Question 1

How does Lakehouse replace dependency on using Data lakes and Data warehouses in a Data and Analytics solution?

○ Open, direct access to data stored in standard data formats.

○ Supports ACID transactions.

○ Supports BI and Machine learning workloads

○ Support for end-to-end streaming and batch workloads

○ All the above

Question 2

You are currently working on storing data you receive from different customer surveys this data is highly unstructured and changes over time, why Lakehouse is a better choice compared to a Data warehouse?

○ Lakehouse supports schema enforcement and evolution, traditional data warehouses lack schema evolution.

○ Lakehouse supports SQL

○ Lakehouse supports ACID

○ Lakehouse enforces data integrity

○ Lakehouse supports primary and foreign keys like a data warehouse

Question 3

Which of the following locations hosts the driver and worker nodes of a Databricks-managed cluster?

○ Data plane

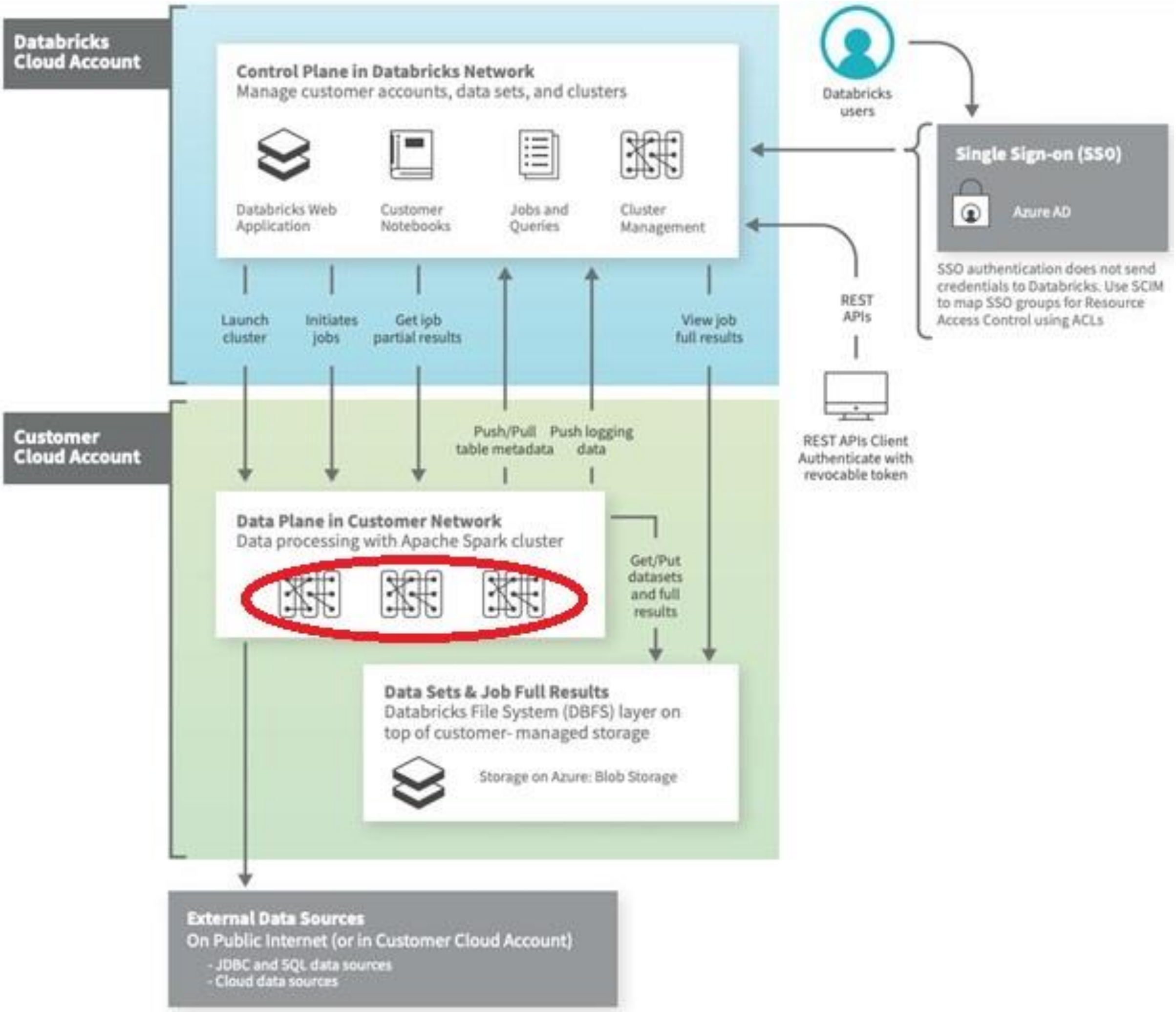
○ Control plane

○ Databricks Filesystem

● JDBC data source

○ Databricks web application

Explanation


Databricks Cloud Account
Control Plane in Databricks Network
Manage customer accounts, data sets, and clusters
Databricks Web Application
Customer Notebooks
Jobs and Queries
Cluster Management
Databricks users
Single Sign-on (SSO)
Azure AD
SSO authentication does not send credentials to Databricks. Use SCIM to map SSO groups for Resource Access Control using ACLs
REST APIs
REST APIs Client Authenticate with revocable token
Launch cluster
Initiates jobs
Get job partial results
View job full results
Customer Cloud Account
Push/Pull table metadata
Push logging data
Data Plane in Customer Network
Data processing with Apache Spark cluster
Get/Put datasets and full results
Data Sets & Job Full Results
Databricks File System (DBFS) layer on top of customer- managed storage
Storage on Azure: Blob Storage
External Data Sources
On Public Internet (or in Customer Cloud Account)
- JDBC and SQL data sources
- Cloud data sources

Question 4

You have written a notebook to generate a summary data set for reporting, Notebook was scheduled using the job cluster, but you realized it takes 8 minutes to start the cluster, what feature can be to start the cluster in a timely fashion?

○

Setup an additional job to run ahead of the actual job so the cluster is running second job starts

○

Use the Databricks cluster pools feature to reduce the startup time

○

Use Databricks Premium edition instead of Databricks standard edition

○

Pin the cluster in the cluster UI page so it is always available to the jobs

○

Disable auto termination so the cluster is always running

Explanation

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

Question 5

Which of the following statement is true about Databricks repos?

○

You can approve the pull request if you are the owner of Databricks repos

○

A workspace can only have one instance of git integration

○

Databricks Repos and Notebook versioning are the same features

○

You cannot create a new branch in Databricks repos

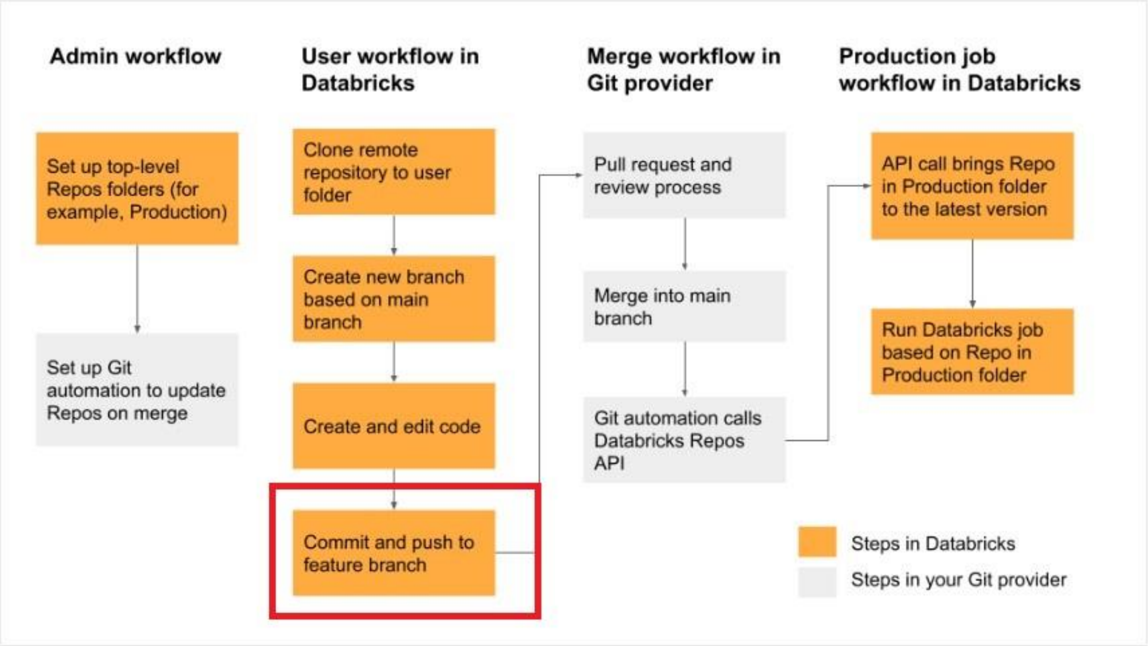
○

Databricks repos allow you to comment and commit code changes and push them to remote branch

Explanation

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

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


Admin workflow
User workflow in Databricks
Merge workflow in Git provider
Production job workflow in Databricks
Set up top-level Repos folders (for example, Production)
Set up Git automation to update Repos on merge
Clone remote repository to user folder
Create new branch based on main branch
Create and edit code
Commit and push to feature branch
Pull request and review process
Merge into main branch
Git automation calls Databricks Repos API
API call brings Repo in Production folder to the latest version
Run Databricks job based on Repo in Production folder
Steps in Databricks
Steps in your Git provider

Question 6

Which of the statement is correct about the cluster pools?

○

Cluster pools allow you to perform load balancing

○

Cluster pools allow you to create a cluster

○

Cluster pools allow you to save time when starting a new cluster

○

Cluster pools are used to share resources among multiple teams

○

Cluster pools allow you to have all the nodes in the cluster from single physical server rack

Question 7

Once a Custer is deleted below additional actions needs to performed by the administrator

○

Remove virtual machines but storage and networking are automatically dropped

○

Drop storage disks but Virtual machines and networking are automatically dropped

○

Remove networking but Virtual machines and storage disks are automatically dropped

○

Remove logs

○

No action needs to be performed. All resources are automatically removed.

Question 8

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

What is Delta?

Delta lake is

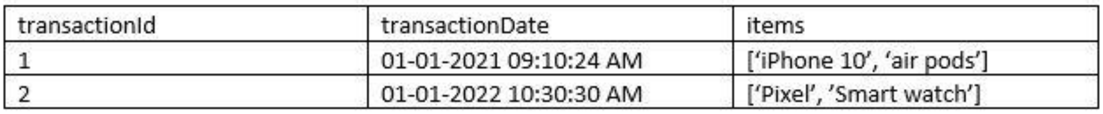
· Open source
· Builds up on standard data format
· Optimized for cloud object storage
· Built for scalable metadata handling

Delta lake is not

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

Question 9

You were asked to create a delta table to store below transaction data?



transactionId	transactionDate	items
1	01-01-2021 09:10:24 AM	[‘iPhone 10’, ‘air pods’]
2	01-01-2022 10:30:30 AM	[‘Pixel’, ’Smart watch’]

○

```
CREATE DELTA TABLE transactions (
transactionId int,
transactionDate timestamp,
items string)
```

○

```
CREATE TABLE transactions (
transactionId int,
transactionDate timestamp,
items string)
FORMAT DELTA
```

○

```
CREATE TABLE transactions (
transactionId int,
transactionDate timestamp,
items string)
```

○

```
CREATE TABLE transactions (
transactionId int,
transactionDate timestamp,
items ARRAY<string>)
```

○

```
CREATE TABLE transactions (
transactionId int,
transactionDate timestamp,
items string<array>)
```

Question 10

Drop a DELTA table

○

`DROP DELTA table_name`

○

`DROP TABLE table_name`

○

`DROP TABLE table_name FORMAT DELTA`

○

`DROP  table_name`

Question 11

Delete records from the transactions Delta table where transactionDate is greater than current timestamp?

○

`DELETE FROM transactions FORMAT DELTA where transactionDate > currenct_timestmap()`

○

`DELETE FROM transactions if transctionDate > current_timestamp()`

○

`DELETE FROM transactions where transactionDate > current_timestamp()`

○

`DELETE FROM transactions where transactionDate > current_timestamp() KEEP_HISTORY`

○

`DELET FROM transactions where transactionDate GE current_timestamp()`

Question 12

Identify one of the below statements that can query a delta table in PySpark Dataframe API

○

`Spark.read.mode(“delta”).table(“table_name)`

○

`Spark.read.table.delta(“table_name)`

○

`Spark.read.table(“table_name”)`

○

`Spark.read.format(“delta”).LoadTableAs(“table_name”)`

○

`Spark.read.format(“delta”).TableAs(“table_name”)`

Question 13

The default threshold of VACUUM is 7 days, internal audit team asked to certain tables to maintain at least 365 days as part of compliance requirement, which of the below setting is needed to implement.

○

```
ALTER TABLE table_name set TBLPROPERTIES
(delta.deletedFileRetentionDuration= ‘interval 365 days’)
```

○

```
MODIFY TABLE table_name set TBLPROPERTY (delta.maxRetentionDays =
‘interval 365 days’)
```

○

```
ALTER TABLE table_name set EXENDED TBLPROPERTIES
(delta.deletedFileRetentionDuration= ‘interval 365 days’)
```

○

```
ALTER TABLE table_name set EXENDED TBLPROPERTIES (delta.vaccum.duration=
‘interval 365 days’)
```

Explanation

```
ALTER TABLE table_name SET TBLPROPERTIES ( property_key [ = ] property_val [, ...]
)
```

TBLPROPERTIES allow you to set key-value pairs

Table properties and table options (Databricks SQL) | Databricks on AWS

Question 14

Which of the following commands can be used to read a delta table in SQL

○

Spark.sql(“select * from table_name”)

○

%sql Select * from table_name

○

Both A & B

○

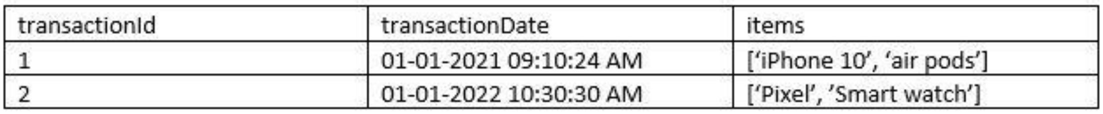
Execute.sql(“select * from table”)

○

Delta.sql(“select * from table”)

Question 15

Create a managed delta table called transactions to store below data



transactionId	transactionDate	items
1	01-01-2021 09:10:24 AM	['iPhone 10', 'air pods']
2	01-01-2022 10:30:30 AM	['Pixel', 'Smart watch']

○

```
CREATE MANAGED TABLE transactions (
transactionid int,
transctionDate timestamp,
items Array<string>)
```

○

```
CREATE TABLE transactions (
transactionid int,
transctionDate timestamp,
items Array<string>)
```

○

```
CREATE MANAGED DELTA TABLE transactions (
transactionid int,
transctionDate timestamp,
items Array<string>)
```

○

```
CREATE TABLE transactions (
transactionid int,
transctionDate timestamp,
items Array<string>)
LOCATION '/mnt/managed/'
```

○

```
CREATE TABLE transactions (
transactionid int,
transctionDate timestamp,
items Array<string>)
LOCATION USING DELTA
```

Explanation

The answer is

```
CREATE TABLE transactions (
transactionid int,
transctionDate timestamp,
items Array<string>)
```

By default, any table without a location keyword is a Delta Managed table.

Question 16

Which of the following SQL statements can be used to update a transactions table, to set a flag on the table from Y to N

○

`MODIFY transactions SET active_flag = 'N' WHERE active_flag = 'Y'`

○

`MERGE transactions SET active_flag = 'N' WHERE active_flag = 'Y'`

○

`UPDATE transactions SET active_flag = 'N' WHERE active_flag = 'Y'`

○

`REPLACE transactions SET active_flag = 'N' WHERE active_flag = 'Y'`

Explanation

The answer is

```
UPDATE transactions SET active_flag = 'N' WHERE active_flag = 'Y'
```

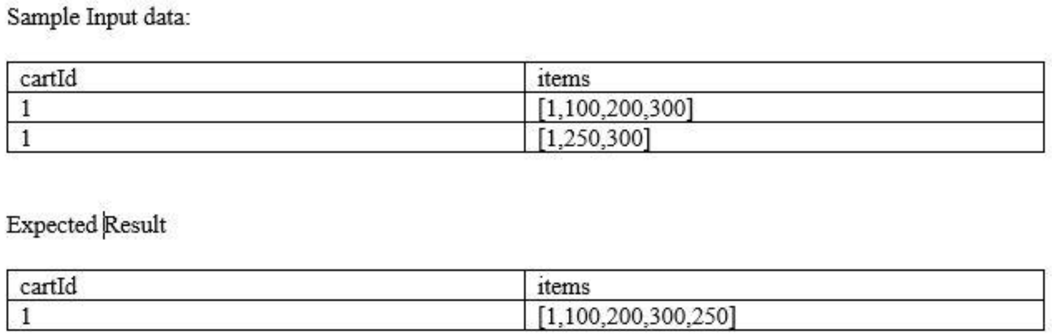
Delta Lake supports UPDATE statements on the delta table, all of the changes as part of the update are ACID compliant.

Question 17
Customers can add many items to the cart and remove them before final checkout, below sample input data cointains every time a customer makes a change to the cart which is stored as an array in the Delta table, Marketing team asked you to create a unique list of item’s that were ever added to the cart by each customer, fill in blanks by choosing the appropriate array function so the query produces below expected result as shown below.

Customer can add many items

Schema: cartId INT, items Array<INT>

Sample Data



Sample Input data:

cartId	items
1	[1,100,200,300]
1	[1,250,300]

Expected Result

cartId	items
1	[1,100,200,300,250]

```
SELECT cartId, ___ (___(items)) as items
FROM carts GROUP BY cartId
```

Expected result
cartId items
1 [1,100,200,300,250]

○
FLATTEN, COLLECT_UNION

○
ARRAY_UNION, FLATTEN

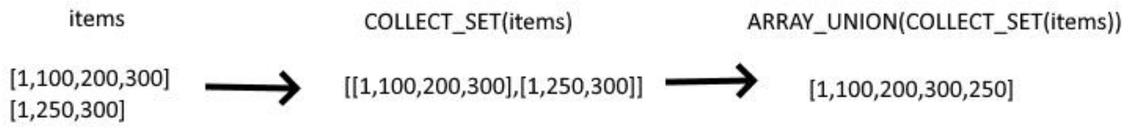
○
ARRAY_UNION, ARRAY_DISTINT

○
ARRAY_UNION, COLLECT_SET

○
ARRAY_DISTINCT, ARRAY_UNION

Explanation
COLLECT SET is a kind of aggregate function that combines a column value from all rows into a unique list

ARRAY_UNION combines and removes any duplicates,



items

[1,100,200,300]
[1,250,300]

COLLECT_SET(items)

[[1,100,200,300],[1,250,300]]

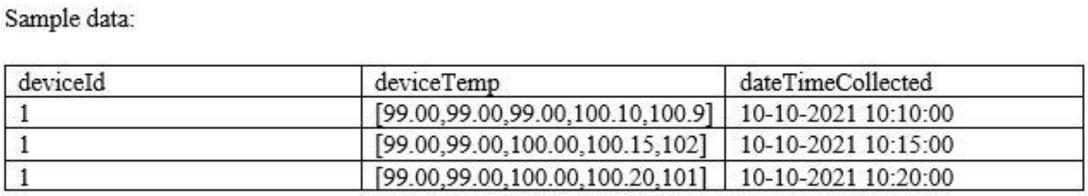
ARRAY_UNION(COLLECT_SET(items))

[1,100,200,300,250]

Question 18

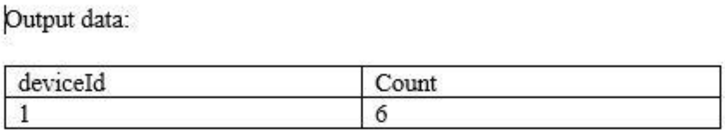
You were asked to identify number of times a temperature sensor exceed threshold temperature (100.00) by each device, each row contains 5 readings collected every 5 minutes, fill in the blank with the appropriate functions

```
Schema: deviceId INT, deviceTemp ARRAY<double>, dateTimeCollected
TIMESTAMP
```



Sample data:

deviceId	deviceTemp	dateTimeCollected
1	[99.00,99.00,99.00,100.10,100.9]	10-10-2021 10:10:00
1	[99.00,99.00,100.00,100.15,102]	10-10-2021 10:15:00
1	[99.00,99.00,100.00,100.20,101]	10-10-2021 10:20:00



Output data:

deviceId	Count
1	6

```
SELECT deviceId, __ (__ (__(deviceTemp], i -> i > 100.00)))
FROM devices
GROUP BY deviceId
```

○

SUM, COUNT, SIZE

○

SUM, SIZE, SLICE

○

SUM, SIZE, ARRAY_CONTAINS

○

SUM, SIZE, ARRAY_FILTER

○

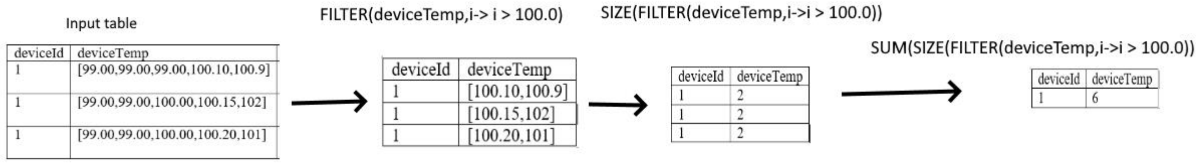
SUM, SIZE, FILTER

Explanation

FILER function can be used to filter an array based on an expression

SIZE function can be used to get size of an array

SUM is used to calculate to total by device



Input table

deviceId	deviceTemp
1	[99.00,99.00,99.00,100.10,100.9]
1	[99.00,99.00,100.00,100.15,102]
1	[99.00,99.00,100.00,100.20,101]

FILTER(deviceTemp,i-> i > 100.0)

deviceId	deviceTemp
1	[100.10,100.9]
1	[100.15,102]
1	[100.20,101]

SIZE(FILTER(deviceTemp,i->i > 100.0))

deviceId	deviceTemp
1	2
1	2
1	2

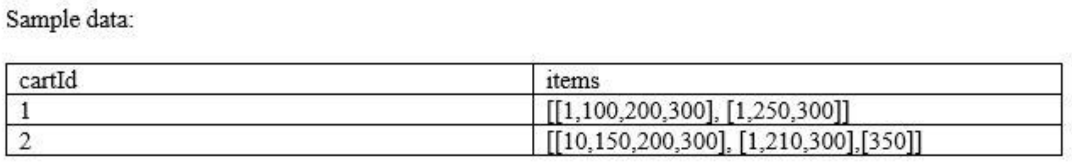
SUM(SIZE(FILTER(deviceTemp,i->i > 100.0))

deviceId	deviceTemp
1	6

Question 19

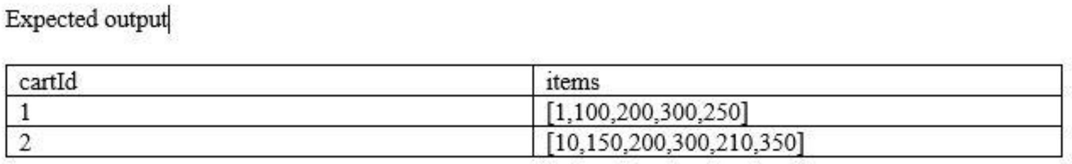
You were asked to create a unique item's list that were added to the cart by user, fill in blanks by choosing the appropriate function

```
Schema: cartId INT, items Array<INT>
```



Sample data:

cartId	items
1	[[1,100,200,300], [1,250,300]]
2	[[10,150,200,300], [1,210,300],[350]]



Expected output

cartId	items
1	[1,100,200,300,250]
2	[10,150,200,300,210,350]

```
SELECT cartId, _(_(items)) FROM carts
```

○ ARRAY_UNION, ARRAY_DISCINT

○ ARRAY_DISTINCT, ARRAY_UNION

○ ARRAY_DISTINCT, FLATTEN

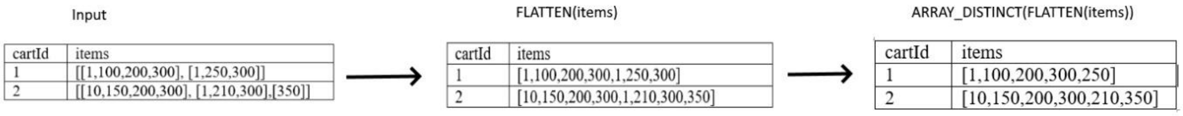
○ FLATTEN, ARRAY_DISTINCT

○ ARRAY_DISTINCT, ARRAY_FLATTEN

Explanation

FLATTEN -> Transforms an array of arrays into a single array.

ARRAY_DISTINCT -> The function returns an array of the same type as the input argument where all duplicate values have been removed.



Input

cartId	items
1	[[1,100,200,300], [1,250,300]]
2	[[10,150,200,300], [1,210,300],[350]]

FLATTEN(items)

cartId	items
1	[1,100,200,300,1,250,300]
2	[10,150,200,300,1,210,300,350]

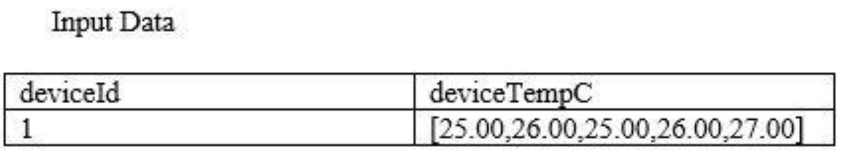
ARRAY_DISTINCT(FLATTEN(items))

cartId	items
1	[1,100,200,300,250]
2	[10,150,200,300,210,350]

Question 20

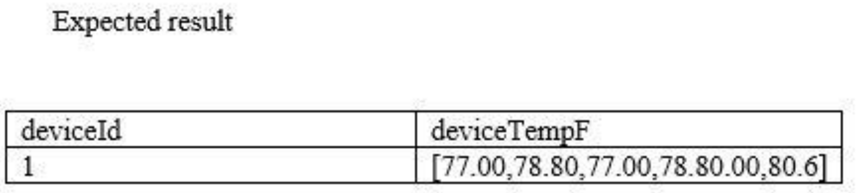
You are working on IOT data where each device has 5 reading in an array collected in Celsius, you were asked to covert each individual reading from Celsius to Fahrenheit, fill in the blank with an appropriate function that can be used in this scenario.

```
Schema: deviceId INT, deviceTemp ARRAY<double>
```



Input Data

deviceId	deviceTempC
1	[25.00,26.00,25.00,26.00,27.00]



Expected result

deviceId	deviceTempF
1	[77.00,78.80,77.00,78.80.00,80.6]

```
SELECT deviceId, __(deviceTempC,i-> (i * 9/5) + 32) as deviceTempF
FROM sensors
```

○ APPLY

○ MULTIPLY

○ ARRAYEXPR

○ TRANSFORM

○ FORALL

Explanation

TRANSFORM -> Transforms elements in an array in `expr` using the function `func`.

```
transform(expr, func)
```

Question 21

Which of the following array functions takes input column return unique list of values in an array?

○

COLLECT_LIST

○

COLLECT_SET

○

COLLECT_UNION

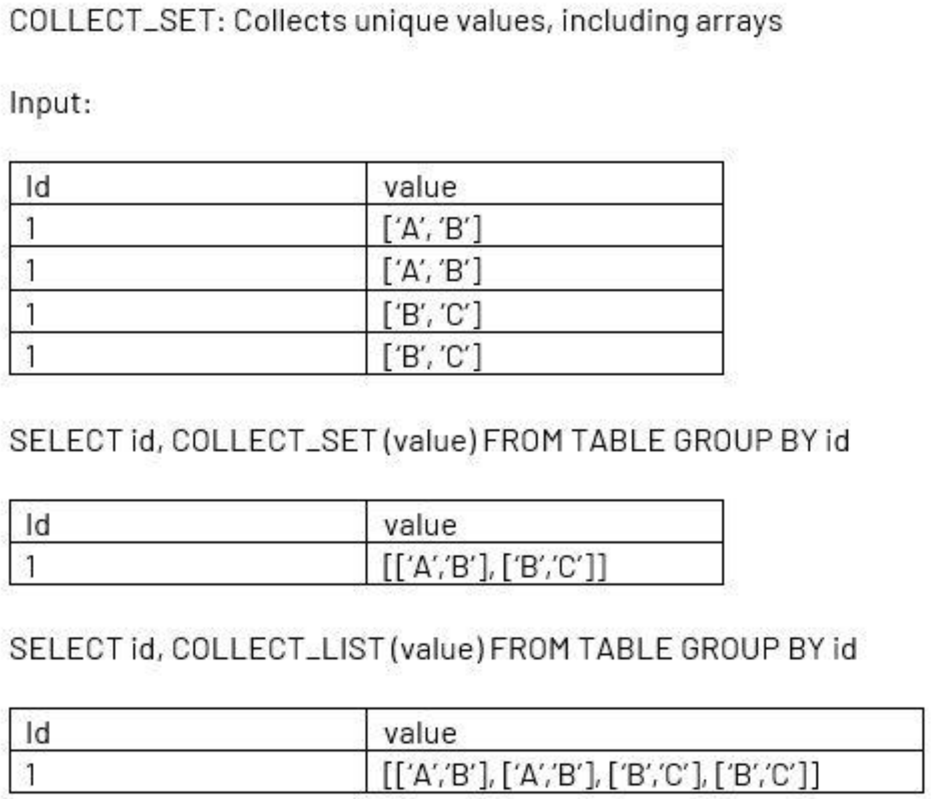
○

ARRAY_INTERSECT

○

ARRAY_UNION

Explanation



COLLECT_SET: Collects unique values, including arrays

Input:

Id	value
1	[‘A’, ’B’]
1	[‘A’, ’B’]
1	[‘B’, ’C’]
1	[‘B’, ’C’]

SELECT id, COLLECT_SET (value) FROM TABLE GROUP BY id

Id	value
1	[[‘A’,’B’], [‘B’,’C’]]

SELECT id, COLLECT_LIST (value) FROM TABLE GROUP BY id

Id	value
1	[[‘A’,’B’], [‘A’,’B’], [‘B’,’C’], [‘B’,’C’]]

Question 22

You are looking to process the data based on two variables, one to check if the department is supply chain or check if process flag is set to True

○

if department = “supply chain” | process:

○

if department == “supply chain” or process = TRUE:

○

if department == “supply chain” | process == TRUE:

○

if department == “supply chain” | if process == TRUE:

○

if department == “supply chain” or process:

Question 23

What is the output of below function when executed with input parameters 1, 3 :

```
def check_input(x,y):
    if x < y:
        x= x+1
        if x>y:
            x= x+1
            if x <y:
            x = x+1
    return x
```

○

1

○

2

○

3

○

4

○

5

Question 24

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

○

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

○

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

○

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

○

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

Explanation

The answer is

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

It is always best to use f strings to replace python variables, rather than using string concatenation.

Question 25

you are currently working on creating a spark stream process to read and write in for a one-time micro batch, and also rewrite the existing target table, fill in the blanks to complete the below command sucesfully.

```
spark.table("source_table")
.writeStream
.option("____", “dbfs:/location/silver")
.outputMode("____")
.trigger(Once=____)
.table("target_table")
```

○

checkpointlocation, complete, True

○

targetlocation, overwrite, True

○

checkpointlocation, True, overwrite

○

checkpointlocation, True, complete

○

checkpointlocation, overwrite, True

Question 26

You were asked to write python code to stop all running streams, which of the following command can be used to get a list of all active streams currently running so we can stop them, fill in the blank.

```
for s in _______________:
  s.stop()
```

○

Spark.getActiveStreams()

○

spark.streams.active

○

activeStreams()

○

getActiveStreams()

○

spark.streams.getActive

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, append

○

cloudfiles.format, cloudfiles.schemalocation, checkpointlocation, overwrite

Question 28

Which of the following scenarios are the best fit for AUTO LOADER?

○

Efficiently process new data incrementally from cloud object storage

○

Efficiently move data incrementally from one delta table to another delta table

○

Incrementally process new streaming from data into delta lake

○

Incrementally process new data from relational databases like MySQL

○

Efficiently copy data from data lake location to another data lake location

Question 29

You are asked to setup an AUTO LOADER to process the incoming data, this data arrives in JSON format and get dropped into cloud object storage and you are required to process the data as soon as it arrives in cloud storage, which of the following statements is correct

○

AUTO LOADER is native to DELTA lake it cannot support external cloud object storage

○

AUTO LOADER has to be triggered from an external process when the file arrives in the cloud storage

○

AUTO LOADER needs to be converted to a Structured stream process

○

AUTO LOADER can only process continuous data when stored in DELTA lake

○

AUTO LOADER can support file notification method so it can process data as it arrives

Explanation

Directory listing: Auto Loader identifies new files by listing the input directory.

File notification: Auto Loader can automatically set up a notification service and queue service that subscribe to file events from the input directory.

Choosing between file notification and directory listing modes | Databricks on AWS

Question 30

What is the main difference between bronze and silver?

○

Duplicates are removed in bronze, schema is applied in silver

○

Silver may contain aggregated data

○

Bronze is raw copy of ingested data, silver contains data with schema and optimized format.

○

Bad data is filtered in Bronze, silver is a copy of bronze data

Question 31

What is the main difference between silver and gold?

○

Silver may contain aggregated data

○

Gold may contain aggregated data

○

Data quality checks are applied in gold

○

Silver is a copy of bronze data

○

God is a copy of silver data

Explanation

Medallion Architecture – Databricks

Silver Layer:

1. Reduces data storage complexity, latency, and redundancy
2. Optimizes ETL throughput and analytic query performance
3. Preserves grain of original data (without aggregation)
4. Eliminates duplicate records
5. production schema enforced
6. Data quality checks, quarantine corrupt data

Question 32

What is the main difference between silver and gold?

○

Silver optimized to perform ETL, Gold is optimized query performance

○

Gold is optimized go perform ETL, Silver is optimized for query performance

○

Silver is copy of Bronze, Gold is a copy of Silver

○

Silver is stored in Delta Lake, Gold is stored in memory

○

Silver may contain aggregated data, gold may preserve the granularity of original data

Explanation

Medallion Architecture – Databricks

Gold Layer:

1. Powers Ml applications, reporting, dashboards, ad hoc analytics
2. Refined views of data, typically with aggregations
3. Reduces strain on production systems
4. Optimizes query performance for business-critical data

Question 33

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

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 dat aset and loaded into a quarantine table.

Question 34

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

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.

Question 35

You are asked to debug a job that is taking too long run-on Sunday's, what are steps you are going to take to see the step that is taking longer to run?

○

A notebook activity of job run is only visible when using all-purpose cluster.

○

Under Workflow UI and jobs select job you want to monitor and select the run, notebook activity can be viewed.

○

Enable debug mode in the Jobs to see the output activity of a job, output should be available to view.

○

Once a job is launched, you cannot access the job's notebook activity.

○

Use the compute's spark UI to monitor the job activity.

Question 36

Your colleague was walking you through how a job was setup, but you noticed a warning message that said, “Jobs running on all-purpose cluster are considered all purpose compute", the colleague was not sure why he was getting the warning message, how do you best explain this warning message?

○

All-purpose clusters cannot be used for Job clusters, due to performance issues.

○

All-purpose clusters take longer to start the cluster vs a job cluster

○

All-purpose clusters are less expensive than the job clusters

○

All-purpose clusters are more expensive than the job clusters

○

All-purpose cluster provide interactive messages that can not be viewed in a job

Question 37

Your team has hundreds of jobs running but it is difficult to track cost of each job run, you are asked to provide a recommendation on how to monitor and track cost across various workloads

○

Create jobs in different workspaces, so we can track the cost easily

○

Use Tags, during job creation so cost can be easily tracked

○

Use job logs to monitor and track the costs

○

Use workspace admin reporting

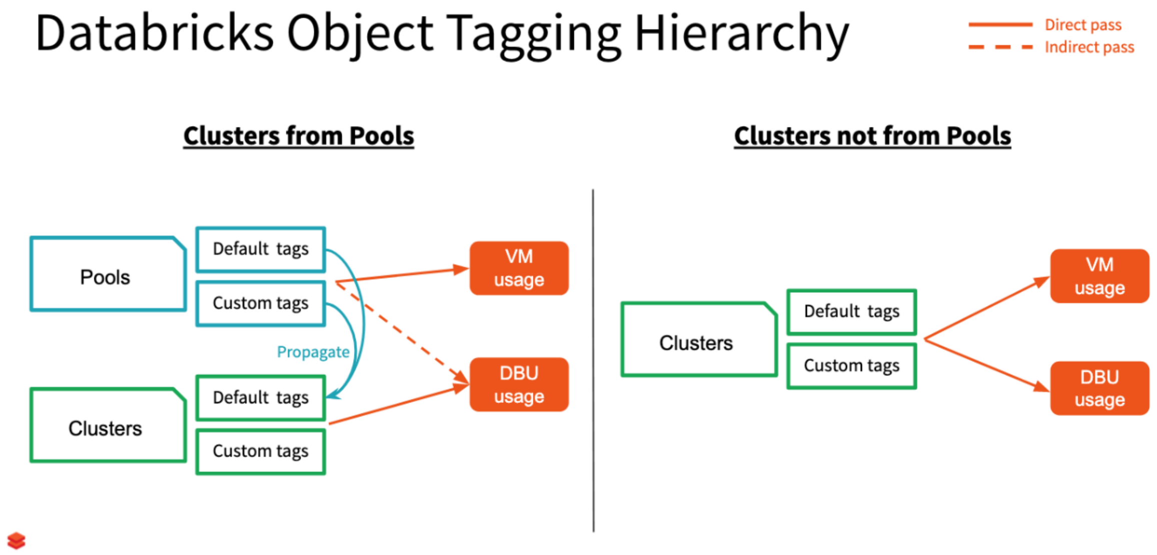
○

Use a single cluster for all the jobs, so cost can be easily tracked

Explanation
The answer is Use Tags, during job creation so cost can be easily tracked

Review below link for more details https://docs.databricks.com/administration-guide/account-settings/usage-detail-tags-aws.html

Here is a view how tags get propagated from pools to clusters and clusters without pools,



Question 38

The sales team has asked the Data engineering team to develop a dashboard that shows sales performance for all stores, but the sales team would like to use the dashboard but would like to select individual store location, which of the following approaches Data Engineering team can use to build this functionality into the dashboard.

○

Use query Parameters which then allow user to choose any location

○

Currently dashboards do not support parameters

○

Use Databricks REST API to create a dashboard for each location

○

Use SQL UDF function to filter the data based on the location

○

Use Dynamic views to filter the data based on the location

Explanation

The answer is

Databricks many types of parameters in the dashboard, a drop-down list can be created based on a query that has a unique list of store locations.

Here is a simple query that takes a parameter for

```
SELECT *  FROM sales WHERE field IN ( {{ Multi Select Parameter }} )
```

Or

```
SELECT *  FROM sales WHERE field =  {{ Single Select Parameter }}
```

Query parameter types

- Text
- Number
- Dropdown List
- Query Based Dropdown List
- Date and Time

Question 39

You are working on dashboard that takes really lot of time to load in the browser, due to each visualization contains lot of data to populate, which of the following approaches can be taken to address this issue?

○

Increase size of the SQL endpoint cluster

○

Increase the scale of maximum range of SQL endpoint cluster

○

Use Databricks SQL Query filter to limit the amount of data in each visualization

○

Remove data from Delta Lake

○

Use Delta cache to store the intermediate results

Explanation

A query filter lets you interactively reduce the amount of data shown in a visualization, similar to query parameter but with a few key differences. A query filter limits data *after* it has been loaded into your browser. This makes filters ideal for smaller datasets and environments where query executions are time-consuming, rate-limited, or costly.
This query filter is different from than filter that needs to be applied at the data level, this filter is at the visualization level so you can toggle how much data you want to see.

```
SELECT action AS `action::filter`, COUNT(0) AS "actions count"
FROM events
GROUP BY action
```

When queries have filters you can also apply filters at the dashboard level. Select the Use Dashboard Level Filters checkbox to apply the filter to all queries.

Dashboard filters

Query filters | Databricks on AWS

Question 40

One of the queries in the Dashboard takes a long time to refresh, which of the below steps can be taken to identify the root cause of this issue?

○

Restart the SQL endpoint

○

Select the SQL endpoint cluster, spark UI, SQL tab to see the execution plan and time spent in each step

○

Run optimize and Z ordering

○

Change the Spot Instance Policy from “Cost optimized” to “Reliability Optimized.”

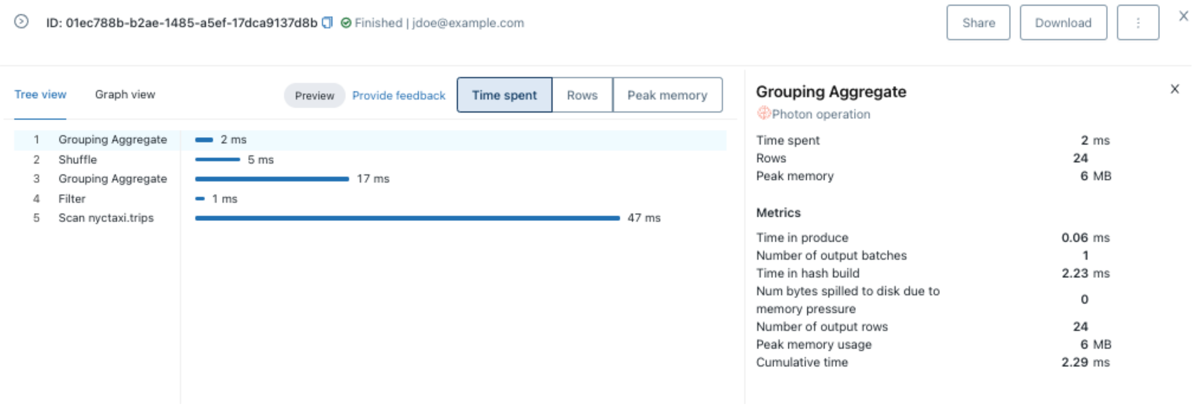
○

Use Query History, to view queries and select query, and check query profile to time spent in each step

Explanation

The answer is, Use Query History, to view queries and select query, and check the query profile to see time spent in each step.

Here is the view of the query profile, for more info use the link https://docs.microsoft.com/en-us/azure/databricks/sql/admin/query-profile



ID: 01ec788b-b2ae-1485-a5ef-17dca9137d8b Finished | jdoe@example.com

1	Grouping Aggregate	2 ms
2	Shuffle	5 ms
3	Grouping Aggregate	17 ms
4	Filter	1 ms
5	Scan nyctaxi.trips	47 ms

Grouping Aggregate

Photon operation

Time spent	2 ms
Rows	24
Peak memory	6 MB

Metrics

Time in produce	0.06 ms
Number of output batches	1
Time in hash build	2.23 ms
Num bytes spilled to disk due to memory pressure	0
Number of output rows	24
Peak memory usage	6 MB
Cumulative time	2.29 ms

Question 41

A SQL Dashboard was built for the supply chain team to monitor the inventory and product orders, but all of the timestamps displayed on the dashboards are showing in UTC format, so they requested to change the time zone to the location of New York. How would you approach resolving this issue?

○

Move the workspace from Central US zone to East US Zone

○

Change the timestamp on the delta tables to America/New_York format

○

Change the spark configuration of SQL endpoint to format the timestamp to America/New_York

○

Under SQL Admin Console, set the SQL configuration parameter time zone to America/New_York

○

Add SET Timezone = America/New_York on every of the SQL queries in the dashboard.

Explanation

The answer is, Under SQL Admin Console, set the SQL configuration parameter time zone to America/New_York

Here are steps you can take this to configure, so the entire dashboard is changed without changing individual queries

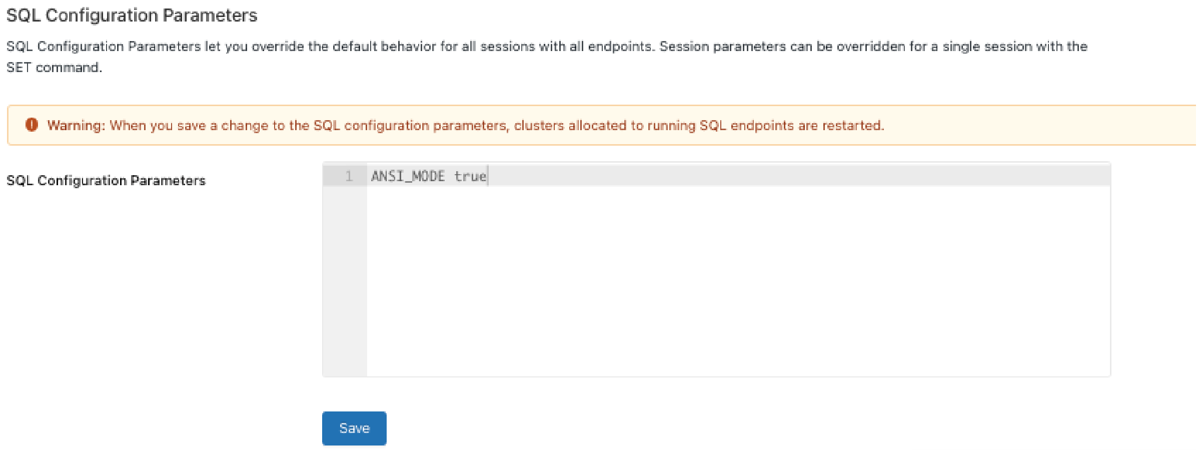
Configure SQL parameters

To configure all warehouses with SQL parameters:

Click ⚙ Settings at the bottom of the sidebar and select SQL Admin Console.

Click the SQL Warehouse Settings tab.

In the SQL Configuration Parameters textbox, specify one key-value pair per line. Separate the name of the parameter from its value using a space. For example, to enable `ANSI_MODE`:



SQL Configuration Parameters

SQL Configuration Parameters let you override the default behavior for all sessions with all endpoints. Session parameters can be overridden for a single session with the SET command.

Warning: When you save a change to the SQL configuration parameters, clusters allocated to running SQL endpoints are restarted.

SQL Configuration Parameters

```
ANSI_MODE true
```

Similarly, we can add a line in the SQL Configuration parameters

timezone America/New_York

[SQL configuration parameters | Databricks on AWS]()

Question 42

Which of the following technique can be used to fine-grained access control to rows and columns based on access?

○

Use Unity catalog to grant access to rows and columns

○

Row and column access control lists

○

Use dynamic view functions

○

Data access control lists

○

Dynamic Access control lists with Unity Catalog

Explanation

The answer is, Use dynamic view functions.

Here is an example that limits access to rows based on the user being part managers group, in the below view if a user is not a part of the manager's group you can only see rows where the total amount is <= 1000000

Dynamic view function to filter rows

```
CREATE VIEW sales_redacted AS
SELECT user_id, country, product, total
FROM sales_raw
WHERE CASE WHEN is_member('managers') THEN TRUE ELSE total <= 1000000 END;
```

Dynamic view function to hide a column

```
CREATE VIEW sales_redacted AS
SELECT user_id,
       CASE WHEN is_member('auditors') THEN email ELSE 'REDACTED' END AS email,
       country,
       product,
       total
FROM sales_raw
```

Please review below for more details
https://docs.microsoft.com/en-us/azure/databricks/security/access-control/table-acls/object-privileges#dynamic-view-functions

Question 43

Unity catalog helps you manage the below resources in Databricks at account level

○

Tables

○

ML Models

○

Dashboards

○

Catalogs

○

All of the above

Question 44

A newly joined team member John Smith in the Marketing team who currently has access read access to sales tables but does not have access to delete rows from the table, which of the following commands help you accomplish this?

○

```
GRANT USAGE ON TABLE table_name TO john.smith@marketing.com
```

○

```
GRANT DELETE ON TABLE table_name TO john.smith@marketing.com
```

○

```
GRANT DELETE TO TABLE table_name ON john.smith@marketing.com
```

○

```
GRANT MODIFY TO TABLE table_name ON john.smith@marketing.com
```

○

```
GRANT MODIFY ON TABLE table_name TO john.smith@marketing.com
```

Explanation

The answer is `GRANT MODIFY ON TABLE table_name TO john.smith@marketing.com`

· SELECT: gives reada 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

Question 45

Kevin is the owner of both the sales table and regional_sales_vw view which uses the sales table as underlying source for the data, and Kevin is looking to grant select privilege on the view regional_sales_vw to one of newly joined team members Steven. Which of the following is a true statement?

○

Kevin can not grant access to Steven since he does not have security admin privilege

○

Kevin although is the owner but does not have ALL PRIVILEGES permission

○

Kevin can grant access to the view, because he is the owner of the view and the underlying table

○

Kevin can not grant access to Steven since he does have workspace admin privilege

○

Steve will also require SELECT access on the underlying table

Explanation

The answer is, Kevin can grant access to the view, because he is the owner of the view and the underlying table,

Ownership determines whether or not you can grant privileges on derived objects to other users, a user who creates a schema, table, view, or function becomes its owner. The owner is granted all privileges and can grant privileges to other users