions ensures consistency as multiple parties concurrently read or write data, typically using SQL.

Schema enforcement and governance: The Lakehouse should have a way to support schema enforcement and evolution, supporting DW schema architectures such as star/snowflake-schemas. The system should be able to [reason about data integrity](#), and it should have robust governance and auditing mechanisms.

BI support: Lakehouses enable using BI tools directly on the source data. This reduces staleness and improves recency, reduces latency, and lowers the cost of having to operationalize two copies of the data in both a data lake and a warehouse.

Storage is decoupled from compute: In practice this means storage and compute use separate clusters, thus these systems are able to scale to many more concurrent users and larger data sizes. Some modern data warehouses also have this property.

Openness: The storage formats they use are open and standardized, such as Parquet, and they provide an API so a variety of tools and engines, including machine learning and Python/R libraries, can efficiently access the data directly.

Support for diverse data types ranging from unstructured to structured data: The lakehouse can be used to store, refine, analyze, and access data types needed for many new data applications, including images, video, audio, semi-structured data, and text.

Support for diverse workloads: including data science, machine learning, and SQL and analytics. Multiple tools might be needed to support all these workloads but they all rely on the same data repository.

End-to-end streaming: Real-time reports are the norm in many enterprises. Support for streaming eliminates the need for separate systems dedicated to serving real-time data applications.

Bottom of Form

Top of Form

Question 2

You are designing an analytical to store structured data from your e-commerce platform and unstructured data from website traffic and app store, how would you approach where you store this data?



Use traditional data warehouse for structured data and use data lakehouse for unstructured data.



Data lakehouse can only store unstructured data but cannot enforce a schema



Data lakehouse can store structured and unstructured data and can enforce schema



Traditional data warehouses are good for storing structured data and enforcing schema

Explanation

The answer is, Data lakehouse can store structured and unstructured data and can enforce schema

[What Is a Lakehouse? - The Databricks Blog](#)

A lakehouse has the following key features:

- **Transaction support:** In an enterprise lakehouse many data pipelines will often be reading and writing data concurrently. Support for ACID transactions ensures consistency as multiple parties concurrently read or write data, typically using SQL.
- **Schema enforcement and governance:** The Lakehouse should have a way to support schema enforcement and evolution, supporting DW schema architectures such as star/snowflake-schemas. The system should be able to **reason about data integrity**, and it should have robust governance and auditing mechanisms.
- **BI support:** Lakehouses enable using BI tools directly on the source data. This reduces staleness and improves recency, reduces latency, and lowers the cost of having to operationalize two copies of the data in both a data lake and a warehouse.
- **Storage is decoupled from compute:** In practice this means storage and compute use separate clusters, thus these systems are able to scale to many more concurrent users and larger data sizes. Some modern data warehouses also have this property.
- **Openness:** The storage formats they use are open and standardized, such as Parquet, and they provide an API so a variety of tools and engines, including machine learning and Python/R libraries, can efficiently access the data **directly**.
- **Support for diverse data types ranging from unstructured to structured data:** The lakehouse can be used to store, refine, analyze, and access data types needed for many new data applications, including images, video, audio, semi-structured data, and text.
- **Support for diverse workloads:** including data science, machine learning, and SQL and analytics. Multiple tools might be needed to support all these workloads but they all rely on the same data repository.
- **End-to-end streaming:** Real-time reports are the norm in many enterprises. Support for streaming eliminates the need for separate systems dedicated to serving real-time data applications.

Bottom of Form

Top of Form

Question 3

You are currently working on a production job failure with a job set up in job clusters due to a data issue, what cluster do you need to start to investigate and analyze the data?

- ☐ A Job cluster can be used to analyze the problem
- ☐ All-purpose cluster/interactive cluster is the recommended way to run commands and view the data.
- ☐ Existing job cluster can be used to investigate the issue
- ☐ Databricks SQL Endpoint can be used to investigate the issue

Explanation

Answer is All-purpose cluster/ interactive cluster is the recommended way to run commands and view the data.

A job cluster can not provide a way for a user to interact with a notebook once the job is submitted, but an Interactive cluster allows to you display data, view visualizations write or edit queries, which makes it a perfect fit to investigate and analyze the data.

Bottom of Form

Top of Form

Question 4

Which of the following describes how Databricks Repos can help facilitate CI/CD workflows on the Databricks Lakehouse Platform?

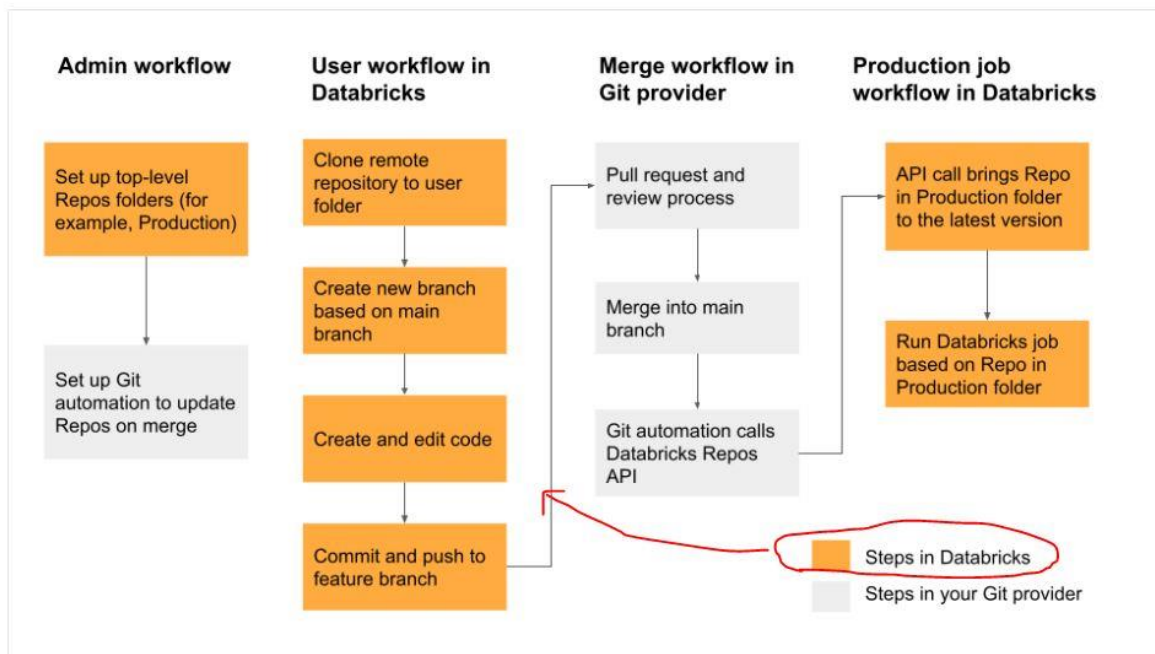
- ☐ Databricks Repos can facilitate the pull request, review, and approval process before merging branches
- ☐ Databricks Repos can merge changes from a secondary Git branch into a main Git branch
- ☐ Databricks Repos can be used to design, develop, and trigger Git automation pipelines
- ☐ Databricks Repos can store the single-source-of-truth Git repository
- ☐ Databricks Repos can commit or push code changes to trigger a CI/CD process

Explanation

Answer is Databricks Repos can commit or push code changes to trigger a CI/CD process

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

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



Bottom of Form

Top of Form

Question 5

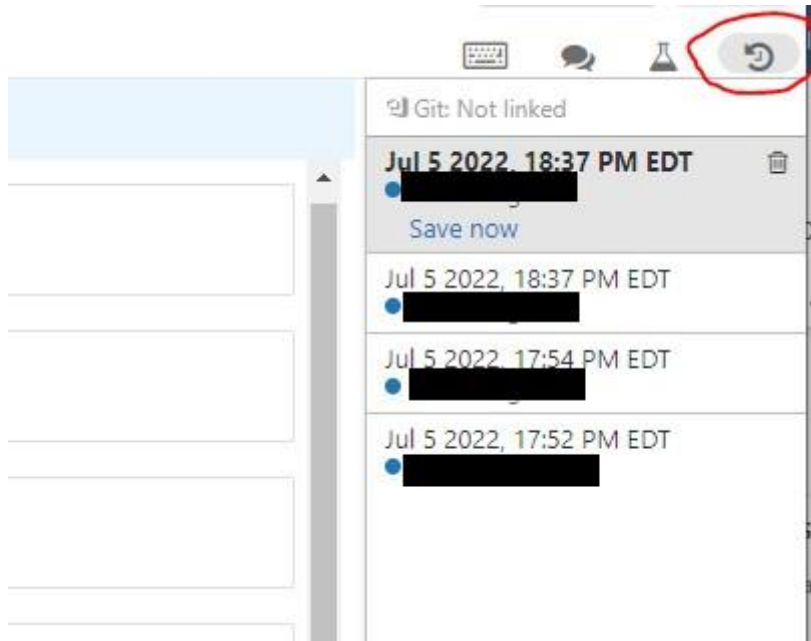
You noticed that colleague is manually copying the notebook with _bkp to store the previous versions, which of the following feature would you recommend instead.

- ☐ Databricks notebooks support change tracking and versioning
- ☐ Databricks notebooks should be copied to a local machine and setup source control locally to version the notebooks
- ☐ Databricks notebooks can be exported into dbc archive files and stored in data lake
- ☐ Databricks notebook can be exported as HTML and imported at a later time

Explanation

Answer is Databricks notebooks support automatic change tracking and versioning.

When you are editing the notebook on the right side check version history to view all the changes, every change you are making is captured and saved.



Bottom of Form

Top of Form

Question 6

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

- ☐ Admin console
- ☐ User settings
- ☐ Data explorer
- ☐ Azure Databricks control pane IAM
- ☐ Azure RBAC

Explanation

Answer is Data Explorer

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

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

To open data explorer, click  Data in the sidebar.

Bottom of Form

Top of Form

Question 7

How does a Delta Lake differ from a traditional data lake?

- ☐ Delta lake is Datawarehouse service on top of data lake that can provide reliability, security, and performance
- ☐ Delta lake is a caching layer on top of data lake that can provide reliability, security, and performance
- ☐ Delta lake is an open storage format like parquet with additional capabilities that can provide reliability, security, and performance
- ☐ Delta lake is an open storage format designed to replace flat files with additional capabilities that can provide reliability, security, and performance
- ☐ Delta lake is proprietary software designed by Databricks that can provide reliability, security, and performance

Explanation

Answer is, Delta lake is an open storage format like parquet with additional capabilities that can provide reliability, security, and performance

Delta lake is

- Open source
- Builds up on standard data format
- Optimized for cloud object storage
- Built for scalable metadata handling

Delta lake is not

- Proprietary technology
- Storage format
- Storage medium
- Database service or data warehouse

Bottom of Form

Top of Form

Question 8

As a Data Engineer, you were asked to create a delta table to store below transaction data?

transactionId	transactionDate	unitsSold
1	01-01-2021 09:10:24 AM	100
2	01-01-2022 10:30:30 AM	10

- ☐ CREATE DELTA TABLE transactions (
 transactionId int,
 transactionDate timestamp,
 unitsSold int)
- ☐ CREATE TABLE transactions (
transactionId int,
transactionDate timestamp,
unitsSold int)
FORMAT DELTA
- ☐ CREATE TABLE transactions (
transactionId int,
transactionDate timestamp,
unitsSold int)
- ☐ CREATE TABLE USING DELTA transactions (
transactionId int,
transactionDate timestamp,
unitsSold int)
- ☐ CREATE TABLE transactions (
transactionId int,
transactionDate timestamp,
unitsSold int)
LOCATION DELTA

Explanation

Answer is

```
CREATE TABLE transactions (  
  transactionId int,  
  transactionDate timestamp,  
  unitsSold int)
```

When creating a table in Databricks by default the table is stored in DELTA format.

Bottom of Form

Top of Form

Question 9

Which of the following is a correct statement on how the data is organized in the storage when managing a DELTA table?

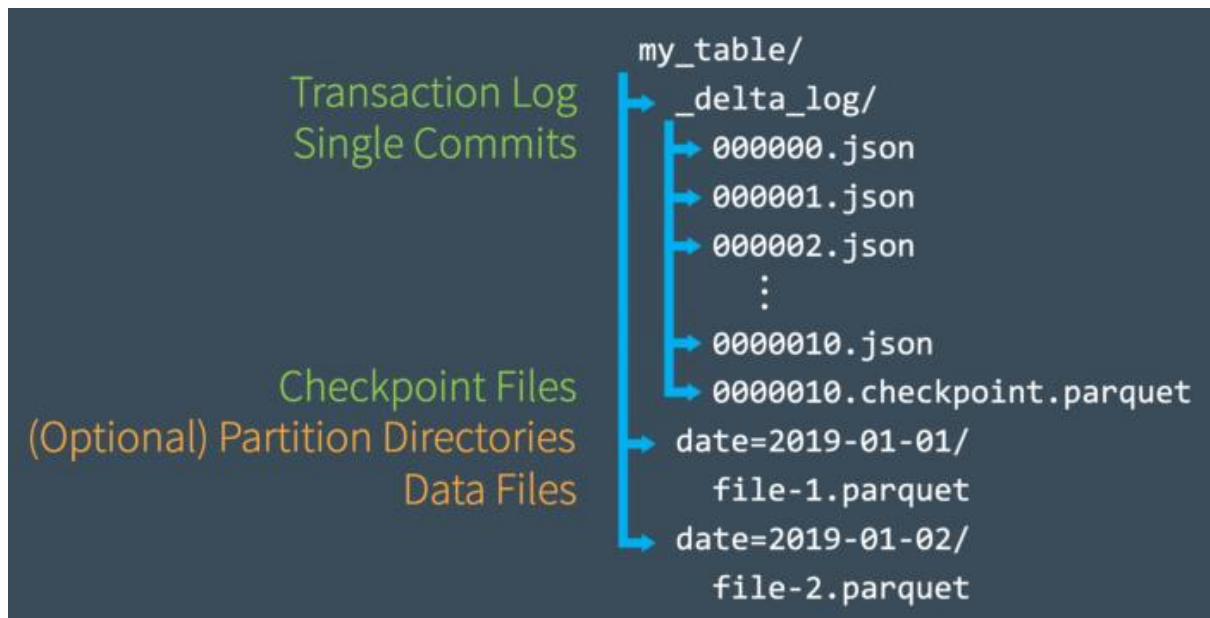
- ☐ All of the data is broken down into one or many parquet files, log files are broken down into one or many JSON files, and each transaction creates a new data file(s) and log file.
- ☐ All of the data and log are stored in a single parquet file
- ☐ All of the data is broken down into one or many parquet files, but the log file is stored as a single json file, and every transaction creates a new data file(s) and log file gets appended.
- ☐ All of the data is broken down into one or many parquet files, log file is removed once the transaction is committed.
- ☐ All of the data is stored into one parquet file, log files are broken down into one or many json files.

Explanation

Answer is

All of the data is broken down into one or many parquet files, log files are broken down into one or many json files, and each transaction creates a new data file(s) and log file.

here is sample layout of how DELTA table might look,



Bottom of Form

Top of Form

Question 10

You are still noticing slowness in query after performing optimize which helped you to resolve the small files problem, the column(transactionId) you are using to filter the data has high cardinality and auto incrementing number. Which delta optimization can you enable to filter data effectively based on this column?

- ☐ Create BLOOM FILTER index on the transactionId
- ☐ Perform Optimize with Zorder on transactionId
- ☐ transactionId has high cardinality, you cannot enable any optimization.
- ☐ Increase the cluster size and enable delta optimization
- ☐ Increase the driver size and enable delta optimization

Explanation

The answer is, perform Optimize with Z-order by transactionid

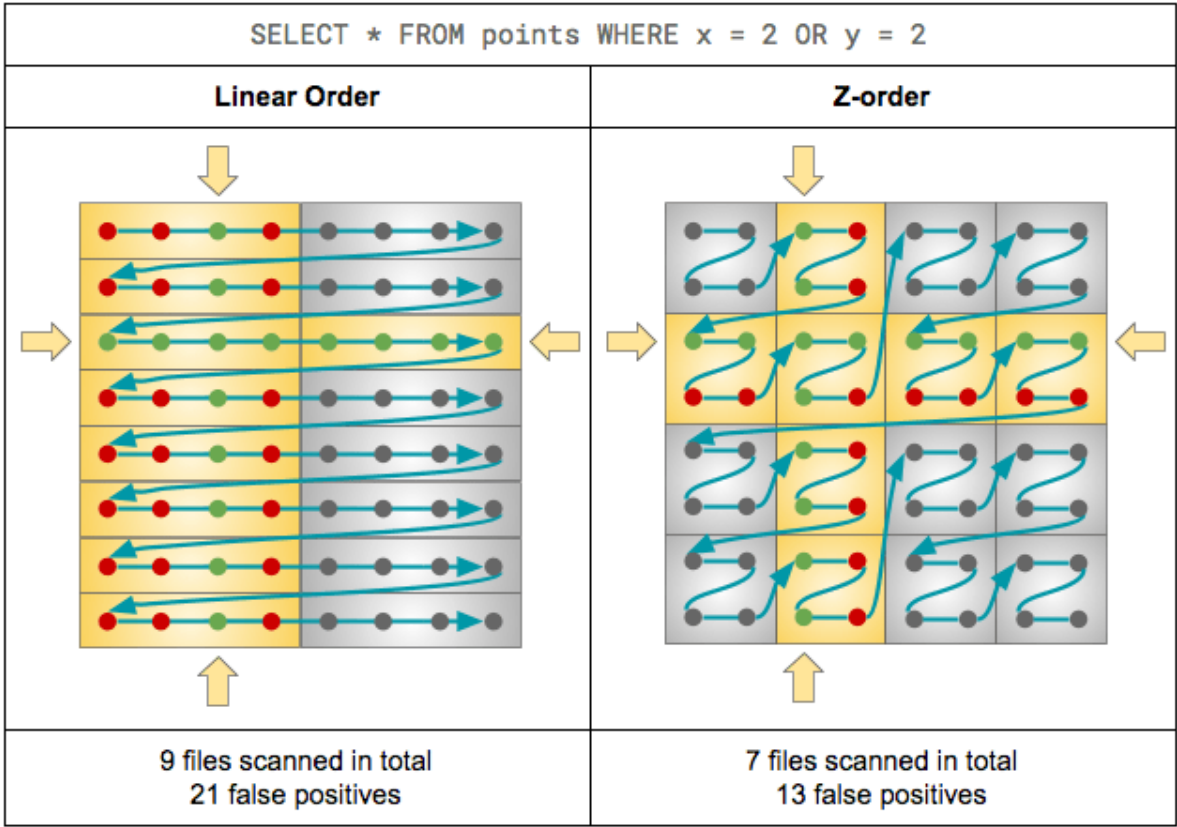
Here is a simple

Explanation of how Z-order works, once the data is naturally ordered, when a file is scanned it only brings the data it needs into spark's memory

Based on the column min and max it knows which data files needs to be scanned.

```
SELECT input_file_name() as "file_name",
       min(col) AS "col_min",
       max(col) AS "col_max"
FROM table
GROUP BY input_file_name()
```

file_name	col_min	col_max
data_file_1	6	8
data_file_2	3	10
data_file_3	1	4



Bottom of Form

Top of Form

Question 11

If you create a database sample_db with the statement CREATE DATABASE sample_db what will be the location of the database in DBFS?

- ☐ Default location, DBFS:/user/
- ☐ The location assigned to parameter warehouse.dir
- ☐ Storage account
- ☐ Statement fails "Unable to create database without location"
- ☐ Default Location, DBFS:/user/databases/

Explanation

Answer is The location assigned to parameter warehouse.dir

Bottom of Form

Top of Form

Question 12

Which of the following results in the creation of an external table?

- ☐ CREATE TABLE transactions (id int, desc string) USING DELTA LOCATION EXTERNAL
- ☐ CREATE TABLE transactions (id int, desc string)
- ☐ CREATE EXTERNAL TABLE transactions (id int, desc string)
- ☐ CREATE TABLE transactions (id int, desc string) TYPE EXTERNAL
- ☐ CREATE TABLE transactions (id int, desc string) LOCATION '/mnt/delta/transactions'

Explanation

Answer is `CREATE TABLE transactions (id int, desc string) USING DELTA LOCATION '/mnt/delta/transactions'`

Anytime a table is created using Location it is considered an external table, below is the current syntax.

Syntax

`CREATE TABLE table_name (column column_data_type...) USING format LOCATION "dbfs:/"`

Bottom of Form

Top of Form

Question 13

When you drop an external DELTA table using `DROP TABLE table_name`, how does it impact metadata, data stored in the storage?

- ☐ Drops table from metastore, metadata and data in storage
- ☐ Drops table from metastore, data but keeps metadata in storage
- ☐ Drops table from metastore, metadata but keeps the data in storage
- ☐ Drops table from metastore, but keeps metadata and data in storage
- ☐ Drops table from metastore and data in storage but keeps meta data

Explanation

The answer is Drops table from metastore, but keeps metadata and data in storage.

When an external table is dropped, only the table definition is dropped from metastore everything including data and metadata remains in the storage.

Bottom of Form

Top of Form

Question 14

Which of the following is a true statement about the global temporary view?

- ☐ A global temporary view is available only on the cluster it was created, when the cluster restarts global temporary view is automatically dropped.
- ☐ A global temporary view is available on all clusters for a given workspace
- ☐ A global temporary view persists even if the cluster is restarted
- ☐ A global temporary view is stored in a user database
- ☐ A global temporary view is automatically dropped after 7 days

Explanation

Answer is, A global temporary view is available only on the cluster it was created.

Two types of temporary views can be created Local and Global

A local temporary view is only available with a spark session, so another notebook in the same cluster can not access it. if a notebook is detached and re attached local temporary view is lost.

A global temporary view is available to all the notebooks in the cluster, if a cluster restarts global temporary view is lost.

Bottom of Form

Top of Form

Question 15

You are trying to create an object by joining two tables that and it is accessible to data scientist's team, so it does not get dropped if the cluster restarts or if the notebook is detached. What type of object are you trying to create?

- ☐ Temporary view
- ☐ Global Temporary view
- ☐ Global Temporary view with cache option
- ☐ External view
- ☐ View

Explanation

Answer is View, A view can be used to join multiple tables but also persist into meta stores so others can access it

Bottom of Form

Top of Form

Question 16

What is the best way to query external csv files located on DBFS Storage to inspect the data using SQL?

- ☐ `SELECT * FROM 'dbfs:/location/csv_files/' FORMAT = 'CSV'`
- ☐ `SELECT CSV. * from 'dbfs:/location/csv_files/'`
- ☐ `SELECT * FROM CSV. 'dbfs:/location/csv_files/'`
- ☐ You can not query external files directly, us COPY INTO to load the data into a table first
- ☐ `SELECT * FROM 'dbfs:/location/csv_files/' USING CSV`

Explanation

Answer is, `SELECT * FROM CSV. 'dbfs:/location/csv_files/'`

you can query external files stored on the storage using below syntax

`SELECT * FROM format. `/Location``

format - CSV, JSON, PARQUET, TEXT

Bottom of Form

Top of Form

Question 17

Direct query on external files limited options, create external tables for CSV files with header and pipe delimited CSV files, fill in the blanks to complete the create table statement

CREATE TABLE sales (id int, unitsSold int, price FLOAT, items STRING)

LOCATION "dbfs:/mnt/sales/*.csv"

☐ FORMAT CSV

OPTIONS ("true", " | ")

☐ USING CSV

TYPE ("true", " | ")

☐ USING CSV

OPTIONS (header ="true", delimiter = " | ")

☐ FORMAT CSV

FORMAT TYPE (header ="true", delimiter = " | ")

☐ FORMAT CSV

TYPE (header ="true", delimiter = " | ")

Explanation

Answer is

USING CSV

OPTIONS (header ="true", delimiter = "|")

Here is the syntax to create an external table with additional options

CREATE TABLE table_name (col_name1 col_typ1,..)

USING data_source

OPTIONS (key='value', key2=vla2)

LOCATION = "/location"

Bottom of Form

Top of Form

Question 18

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

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

- ☐ 3
- ☐ 2
- ☐ 1
- ☐ 0
- ☐ NULL

Explanation

The answer is 2,

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

Bottom of Form

Top of Form

Question 19

You are working on a table called orders which contains data for 2021 and you have the second table called orders_2020 which contains data for 2020, you need to combine the data from two tables and there could be duplicate rows between the tables, which of the following helps you accomplish that. There could be duplicate rows between the tables.

- ☒ SELECT * FROM orders UNION SELECT * FROM orders_2020
- ☐ SELECT * FROM orders INTERSECT SELECT * FROM orders2020
- ☐ SELECT * FROM orders UNION ALL SELECT * FROM orders_2020
- ☐ SELECT * FROM orders_2020 MINUS SELECT * FROM orders
- ☐ SELECT * FROM orders JOIN orders_2020

Explanation

Answer is `SELECT * FROM orders UNION SELECT * FROM orders_2020`

UNION and UNION ALL are set operators,

UNION combines the output from both queries but also eliminates the duplicates.

UNION ALL combines the output from both queries.

Bottom of Form

Top of Form

Question 20

Which of the following python statement can be used to replace the schema name and table name in the query statement?

☐

```
table_name = "sales"
```

```
schema_name = "bronze"
```

```
query = f"select * from schema_name.table_name"
```

☐

```
table_name = "sales"
```

```
schema_name = "bronze"
```

```
query = "select * from {schema_name}.{table_name}"
```

☐

```
table_name = "sales"
```

```
schema_name = "bronze"
```

```
query = f"select * from { schema_name}.{table_name}"
```

☐

```
table_name = "sales"
```

```
schema_name = "bronze"
```

```
query = f"select * from + schema_name + "." + table_name"
```

Explanation

Answer is

```
table_name = "sales"
```

```
query = f"select * from {schema_name}.{table_name}"
```

f strings can be used to format a string. f" This is string {python variable}"

<https://realpython.com/python-f-strings/>

Bottom of Form

Top of Form

Question 21

Which of the following SQL statements can replace python variables in Databricks SQL code, when the notebook is set in SQL mode?

```
%python
```

```
table_name = "sales"
```

```
schema_name = "bronze"
```

```
%sql
```

```
SELECT * FROM _____
```

- ☐ SELECT * FROM f{schema_name.table_name}
- ☐ SELECT * FROM {schem_name.table_name}
- ☐ SELECT * FROM \${schema_name}.\${table_name}
- ☐ SELECT * FROM schema_name.table_name

Explanation

The answer is, `SELECT * FROM ${schema_name}.${table_name}`

`%python`

`table_name = "sales"`

`schema_name = "bronze"`

`%sql`

`SELECT * FROM ${schema_name}.${table_name}`

`${python variable}` -> Python variables in Databricks SQL code

Bottom of Form

Top of Form

Question 22

Your notebook accepts an input parameter called department and you are looking to control the flow of the code using department, if the department is passed then execute code and if no department is passed skip code execution.



if department is not None:

 #Execute code

else:

 pass



if (department is not None)

 #Execute code

else

 pass



if department is not None:

 #Execute code

end:

 pass



if department is not None:

 #Execute code

then:

 pass



if department is None:

 #Execute code

else:

 pass

Explanation

The answer is,

if department is not None:

 #Execute code

else:

 pass

Bottom of Form

Top of Form

Question 23

Which of the following operations are not supported on a streaming dataset view ?

```
spark.readStream.table("sales").createOrReplaceTempView("streaming_view")
```

- ☐ SELECT sum(unitssold) FROM streaming_view
- ☐ SELECT max(unitssold) FROM streaming_view
- ☐ SELECT id, sum(unitssold) FROM streaming_view
- ☐ SELECT id, count(*) FROM streaming_view group by id
- ☐ SELECT * FROM streadming_view order by id

Explanation

The answer is `SELECT id, count(*) FROM streaming_view group by id`

Certain operations are not allowed on streaming data,

<https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html#unsupported-operations>

Multiple streaming aggregations (i.e. a chain of aggregations on a streaming DF) are not yet supported on streaming Datasets.

Limit and take the first N rows are not supported on streaming Datasets.

Distinct operations on streaming Datasets are not supported.

Deduplication operation is not supported after aggregation on a streaming Datasets.

Sorting operations are supported on streaming Datasets only after an aggregation and in Complete Output Mode.

Bottom of Form

Top of Form

Question 24

Which of the following techniques structured streaming uses to ensure recovery of failures during stream processing?

- ☐ Checkpointing and Watermarking
- ☐ Write ahead logging and watermarking
- ☐ Checkpointing and write-ahead logging
- ☐ Delta time travel
- ☐ The stream will failover to available nodes in the cluster
- ☐ Checkpointing and Idempotent sinks

Explanation

The answer is Checkpointing and write-ahead logging.

Structured Streaming uses checkpointing and write-ahead logs to record the offset range of data being processed during each trigger interval.

Bottom of Form

Top of Form

Question 25

What is the underlying process that makes the Auto Loader work?

- ☐ Loader
- ☐ Delta Live Tables
- ☐ Structured Streaming
- ☐ DataFrames
- ☐ Live DataFrames

Explanation

The answer is Structured Streaming

Auto Loader is built on top of Structured Streaming, Auto Loader provides a Structured Streaming source called cloudFiles. Given an input directory path on the cloud file storage, the cloudFiles source automatically processes new files as they arrive, with the option of also processing existing files in that directory

Bottom of Form

Top of Form

Question 26

Thousands of files get uploaded to the cloud object storage for consumption, you are asked to build a process to ingest data which of the following method can be used to ingest the data incrementally, the schema of the file is expected to change overtime ingestion process should be able to handle these changes automatically.

- ☐ AUTO APPEND
- ☐ AUTO LOADER
- ☐ COPY INTO
- ☐ Structured Streaming
- ☐ Checkpoint

Explanation

The answer is AUTO LOADER,

Use Auto Loader instead of the [COPY INTO SQL command](#) when:

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

Your data schema evolves frequently. Auto Loader provides better support for schema inference and evolution. See [Configuring schema inference and evolution in Auto Loader](#).

Bottom of Form

Top of Form

Question 27

At the end of the inventory process, a file gets uploaded to the cloud object storage, you are asked to build a process to ingest data which of the following method can be used to ingest the data incrementally, schema of the file is expected to change overtime ingestion process should be able to handle these changes automatically. Below is the auto loader to command to load the data, fill in the blanks for successful execution of below code.

```
spark.readStream
  .format("cloudfiles")
  .option("_____", "csv")
  .option("_____", 'dbfs:/location/checkpoint/')
  .load(data_source)
  .writeStream
  .option("_____", 'dbfs:/location/checkpoint/')
  .option("_____", "true")
  .table(table_name))
```

- ☐ format, checkpointlocation, schemalocation, overwrite
- ☐ cloudfiles.format, checkpointlocation, cloudfiles.schemalocation, overwrite
- ☐ cloudfiles.format, cloudfiles.schemalocation, checkpointlocation, mergeSchema
- ☐ cloudfiles.format, cloudfiles.schemalocation, checkpointlocation, overwrite
- ☐ cloudfiles.format, cloudfiles.schemalocation, checkpointlocation, append

Explanation

The answer is cloudfiles.format, cloudfiles.schemalocation, checkpointlocation, mergeSchema.

Here is the end to end syntax of streaming ELT, below link contains complete options [Auto Loader options | Databricks on AWS](#)

```
spark.readStream
```

```
.format("cloudfiles") # Returns a stream data source, reads data as it arrives based on the trigger.
```

```
.option("cloudfiles.format", "csv") # Format of the incoming files
```

```
.option("cloudfiles.schemalocation", "dbfs:/location/checkpoint/") # The location to store the inferred schema and subsequent changes
```

```
.load(data_source)
```

```
.writeStream
```

```
.option("checkpointlocation", "dbfs:/location/checkpoint/") # The location of the stream's checkpoint
```

```
.option("mergeSchema", "true") # Infer the schema across multiple files and to merge the schema of each file. Enabled by default for Auto Loader when inferring the schema.
```

```
.table(table_name)) # target table
```

Bottom of Form

Top of Form

Question 28

What is the purpose of the bronze layer in a Multi-hop architecture?

- ☐ Eliminates duplicate records
- ☐ Powers ML applications
- ☐ Data quality checks, corrupt data quarantined
- ☐ Contain aggregated data that is to be consumed into Silver
- ☐ Efficient storage and querying of full, unprocessed history

Explanation

The answer is Efficient storage and querying of full, unprocessed history.

[Medallion Architecture – Databricks](#)

Bronze Layer:

Raw copy of ingested data

Replaces traditional data lake

Provides efficient storage and querying of full, unprocessed history of data

No schema is applied at this layer

Bottom of Form

Top of Form

Question 29

What is the purpose of a silver layer in Multi hop architecture?

- ☐ Replaces a traditional data lake
- ☐ Efficient storage and querying of full, unprocessed history of data
- ☐ A schema is enforced, with data quality checks.
- ☐ Refined views with aggregated data
- ☐ Optimized query performance for business-critical data

Explanation

The answer is, A schema is enforced, with data quality checks.

[Medallion Architecture – Databricks](#)

Silver Layer:

Reduces data storage complexity, latency, and redundancy

Optimizes ETL throughput and analytic query performance

Preserves grain of original data (without aggregation)

Eliminates duplicate records

production schema enforced

Data quality checks, quarantine corrupt data

Bottom of Form

Top of Form

Question 30

What is the purpose of a gold layer in Multi-hop architecture?

- ☐ Optimizes ETL throughput and analytic query performance
- ☐ Eliminate duplicate records
- ☐ Preserves grain of original data, without any aggregations
- ☐ Data quality checks and schema enforcement
- ☐ Powers ML applications, reporting, dashboards and adhoc reports.

Explanation

The answer is Powers ML applications, reporting, dashboards and adhoc reports.

Review the below link for more info,

[Medallion Architecture – Databricks](#)

Gold Layer:

Powers ML applications, reporting, dashboards, ad hoc analytics

Refined views of data, typically with aggregations

Reduces strain on production systems

Optimizes query performance for business-critical data

Bottom of Form

Top of Form

Question 31

You are currently asked to work on building a data pipeline, you have noticed that you are currently working on a very large scale ETL many data dependencies, which of the following tools can be used to address this problem?

- ☐ AUTO LOADER
- ☐ JOBS and TASKS
- ☐ SQL Endpoints
- ☐ DELTA LIVE TABLES
- ☐ STRUCTURED STREAMING with MULTI HOP

Explanation

The answer is, DELTA LIVE TABLES

DLT simplifies data dependencies by building DAG-based joins between live tables. Here is a view of how the dag looks with data dependencies without additional meta data,

```
create or replace live view customers
```

```
select * from customers;
```

```
create or replace live view sales_orders_raw
```

```
select * from sales_orders;
```

```
create or replace live view sales_orders_cleaned
```

```
as
```

```
select sales.* from
```

```
live.sales_orders_raws
```

```
join live.customers c
```

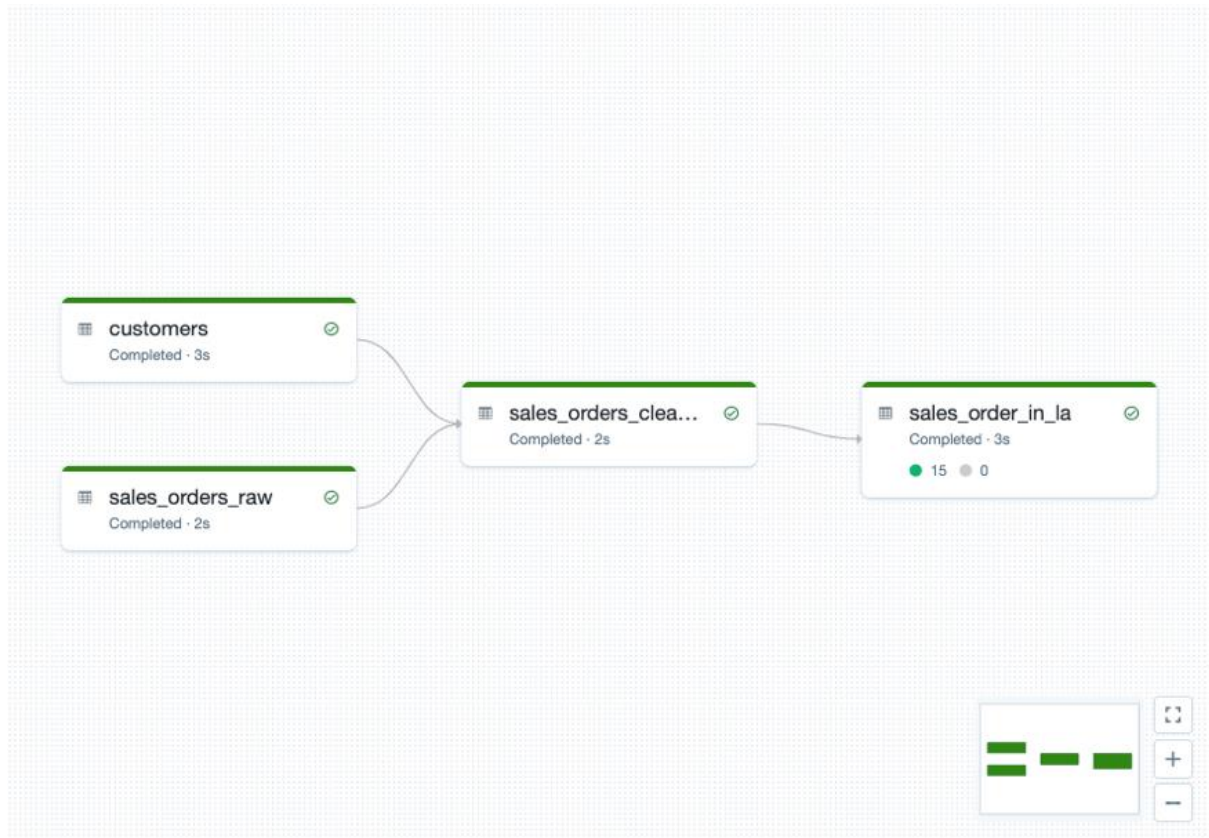
```
on c.customer_id=s.customer_id
```

```
where c.city= 'LA';
```

```
create or replace live table sales_orders_in_la
```

```
select * from sales_orders_cleaned;
```

Above code creates below dag



Documentation on DELTA LIVE TABLES,

<https://databricks.com/product/delta-live-tables>

<https://databricks.com/blog/2022/04/05/announcing-generally-availability-of-databricks-delta-live-tables-dlt.html>

DELTA LIVE TABLES, addresses below challenges when building ETL processes

Complexities of large scale ETL

Hard to build and maintain dependencies

Difficult to switch between batch and stream

Data quality and governance

Difficult to monitor and enforce data quality

Impossible to trace data lineage

Difficult pipeline operations

Poor observability at granular data level

Error handling and recovery is laborious

Bottom of Form

Top of Form

Question 32

How do you create a delta live tables pipeline and deploy using DLT UI?

- ☐ Within the Workspace UI, click on Workflows, select Delta Live tables and create a pipeline and select the notebook with DLT code.
- ☐ Under Cluster UI, select SPARK UI and select Structured Streaming and click create pipeline and select the notebook with DLT code.
- ☐ There is no UI, you can only setup DELTA LIVE TABLES using Python and SQL API and select the notebook with DLT code.
- ☐ Use VS Code and download DBX plugin, once the plugin is loaded you can build DLT pipelines and select the notebook with DLT code.
- ☐ Within the Workspace UI, click on SQL Endpoint, select Delta Live tables and create pipeline and select the notebook with DLT code.

Explanation

The answer is, Within the Workspace UI, click on Workflows, select Delta Live tables and create a pipeline and select the notebook with DLT code.

<https://docs.databricks.com/data-engineering/delta-live-tables/delta-live-tables-quickstart.html>

Bottom of Form

Top of Form

Question 33

You are noticing job cluster is taking 6 to 8 mins to start which is delaying your job to finish on time, what steps you can take to reduce the amount of time cluster startup time

- ☐ Setup a second job ahead of first job to start the cluster, so the cluster is ready with resources when the job starts
- ☐ Use All purpose cluster instead to reduce cluster start up time
- ☐ Reduce the size of the cluster, smaller the cluster size shorter it takes to start the cluster
- ☐ Use cluster pools to reduce the startup time of the jobs
- ☐ Use SQL endpoints to reduce the startup time

Explanation

The answer is, Use cluster pools to reduce the startup time of the jobs.

Cluster pools allow us to reserve VM's ahead of time, when a new job cluster is created VM are grabbed from the pool. Note: when the VM's are waiting to be used by the cluster only cost incurred is Azure. Databricks run time cost is only billed once VM is allocated to a cluster.

Here is a demo of how to setup and follow some best practices,

https://www.youtube.com/watch?v=FVtITxOabxg&ab_channel=DatabricksAcademy

Bottom of Form

Top of Form

Question 34

Data engineering team has a job currently setup to run a task load data into a reporting table every day at 8: 00 AM takes about 20 mins, Operations teams are planning to use that data to run a second job, so they access latest complete set of data. What is the best to way to orchestrate this job setup?

- ☐ Add Operation reporting task in the same job and set the Data Engineering task to depend on Operations reporting task
- ☐ Setup a second job to run at 8:20 AM in the same workspace
- ☐ Add Operation reporting task in the same job and set the operations reporting task to depend on Data Engineering task
- ☐ Use Auto Loader to run every 20 mins to read the initial table and set the trigger to once and create a second job
- ☐ Setup a Delta live to table based on the first table, set the job to run in continuous mode

Explanation

The answer is Add Operation reporting task in the same job and set the operations reporting task to depend on Data Engineering task.

Task name * ?

OperationsReporting

Type *

Notebook

Source * ?

Workspace

Path * ?

/Users/ [REDACTED]

Cluster * ?

cluster (126.00 GB | 36 Cores | DBR 10.4 LTS | Spark 3.2.1 | Scala 2.12)

Parameters ?

UI | JSON

Add

Depends on

Dataengineering x

Advanced options

Cancel

Create task

Job View



Bottom of Form

Top of Form

Question 35

The data engineering team noticed that one of the job normally finishes in 15 mins but gets stuck randomly when reading remote databases due to network packet drops and the job takes really long time to finish, which of the following option can be used to fix the problem?

- ☐ Use Databrick REST API to monitor long running jobs and issue a kill command
- ☐ Use Jobs runs, active runs UI section to monitor and kill long running job
- ☐ Modify the task, to include a timeout to kill the job if it runs more than 15 mins.
- ☐ Use Spark job time out setting in the Spark UI
- ☐ Use Cluster timeout setting in the Job cluster UI

Explanation

The answer is, Modify the task, to include time out to kill the job if it runs more than 15 mins.

<https://docs.microsoft.com/en-us/azure/databricks/data-engineering/jobs/jobs#timeout>

Bottom of Form

Top of Form

Question 36

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

- ☐ Python
- ☐ Scala
- ☐ SQL
- ☐ R
- ☐ All of the above

Explanation

The answer is SQL

Bottom of Form

Top of Form

Question 37

You have noticed that Databricks SQL queries are running slow, you were asked to look reason why queries are running slow and identify steps to improve the performance, when you looked at the code you noticed all the queries are running sequentially and using a SQL endpoint cluster. Which of the following steps can be taken to resolve the issue?

- ☐ Turn on the Serverless feature for the SQL endpoint.
- ☐ Increase the maximum bound of the SQL endpoint's scaling range.
- ☐ Increase the cluster size of the SQL endpoint.
- ☐ Turn on the Auto Stop feature for the SQL endpoint.
- ☐ Turn on the Serverless feature for the SQL endpoint and change the Spot Instance Policy to "Reliability Optimized."

Explanation

The answer is increase the cluster size of the SQL Endpoint,

SQL endpoint scales horizontally(scale-out) and vertical (scale-up), you have to understand when to use what.

Scale-out -> Add more clusters, change max number of clusters

If you are trying to improve the throughput, being able to run as many queries as possible then having an additional cluster(s) will improve the performance.

Scale-up-> Increase the size of the cluster from x-small to small, to medium, X Large....

If you are trying to improve the performance of a single query having additional memory, additional nodes and cpu in the cluster will improve the performance.

SQL endpoint

Starter Endpoint

The screenshot shows the 'Starter Endpoint' configuration form. It includes a 'Name' field with the value 'Starter Endpoint'. Below this is a 'Cluster size' section with a dropdown menu currently set to 'X-Small' and a label '6 DBU / cluster'. An arrow points from an oval labeled 'Scale up' to this dropdown. Further down is an 'Auto stop' section with a toggle switch turned on and a text field set to '60 minutes of inactivity'. At the bottom is a 'Scaling' section with 'Min.' set to '1' and 'Max.' set to '2', with a label 'clusters (6 to 12 DBU)'. An arrow points from an oval labeled 'Scale out' to the 'Max.' field.

Bottom of Form

Top of Form

Question 38

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

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

- Create ____ query that calculates total units sold
 - Setup ____ with query on trigger condition Units Sold > 10,000
 - Setup ____ to run every 5 mins
 - Add destination _____
- ☐ Python, Job, SQL Cluster, email address
 - ☐ SQL, Alert, Refresh, email address
 - ☐ SQL, Job, SQL Cluster, email address
 - ☐ SQL, Job, Refresh, email address
 - ☐ Python, Job, Refresh, email address

Explanation


The answer is SQL, Alert, Refresh, email address


Here the steps from Databricks documentation,

[Create an alert](#)

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

Do one of the following:

Click  Create in the sidebar and select Alert.

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

Search for a target query.

New Alert

Start by selecting the query that you would like to monitor using the search bar. [Setup Instructions ?](#)
Keep in mind that Alerts do not work with queries that use parameters.

Query:

Create Alert

To alert on multiple columns, you need to modify your query. See [Alert on multiple columns](#).

In the Trigger when field, configure the alert.

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

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

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

Start by selecting the query that you would like to monitor using the search bar. Keep in mind that Alerts do not work with queries that use parameters. [Setup Instructions](#)

Query:

▲ This query has no refresh schedule. [Why it's recommended](#)

Value column	Condition	Threshold
Trigger when: <input type="text" value="value"/>	<input type="text" value=">"/>	<input type="text" value="1"/>

Top row value is ●

When triggered, send notification:

Template:

[Create Alert](#)

Note

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

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

Just once: Send a notification when the [alert status](#) changes from OK to TRIGGERED.

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

At most every: Send a notification whenever the alert status is TRIGGERED at a specific interval. This choice lets you avoid notification spam for alerts that trigger often.

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

In the Template drop-down, choose a template:

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

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

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

ALERT_STATUS: The evaluated alert status (string).

ALERT_CONDITION: The alert condition operator (string).

ALERT_THRESHOLD: The alert threshold (string or number).

ALERT_NAME: The alert name (string).

ALERT_URL: The alert page URL (string).

QUERY_NAME: The associated query name (string).

QUERY_URL: The associated query page URL (string).

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

QUERY_RESULT_ROWS: The query result rows (value array).

QUERY_RESULT_COLS: The query result columns (string array).

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

Click the Preview toggle button to preview the rendered result.

Important

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

Click the Save Changes button.

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

If the query is a Run as owner query, the query runs using the query owner's credential on the alert's refresh schedule.

If the query is a Run as viewer query, the query runs using the alert creator's credential on the alert's refresh schedule.

Click Create Alert.

Choose an [alert destination](#).

Important

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

Generic Alert [Edit] [More]

STATUS: TRIGGERED
Last triggered 7 minutes ago

Query: [Generic Query 1](#)
Scheduled to refresh every minute

Trigger when: Value column: value Condition: = Threshold: 1808
Top row value is 1808

Notifications: Notifications are sent just once, until back to normal.
Set to default notification template.

Destinations [Add]

<input checked="" type="checkbox"/> Email	[X]
<input type="checkbox"/> #ops	[X]
<input type="checkbox"/> #Platform	[X]
<input type="checkbox"/> Test Webhook	[X]

Bottom of Form

Top of Form

Question 39

The marketing team is launching a new campaign to monitor the performance of the new campaign for the first two weeks, they would like to set up a dashboard with a refresh schedule to run every 5 minutes, which of the below steps can be taken to reduce of the cost of this refresh over time?

- ☐ Reduce the size of the SQL Cluster size
- ☐ Reduce the max size of auto scaling from 10 to 5
- ☐ Setup the dashboard refresh schedule to end in two weeks
- ☐ Change the spot instance policy from reliability optimized to cost optimized
- ☐ Always use X-small cluster

Explanation

The answer is Setup the dashboard refresh schedule to end in two weeks

Bottom of Form

Top of Form

Question 40

Which of the following tool provides Data Access control, Access Audit, Data Lineage, and Data discovery?

- ☐ DELTA LIVE Pipelines
- ☐ Unity Catalog
- ☐ Data Governance
- ☐ DELTA lake
- ☐ Lakehouse

Explanation

The answer is Unity Catalog

Bottom of Form

Top of Form

Question 41

Data engineering team is required to share the data across with Data science team, both the teams are using different workspaces which of the following techniques can be used to simplify sharing data across?

- ☐ Data Sharing
- ☐ Unity Catalog
- ☐ DELTA lake
- ☐ Use single storage location
- ☐ DELTA LIVE Pipelines

Explanation

The answer is Unity catalog.



Review product features

<https://databricks.com/product/unity-catalog>

Bottom of Form

Top of Form

Question 42

A newly joined team member John Smith in the Marketing team who currently does not have any access to the data requires read access to customers table, which of the following statements can be used to grant access.

- ☐ GRANT SELECT, USAGE TO john.smith@marketing.com ON TABLE customers
- ☐ GRANT READ, USAGE TO john.smith@marketing.com ON TABLE customers
- ☐ GRANT SELECT, USAGE ON TABLE customers TO john.smith@marketing.com
- ☐ GRANT READ, USAGE ON TABLE customers TO john.smith@marketing.com
- ☐ GRANT READ, USAGE ON customers TO john.smith@marketing.com

Explanation

The answer is GRANT SELECT, USAGE ON TABLE customers TO john.smith@marketing.com

[Data object privileges - Azure Databricks | Microsoft Docs](#)

Bottom of Form

Top of Form

Question 43

Grant full privileges to new marketing user Kevin Smith to table sales

- ☐ GRANT FULL PRIVILEGES TO kevin.smith@marketing.com ON TABLE sales
- ☐ GRANT ALL PRIVILEGES TO kevin.smith@marketing.com ON TABLE sales
- ☐ GRANT FULL PRIVILEGES ON TABLE sales TO kevin.smith@marketing.com
- ☐ GRANT ALL PRIVILEGES ON TABLE sales TO kevin.smith@marketing.com
- ☐ GRANT ANY PRIVILEGE ON TABLE sales TO kevin.smith@marketing.com

Explanation

The answer is GRANT ALL PRIVILEGE ON TABLE sales TO kevin.smith@marketing.com

GRANT <privilege> ON <securable_type> <securable_name> TO <principal>

Here are the available privileges and ALL Privileges gives full access to an object.

Privileges

SELECT: gives read access to an object.

CREATE: gives ability to create an object (for example, a table in a schema).

MODIFY: gives ability to add, delete, and modify data to or from an object.

USAGE: does not give any abilities, but is an additional requirement to perform any action on a schema object.

READ_METADATA: gives ability to view an object and its metadata.

CREATE_NAMED_FUNCTION: gives ability to create a named UDF in an existing catalog or schema.

MODIFY_CLASSPATH: gives ability to add files to the Spark class path.

ALL PRIVILEGES: gives all privileges (is translated into all the above privileges).

Bottom of Form

Top of Form

Question 44

Which of the following locations in the Databricks product architecture hosts the notebooks and jobs?

- ☐ Data plane
- ☐ Control plane
- ☐ Databricks Filesystem
- ☐ JDBC data source
- ☐ Databricks web application

Explanation

The answer is Control Pane,

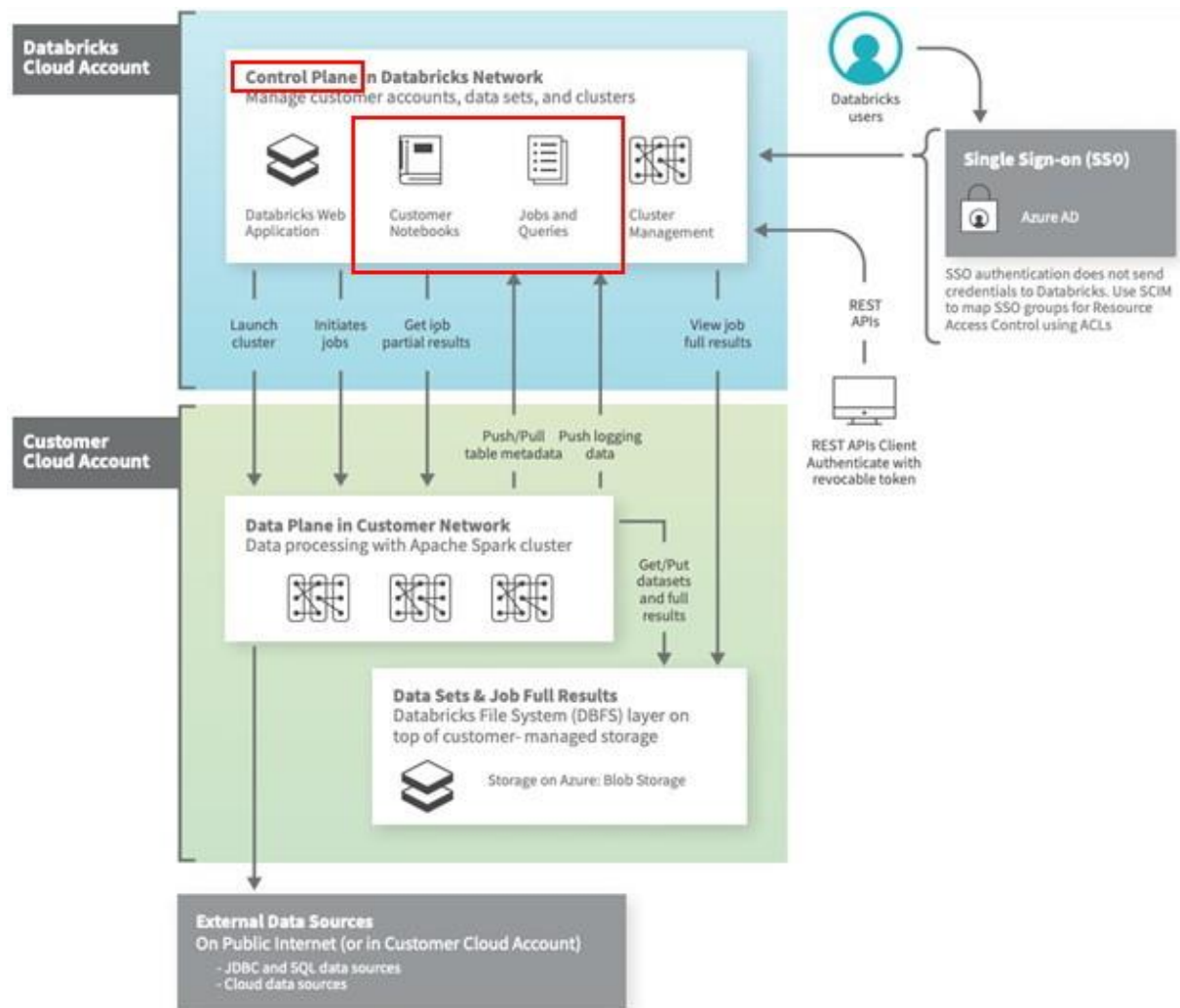
Databricks operates most of its services out of a control plane and a data plane, please note serverless features like SQL Endpoint and DLT compute use shared compute in Control plane.

Control Plane: Stored in Databricks Cloud Account

The control plane includes the backend services that Databricks manages in its own Azure account. Notebook commands and many other workspace configurations are stored in the control plane and encrypted at rest.

Data Plane: Stored in Customer Cloud Account

The data plane is managed by your Azure account and is where your data resides. This is also where data is processed. You can use Azure Databricks connectors so that your clusters can connect to external data sources outside of your Azure account to ingest data or for storage.



Bottom of Form

Top of Form

Question 45

A dataset has been defined using Delta Live Tables and includes an expectations clause: `CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION FAIL UPDATE`

What is the expected behavior when a batch of data containing data that violates these constraints is processed?

- ☐ Records that violate the expectation are added to the target dataset and recorded as invalid in the event log.
- ☐ Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.
- ☐ Records that violate the expectation cause the job to fail
- ☐ Records that violate the expectation are added to the target dataset and flagged as invalid in a field added to the target dataset.
- ☐ Records that violate the expectation are dropped from the target dataset and loaded into a quarantine table.

Explanation

The answer is Records that violate the expectation cause the job to fail

See below notes for additional notes,

There are three types of DLT expectations

Invalid records:

Use the expect operator when you want to keep records that violate the expectation. Records that violate the expectation are added to the target dataset along with valid records:

SQL

```
CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01')
```

Drop invalid records:

Use the expect or drop operator to prevent the processing of invalid records. Records that violate the expectation are dropped from the target dataset:

SQL

```
CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION DROP ROW
```

Fail on invalid records:

When invalid records are unacceptable, use the expect or fail operator to halt execution immediately when a record fails validation. If the operation is a table update, the system atomically rolls back the transaction:

SQL

```
CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION FAIL UPDATE
```

Bottom of Form