# Databricks.Databricks-Certified-Professional-Data-Engineer.v2023-05-23.q104

| | |
|---|---|
| **Exam Code:** | Databricks-Certified-Professional-Data-Engineer |
| **Exam Name:** | Databricks Certified Professional Data Engineer Exam |
| **Certification Provider:** | Databricks |
| **Free Question Number:** | 104 |
| **Version:** | v2023-05-23 |
| **# of views:** | 1105 |
| **# of Questions views:** | 1040 |
| https://www.freeqas.com/qa/Databricks/Databricks-Certified-Professional-Data-Engineer/Databricks.Databricks-Certified-Professional-Data-Engineer.v2023-05-23.q104.html | |

**NEW QUESTION: 1**
You were asked to setup a new all-purpose cluster, but the cluster is unable to start which of the following steps do you need to take to identify the root cause of the issue and the reason why the cluster was unable to start?
**A.** Check the cluster driver logs
**B.** Check the cluster event logs
(Correct)
**C.** Workspace logs
**D.** Storage account
**E.** Data plane
**Answer: (SHOW ANSWER)**
Explanation
Cluster event logs are very useful, to identify issues pertaining to cluster availability. Cluster may not start due to resource limitations or issues with the cloud providers.
Some of the common issues include a subnet for compute VM reaching its limits or exceeding the subscription or cloud account CPU quota limit.
Here is an example where the cluster did not start due to subscription reaching the quota limit on a certain type of cpu cores for a VM type.
Graphical user interface, text, application, email Description automatically generated


Click on event logs

Graphical user interface, text, application, email Description automatically generated


Click on the message to see the detailed error message on why the cluster did not start.

Graphical user interface, text, application, email Description automatically generated


**NEW QUESTION: 2**

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

**A.** Replaces a traditional data lake

**B.** Efficient storage and querying of full, unprocessed history of data

**C.** Eliminates duplicate data, quarantines bad data

**D.** Refined views with aggregated data

**E.** Optimized query performance for business-critical data

**Answer: (SHOW ANSWER)**

Explanation

Medallion Architecture - Databricks

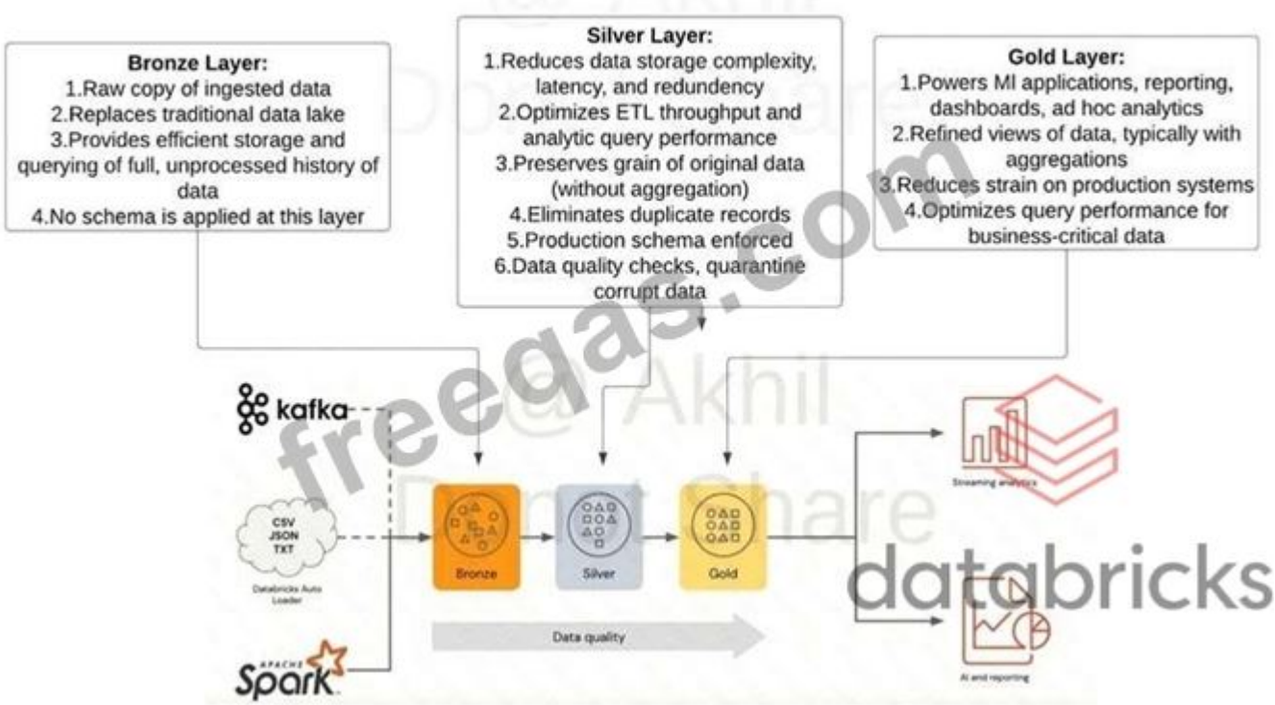
Silver Layer:

1. Reduces data storage complexity, latency, and redundency
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

Exam focus: Please review the below image and understand the role of each layer(bronze, silver, gold) in medallion architecture, you will see varying questions targeting each layer and its purpose.

Sorry I had to add the watermark some people in Udemy are copying my content.

A diagram of a house Description automatically generated with low confidence



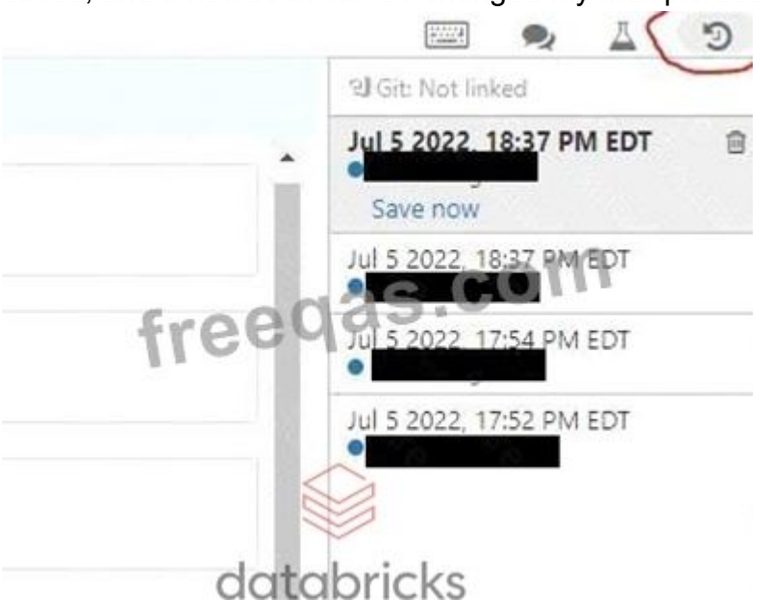

**NEW QUESTION: 3**

You are currently working with the second team and both teams are looking to modify the same notebook, you noticed that the second member is copying the notebooks to the personal folder to edit and replace the collaboration notebook, which notebook feature do you recommend to make the process easier to collaborate.

**A.** Databricks notebooks should be copied to a local machine and setup source control lo-cally to version the notebooks

**B.** Databricks notebooks support automatic change tracking and versioning

**C.** Databricks Notebooks support real-time coauthoring on a single notebook
**D.** Databricks notebooks can be exported into dbc archive files and stored in data lake
**E.** Databricks notebook can be exported as HTML and imported at a later time
**Answer: (SHOW ANSWER)**
Explanation
Answer is Databricks Notebooks support real-time coauthoring on a single notebook Every change is saved, and a notebook can be changed my multiple users.

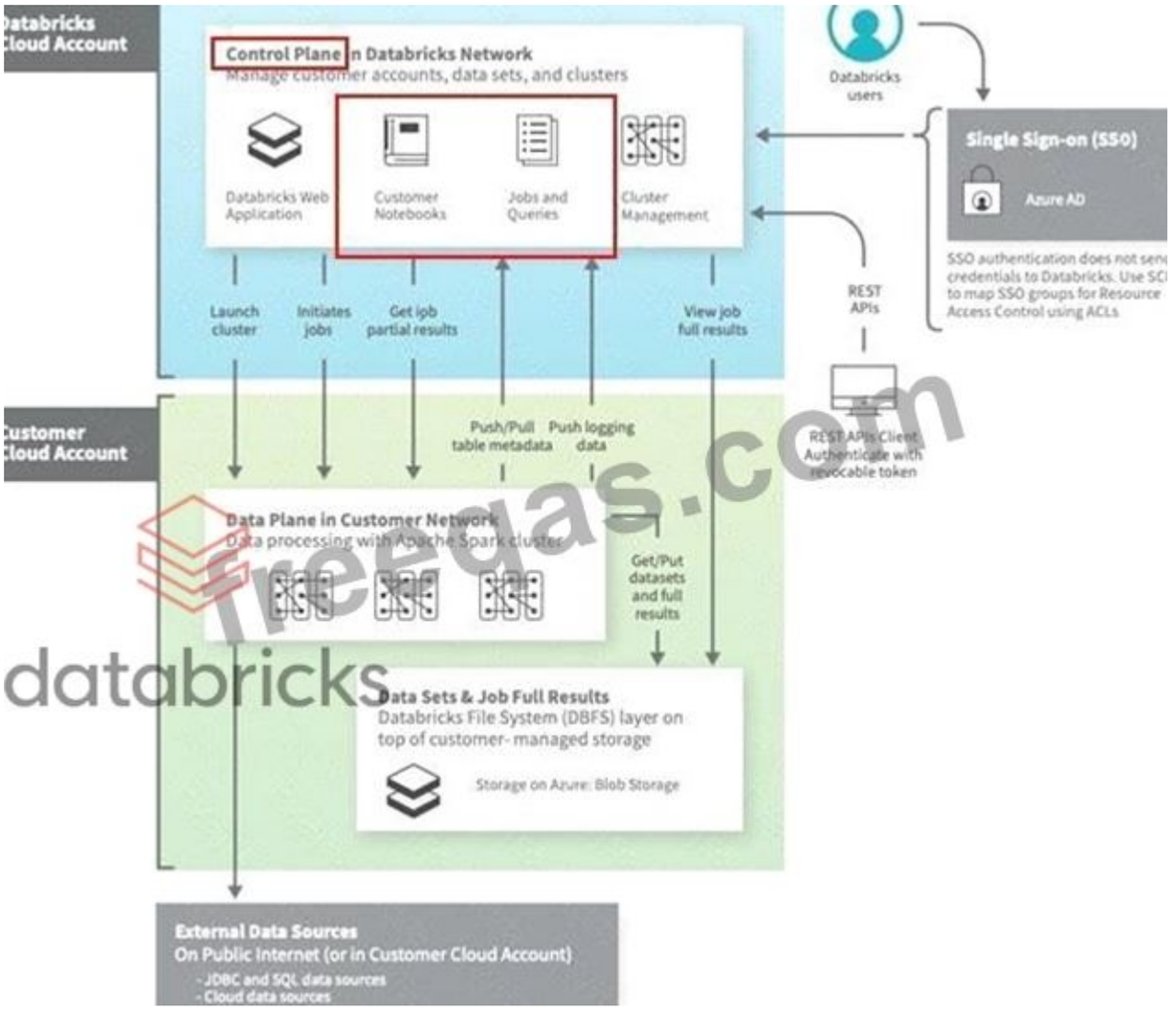



**NEW QUESTION: 4**
The current ELT pipeline is receiving data from the operations team once a day so you had setup an AUTO LOADER process to run once a day using trigger (Once = True) and scheduled a job to run once a day, operations team recently rolled out a new feature that allows them to send data every 1 min, what changes do you need to make to AUTO LOADER to process the data every 1 min.
**A.** Change AUTO LOADER trigger to ("1 minute")
**B.** Convert AUTO LOADER to structured streaming
**C.** Setup a job cluster run the notebook once a minute
**D.** Enable stream processing
**E.** Change AUTO LOADER trigger to .trigger(ProcessingTime = "1 minute")
**Answer: E (LEAVE A REPLY)**

**NEW QUESTION: 5**
When writing streaming data, Spark's structured stream supports the below write modes
**A.** Append, Delta, Complete
**B.** Delta, Complete, Continuous
**C.** Append, Complete, Update
**D.** Complete, Incremental, Update
**E.** Append, overwrite, Continuous
**Answer: (SHOW ANSWER)**

Explanation
The answer is Append, Complete, Update
*Append mode (default) - This is the default mode, where only the new rows added to the Result Table since the last trigger will be outputted to the sink. This is supported for only those queries where rows added to the Result Table is never going to change. Hence, this mode guarantees that each row will be output only once (assuming fault-tolerant sink). For example, queries with only select, where, map, flatMap, filter, join, etc. will support Append mode.
*Complete mode - The whole Result Table will be outputted to the sink after every trigger. This is supported for aggregation queries.
*Update mode - (Available since Spark 2.1.1) Only the rows in the Result Table that were updated since the last trigger will be outputted to the sink. More information to be added in future releases.

**NEW QUESTION: 6**
Which of the following locations in the Databricks product architecture hosts the notebooks and jobs?
**A.** Data plane
**B.** Control plane
**C.** Databricks Filesystem
**D.** JDBC data source
**E.** Databricks web application
**Answer: (SHOW ANSWER)**
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 pane.
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.
Timeline Description automatically generated






**NEW QUESTION: 7**

Which of the following table constraints that can be enforced on Delta lake tables are supported?

**A.** Primary key, foreign key, Not Null, Check Constraints

**B.** Primary key, Not Null, Check Constraints

**C.** Default, Not Null, Check Constraints

**D.** Not Null, Check Constraints

**E.** Unique, Not Null, Check Constraints

**Answer: (SHOW ANSWER)**

Explanation

The answer is Not Null, Check Constraints

https://docs.microsoft.com/en-us/azure/databricks/delta/delta-constraints

* CREATE TABLE events( id LONG,

* date STRING,

* location STRING,

* description STRING

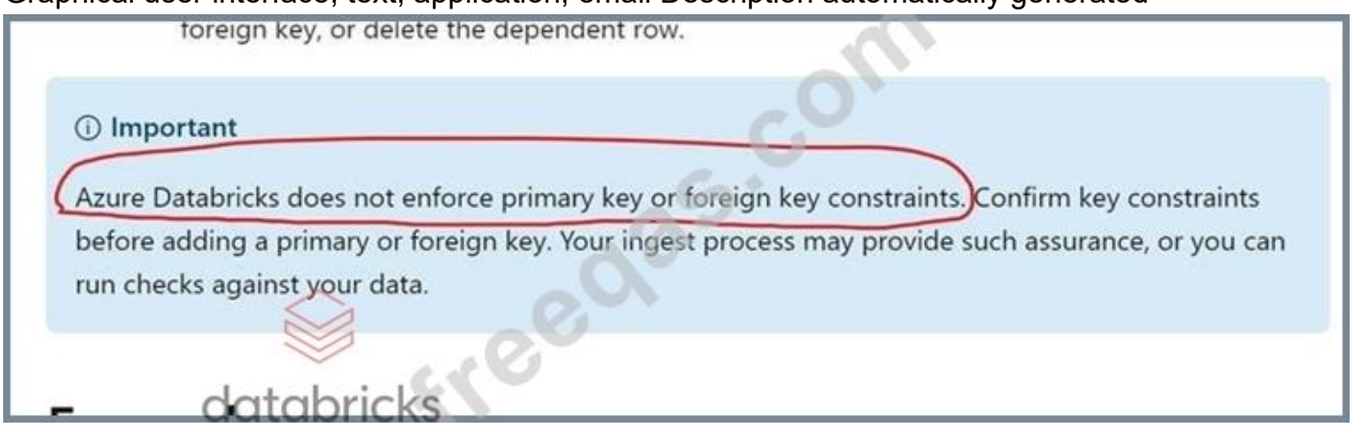
* ) USING DELTA;

ALTER TABLE events CHANGE COLUMN id SET NOT NULL;

ALTER TABLE events ADD CONSTRAINT dateWithinRange CHECK (date > '1900-01-01'); Note: Databricks as of DBR 11.1 added support for Primary Key and Foreign Key when Unity Catalog is enabled but this is for information purposes only these are not actually enforced. You may ask then why are we defining these if they are not enforced, so especially these information constraints are very helpful if you have a BI tool that can benefit from knowing the relationship between the tables, so it will be easy when creating reports/dashboards or understanding the data model when using any Data modeling tool.

Primary and Foreign Key

Graphical user interface, text, application, email Description automatically generated




**NEW QUESTION: 8**

Which of the following SQL keywords can be used to append new rows to an existing Delta table?

**A.** DELETE

**B.** UNION

**C.** INSERT INTO

**D.** COPY

**E.** UPDATE

**Answer: (SHOW ANSWER)**

**NEW QUESTION: 9**

Which of the following features of data lakehouse can help you meet the needs of both workloads?

**A.** Data lakehouse requires very little data modeling.

**B.** Data lakehouse combines compute and storage for simple governance.

**C.** Data lakehouse provides autoscaling for compute clusters.

**D.** Data lakehouse can store unstructured data and support ACID transactions.

**E.** Data lakehouse fully exists in the cloud.

**Answer: D (LEAVE A REPLY)**

Explanation

The answer is A data lakehouse stores unstructured data and is ACID-compliant,

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.



**NEW QUESTION: 10**

You noticed that a team member started using an all-purpose cluster to develop a notebook and used the same all-purpose cluster to set up a job that can run every 30 mins so they can update un-derlying tables which are used in a dashboard. What would you recommend for reducing the overall cost of this approach?

**A.** Reduce the size of the cluster

**B.** Reduce the number of nodes and enable auto scale

**C.** Enable auto termination after 30 mins

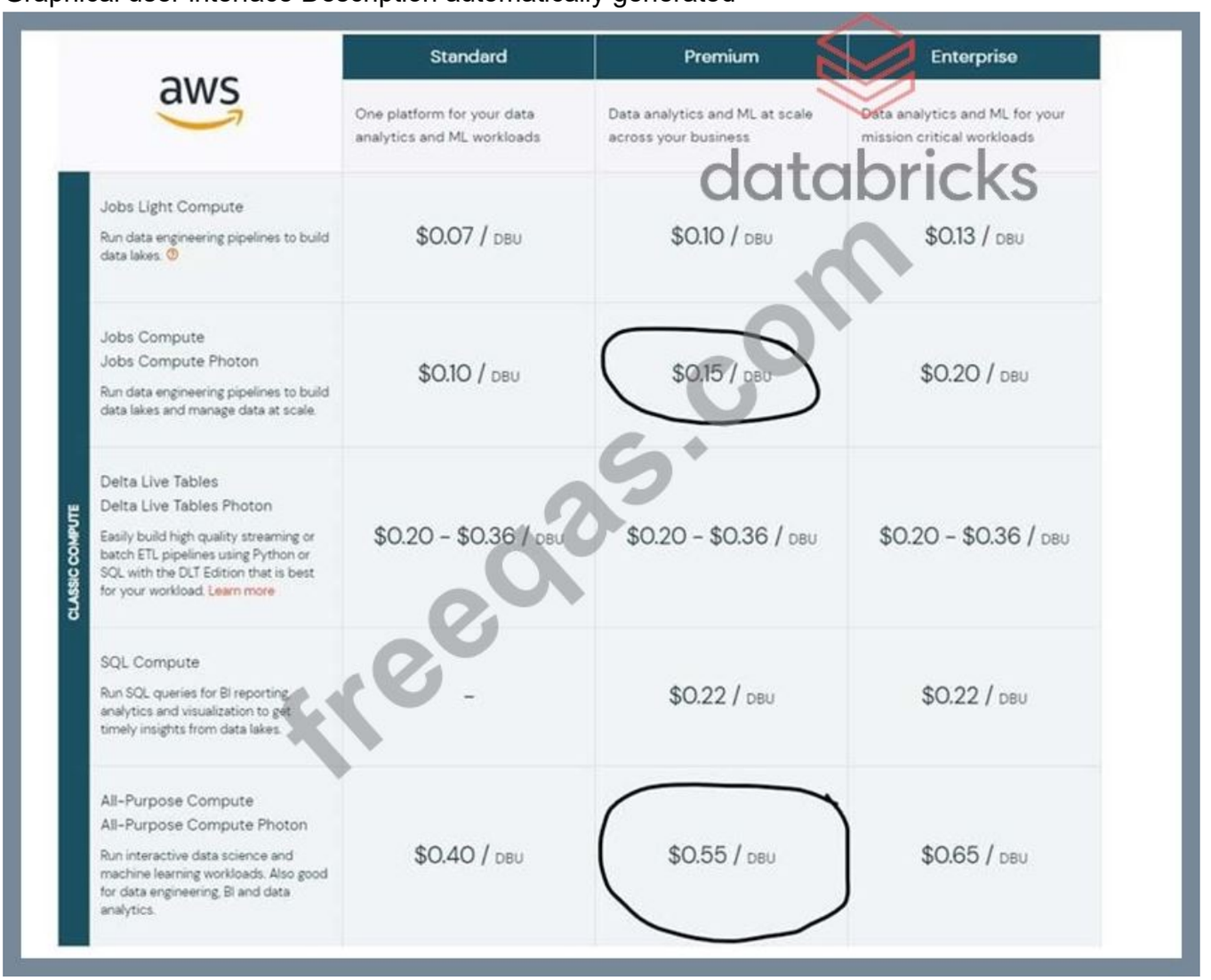
**D.** Change the cluster all-purpose to job cluster when scheduling the job

**E.** Change the cluster mode from all-purpose to single-mode

**Answer: D (LEAVE A REPLY)**

Explanation

While using an all-purpose cluster is ok during development but anytime you don't need to interact with a notebook, especially for a scheduled job it is less expensive to use a job cluster. Using an all-purpose cluster can be twice as expensive as a job cluster.
Please note: The compute cost you pay the cloud provider for the same cluster type and size be-tween an all-purpose cluster and job cluster is the same the only difference is the DBU cost.
The total cost of cluster = Total cost of VM compute(Azure or AWS or GCP) + Cost per DBU The per DBU cost varies between all-purpose and Job Cluster Here is the recent cost estimate from AWS between Jobs Cluster and all-purpose Cluster, for jobs compute its
$0.15 cents per DBU v$0.55 cents per DBU for all-purpose
Graphical user interface Description automatically generated




How do I check how much the DBU cost for my cluster?
When you click on an exister cluster or when you look at the cluster details you will see this in the top right corner Graphical user interface, text, application, email Description automatically generated


**NEW QUESTION: 11**
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
**A.** Setup a second job ahead of first job to start the cluster, so the cluster is ready with re-sources when the job starts
**B.** Use All purpose cluster instead to reduce cluster start up time
**C.** Reduce the size of the cluster, smaller the cluster size shorter it takes to start the clus-ter
**D.** Use cluster pools to reduce the startup time of the jobs
**E.** Use SQL endpoints to reduce the startup time
**Answer: (SHOW ANSWER)**
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

**NEW QUESTION: 12**
Data engineering team is required to share the data with Data science team and both the teams are using different workspaces in the same organizationwhich of the following techniques can be used to simplify sharing data across?
*Please note the question is asking how data is shared within an organization across multiple workspaces.
**A.** Data Sharing
**B.** Unity Catalog

**C.** DELTA lake

**D.** Use a single storage location

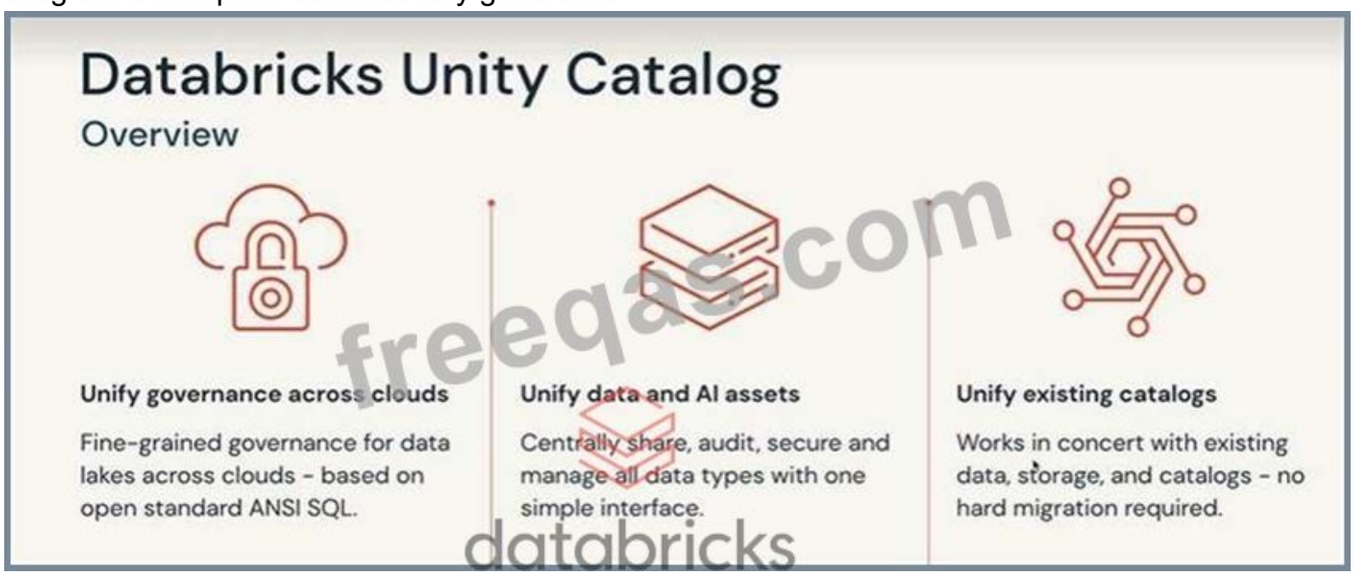
**E.** DELTA LIVE Pipelines

**Answer: B (LEAVE A REPLY)**

Explanation

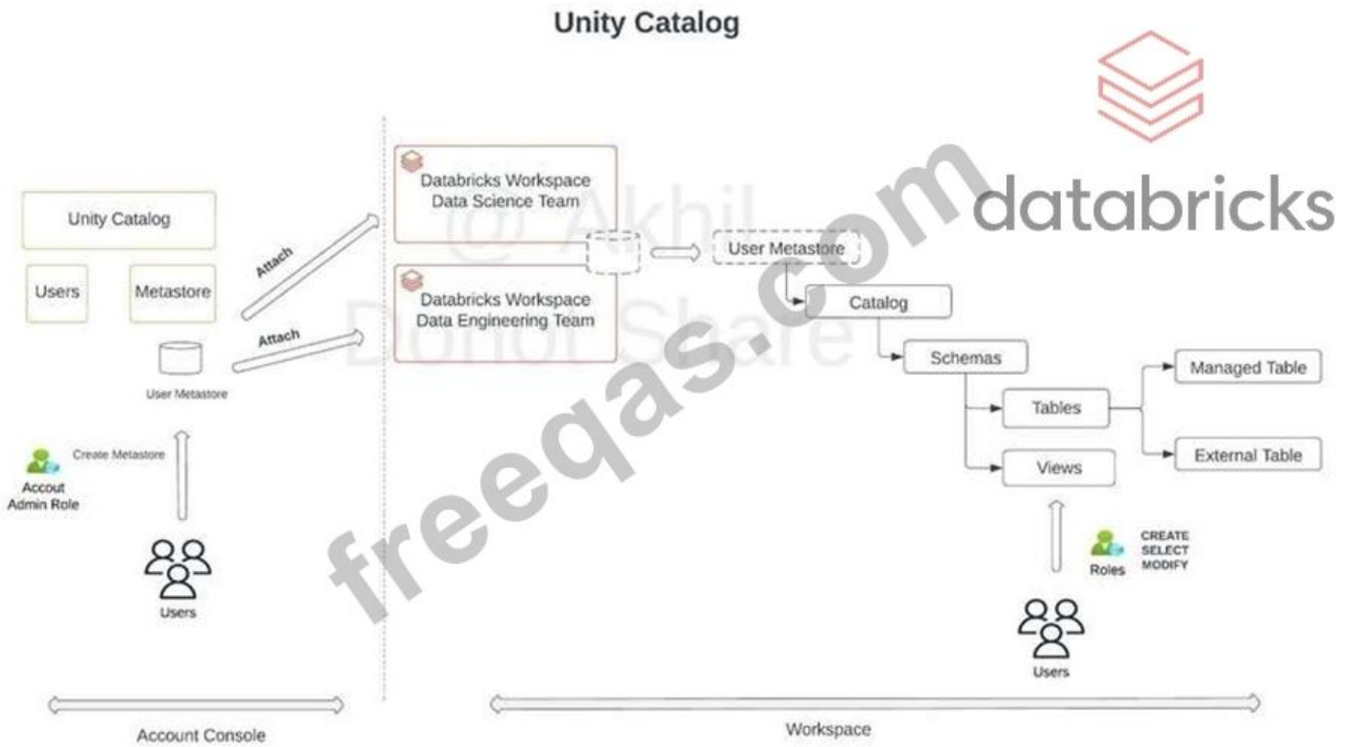
The answer is the Unity catalog.

Diagram Description automatically generated




Unity Catalog works at the Account level, it has the ability to create a meta store and attach that meta store to many workspaces see the below diagram to understand how Unity Catalog Works, as you can see a metastore can now be shared with both workspaces using Unity Catalog, prior to Unity Catalog the options was to use single cloud object storage manually mount in the second databricks workspace, and you can see here Unity Catalog really simplifies that.

Diagram Description automatically generated with medium confidence






sorry for the inconvenience watermark was added because other people on Udemy are copying my questions and images.
duct features
https://databricks.com/product/unity-catalog

**NEW QUESTION: 13**
Which of the following statements can successfully read the notebook widget and pass the python variable to a SQL statement in a Python notebook cell?
**A.** 1.order_date = dbutils.widgets.get("widget_order_date")
2.
3.spark.sql(f"SELECT * FROM sales WHERE orderDate = '${order_date }' ")
**B.** 1.order_date = dbutils.widgets.get("widget_order_date")
2.
3.spark.sql(f"SELECT * FROM sales WHERE orderDate = 'f{order_date }'")
**C.** 1.order_date = dbutils.widgets.get("widget_order_date")
2.
3.spark.sql(f"SELECT * FROM sales WHERE orderDate = '{order_date}' ")
(Correct)
**D.** 1.order_date = dbutils.widgets.get("widget_order_date")
2.
3.spark.sql("SELECT * FROM sales WHERE orderDate = order_date")
**E.** 1.order_date = dbutils.widgets.get("widget_order_date")
2.

3.spark.sql(f"SELECT * FROM sales WHERE orderDate = 'order_date' ")
**Answer: C (LEAVE A REPLY)**

**NEW QUESTION: 14**
Which of the following python statement can be used to replace the schema name and table name in the query statement?
**A.** 1.table_name = "sales"
2.schema_name = "bronze"
3.query = f"select * from schema_name.table_name"
**B.** 1.table_name = "sales"
2.schema_name = "bronze"
3.query = "select * from {schema_name}.{table_name}"
**C.** 1.table_name = "sales"
2.schema_name = "bronze"
3.query = f"select * from { schema_name}.{table_name}"
**D.** 1.table_name = "sales"
2.schema_name = "bronze"
3.query = f"select * from + schema_name +"."+table_name"
**Answer: C (LEAVE A REPLY)**
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/

**NEW QUESTION: 15**
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 in-crementally, 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.
1.spark.readStream
2..format("cloudfiles")
3..option("_____","csv)
4..option("_____", 'dbfs:/location/checkpoint/')
5..load(data_source)
6..writeStream
7..option("_____"," dbfs:/location/checkpoint/')
8..option("_____", "true")

9..table(table_name))
**A.** format, checkpointlocation, schemalocation, overwrite
**B.** cloudfiles.format, checkpointlocation, cloudfiles.schemalocation, overwrite
**C.** cloudfiles.format, cloudfiles.schemalocation, checkpointlocation, mergeSchema
**D.** cloudfiles.format, cloudfiles.schemalocation, checkpointlocation, overwrite
**E.** cloudfiles.format, cloudfiles.schemalocation, checkpointlocation, append
**Answer: C (LEAVE A REPLY)**
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
1.spark.readStream
2..format("cloudfiles") # Returns a stream data source, reads data as it arrives based on the trigger.
3..option("cloudfiles.format","csv") # Format of the incoming files
4..option("cloudfiles.schemalocation", "dbfs:/location/checkpoint/") The location to store the inferred schema and subsequent changes
5..load(data_source)
6..writeStream
7..option("checkpointlocation","dbfs:/location/checkpoint/") # The location of the stream's checkpoint
8..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.
9..table(table_name)) # target table

**NEW QUESTION: 16**
A data engineer has a Job with multiple tasks that runs nightly. One of the tasks unexpectedly fails during 10
percent of the runs.
Which of the following actions can the data engineer perform to ensure the Job completes each night while
minimizing compute costs?
**A.** They can set up the Job to run multiple times ensuring that at least one will complete
**B.** They can observe the task as it runs to try and determine why it is failing
**C.** They can institute a retry policy for the entire Job
**D.** They can institute a retry policy for the task that periodically fails
**E.** They can utilize a Jobs cluster for each of the tasks in the Job
**Answer: (SHOW ANSWER)**

**NEW QUESTION: 17**
Which of the statement is correct about the cluster pools?
**A.** Cluster pools allow you to create a cluster
**B.** Cluster pools allow you to save time when starting a new cluster
**C.** Cluster pools are used to share resources among multiple teams
**D.** Cluster pools allow you to perform load balancing
**E.** Cluster pools allow you to have all the nodes in the cluster from single physical server rack
**Answer: B (LEAVE A REPLY)**

**NEW QUESTION: 18**
A denote the event 'student is female' and let B denote the event 'student is French'. In a class of 100 students
suppose 60 are French, and suppose that 10 of the French students are females. Find the probability that if I
pick a French student, it will be a girl, that is, find P(A|B).
**A.** 1/3
**B.** 2/3
**C.** 1/6
**D.** 2/6
**Answer: (SHOW ANSWER)**
Explanation
Since 10 out of 100 students are both French and female, then
P(AandB)=10100
Also. 60 out of the 100 students are French, so
P(B)=60100
So the required probability is:
P(A|B)=P(AandB)P(B)=10/10060/100=16

**NEW QUESTION: 19**
Which method is used to solve for coefficients bO, b1, ... bn in your linear regression model:
$Y = b0 + b1x1+b2x2+ .... +bnxn$
**A.** Apriori Algorithm
**B.** Ridge and Lasso

**C.** Ordinary Least squares
**D.** Integer programming
**Answer: (SHOW ANSWER)**
Explanation : RY = b0 + b1x1+b2x2+ .... +bnxn
In the linear model, the bi's represent the unknown p parameters. The estimates for these unknown parameters
are chosen so that, on average, the model provides a reasonable estimate of a person's income based on age
and education. In other words, the fitted model should minimize the overall error between the linear model and
the actual observations. Ordinary Least Squares (OLS) is a common technique to estimate the parameters

**NEW QUESTION: 20**
A data engineer needs to create a database called customer360 at the loca-
tion /customer/customer360. The
data engineer is unsure if one of their colleagues has already created the database.
Which of the following commands should the data engineer run to complete this task?
**A.** CREATE DATABASE IF NOT EXISTS customer360 LOCATION '/customer/customer360';
**B.** CREATE DATABASE IF NOT EXISTS customer360 DELTA LOCATION '/customer/customer360';
**C.** CREATE DATABASE customer360 DELTA LOCATION '/customer/customer360';
**D.** CREATE DATABASE customer360 LOCATION '/customer/customer360';
**E.** CREATE DATABASE IF NOT EXISTS customer360;
**Answer: (SHOW ANSWER)**

**NEW QUESTION: 21**
When using the complete mode to write stream data, how does it impact the target table?
**A.** Entire stream waits for complete data to write
**B.** Stream must complete to write the data
**C.** Target table cannot be updated while stream is pending
**D.** Target table is overwritten for each batch
**E.** Delta commits transaction once the stream is stopped
**Answer: D (LEAVE A REPLY)**
Explanation
The answer is Target table is overwritten for each batch
Complete mode - The whole Result Table will be outputted to the sink after every trigger. This is supported for aggregation queries

**NEW QUESTION: 22**
Which of the following section in the UI can be used to manage permissions and grants to tables?
**A.** User Settings

**B.** Admin UI
**C.** Workspace admin settings
**D.** User access control lists
**E.** Data Explorer
**Answer: E (LEAVE A REPLY)**
Explanation
The answer is Data Explorer

**NEW QUESTION: 23**
You are asked to debug a databricks job that is taking too long to run on Sunday's, what are the steps you are going to take to identify the step that is taking longer to run?
**A.** A notebook activity of job run is only visible when using all-purpose cluster.
**B.** Under Workflow UI and jobs select job you want to monitor and select the run, notebook activity can be viewed.
**C.** Enable debug mode in the Jobs to see the output activity of a job, output should be available to view.
**D.** Once a job is launched, you cannot access the job's notebook activity.
**E.** Use the compute's spark UI to monitor the job activity.
**Answer: (SHOW ANSWER)**
Explanation
The answer is, Under Workflow UI and jobs select job you want to monitor and select the run, notebook activity can be viewed.
You have the ability to view current active runs or completed runs, once you click the run you can see the A picture containing graphical user interface Description automatically generated

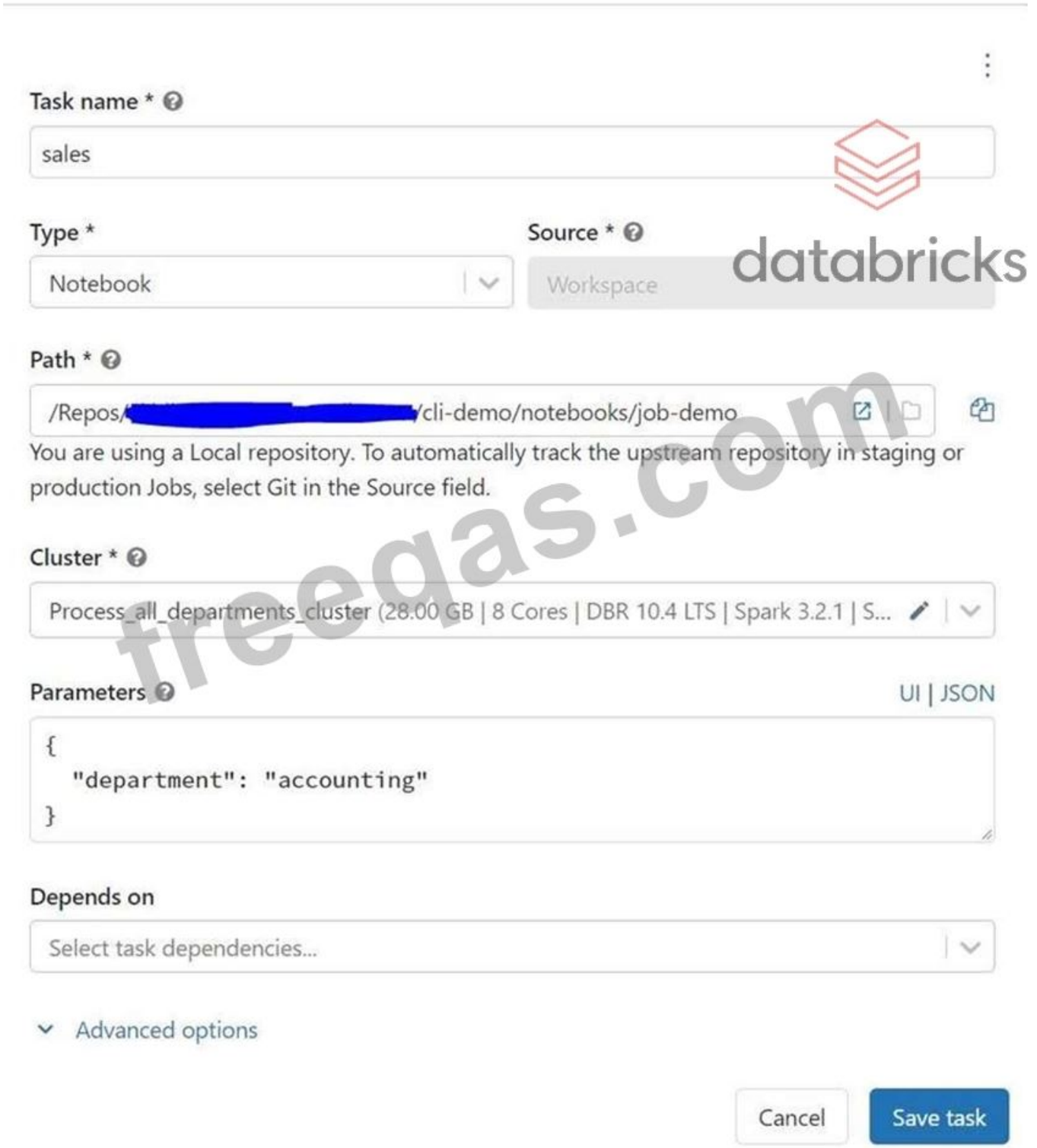

Click on the run to view the notebook output
Graphical user interface, text, application, email Description automatically generated



**NEW QUESTION: 24**
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.
**A.** ALTER TABLE table_name set TBLPROPERTIES (del-ta.deletedFileRetentionDuration= 'interval 365 days')
**B.** MODIFY TABLE table_name set TBLPROPERTY (delta.maxRetentionDays = 'inter-val 365 days')
**C.** ALTER TABLE table_name set EXENDED TBLPROPERTIES (del-ta.deletedFileRetentionDuration= 'interval 365 days')
**D.** ALTER TABLE table_name set EXENDED TBLPROPERTIES (delta.vaccum.duration= 'interval 365 days')
**Answer: A (LEAVE A REPLY)**
Explanation
1.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

**NEW QUESTION: 25**
You are tasked to set up a set notebook as a job for six departments and each department can run the task parallelly, the notebook takes an input parameter dept number to process the data by department, how do you go about to setup this up in job?

**A.** Use a single notebook as task in the job and use dbutils.notebook.run to run each note-book with parameter in a different cell
**B.** A task in the job cannot take an input parameter, create six notebooks with hardcoded dept number and setup six tasks with linear dependency in the job
**C.** A task accepts key-value pair parameters, creates six tasks pass department number as parameter foreach task with no dependency in the job as they can all run in parallel.
(Correct)
**D.** A parameter can only be passed at the job level, create six jobs pass department number to each job with linear job dependency
**E.** A parameter can only be passed at the job level, create six jobs pass department number to each job with no job dependency
**Answer: (SHOW ANSWER)**
Explanation
Here is how you setup
Create a single job and six tasks with the same notebook and assign a different parameter for each task , Graphical user interface, text, application, email Description automatically generated




All tasks are added in a single job and can run parallel either using single shared cluster or with individual clusters.

Graphical user interface, application, Teams Description automatically generated


**NEW QUESTION: 26**
Kevin is the owner of both the sales table and regional_sales_vw view which uses the sales table as the 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?
**A.** Kevin can not grant access to Steven since he does not have security admin privilege
**B.** Kevin although is the owner but does not have ALL PRIVILEGES permission
**C.** Kevin can grant access to the view, because he is the owner of the view and the under-lying table
**D.** Kevin can not grant access to Steven since he does have workspace admin privilege
**E.** Steve will also require SELECT access on the underlying table
**Answer: (SHOW ANSWER)**
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

**NEW QUESTION: 27**
John Smith is a newly joined team member 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?
**A.** GRANT USAGE ON TABLE table_name TO john.smith@marketing.com
**B.** GRANT DELETE ON TABLE table_name TO john.smith@marketing.com
**C.** GRANT DELETE TO TABLE table_name ON john.smith@marketing.com
**D.** GRANT MODIFY TO TABLE table_name ON john.smith@marketing.com
**E.** GRANT MODIFY ON TABLE table_name TO john.smith@marketing.com
**Answer: E (LEAVE A REPLY)**
Explanation
The answer is GRANT MODIFY ON TABLE table_name TO john.smith@marketing.com , please note INSERT, UPDATE, and DELETE are combined into one role called MODIFY.
Below are the list of privileges that can be granted to a user or a group, SELECT: gives read access to an object.
CREATE: gives the ability to create an object (for example, a table in a schema).
MODIFY: gives the ability to add, delete, and modify data to or from an object.

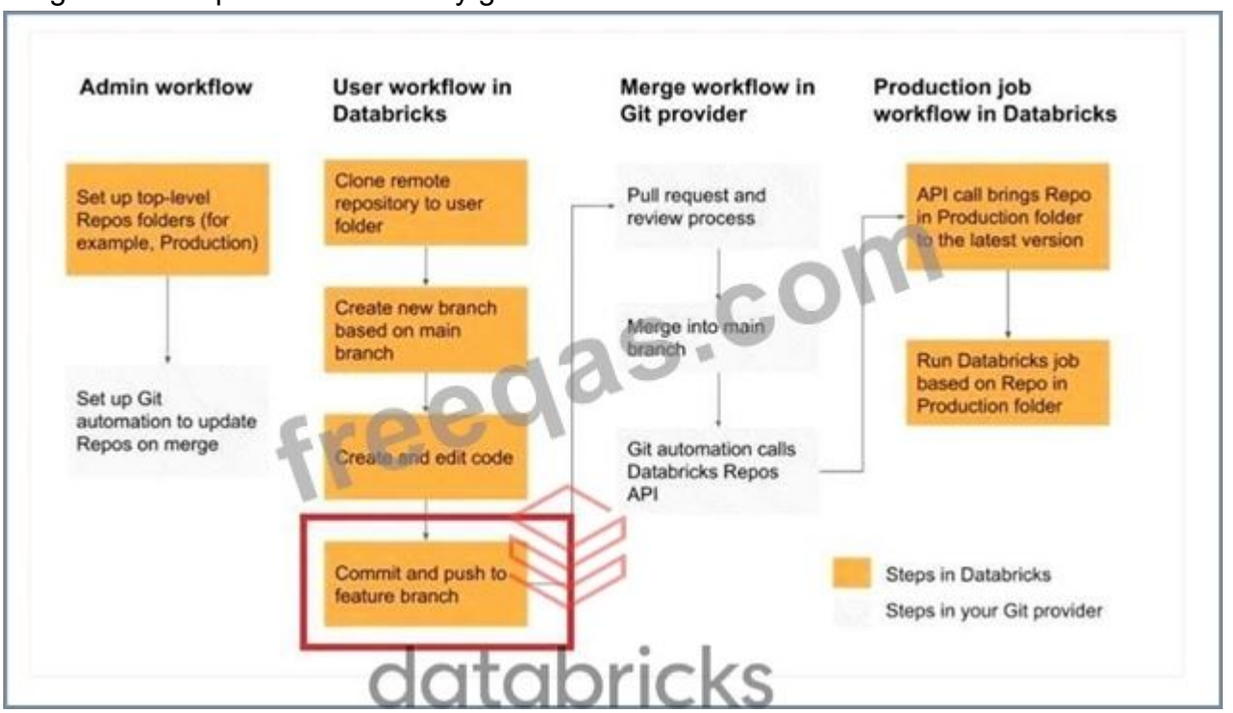
USAGE: does not give any abilities, but is an additional requirement to perform any action on a schema object.
READ_METADATA: gives the ability to view an object and its metadata.
CREATE_NAMED_FUNCTION: gives the ability to create a named UDF in an existing catalog or schema.
MODIFY_CLASSPATH: gives the ability to add files to the Spark classpath.
ALL PRIVILEGES: gives all privileges (is translated into all the above privileges

**NEW QUESTION: 28**
Which of the following commands can be used to query a delta table?
**A.** 1.%python
2.spark.sql("select * from table_name")
**B.** 1.%sql
2.Select * from table_name
**C.** Both A & B
(Correct)
**D.** 1.%python
2.execute.sql("select * from table")
**E.** 1.%python
2.delta.sql("select * from table")
**Answer: (SHOW ANSWER)**
Explanation
The answer is both options A and B
Options C and D are incorrect because there is no command in Spark called execute.sql or delta.sql

**NEW QUESTION: 29**
Which of the following developer operations in CI/CD flow can be implemented in Databricks Re-pos?
**A.** Delete branch
**B.** Trigger Databricks CICD pipeline
**C.** Commit and push code
**D.** Create a pull request
**E.** Approve the pull request
**Answer: C (LEAVE A REPLY)**
Explanation
The answer is Commit and push code.
See the 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.

Exam focus: Please study the below image carefully to understand all of the steps in the CI/CD flow to understand the tasks that are implemented in Databricks Repo vs Git Provider, exam may ask a different type of questions based on this flow.

Diagram Description automatically generated




**NEW QUESTION: 30**

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

**A.** COLLECT_LIST

**B.** COLLECT_SET

**C.** COLLECT_UNION

**D.** ARRAY_INTERSECT

**E.** ARRAY_UNION

**Answer: B (LEAVE A REPLY)**

Explanation

Table Description automatically generated

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']] |



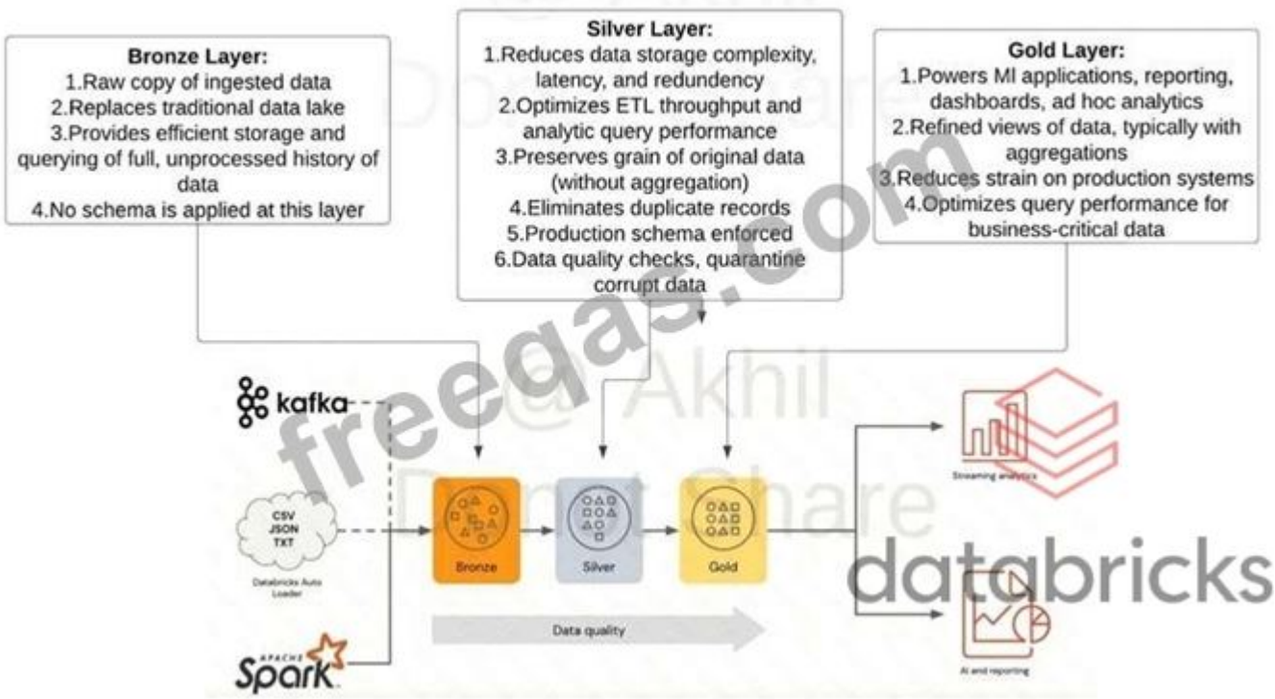
**NEW QUESTION: 31**
Create a sales database using the DBFS location 'dbfs:/mnt/delta/databases/sales.db/'
**A.** CREATE DATABASE sales FORMAT DELTA LOCATION 'dbfs:/mnt/delta/databases/sales.db/"
**B.** CREATE DATABASE sales USING LOCATION 'dbfs:/mnt/delta/databases/sales.db/'
**C.** CREATE DATABASE sales LOCATION 'dbfs:/mnt/delta/databases/sales.db/'
**D.** The sales database can only be created in Delta lake
**E.** CREATE DELTA DATABASE sales LOCATION 'dbfs:/mnt/delta/databases/sales.db/'
**Answer: (SHOW ANSWER)**
Explanation
The answer is
CREATE DATABASE sales LOCATION 'dbfs:/mnt/delta/databases/sales.db/'
Note: with the introduction of the Unity catalog and three-layer namespace usage of SCHEMA and DATABASE is interchangeable

**NEW QUESTION: 32**
You have configured AUTO LOADER to process incoming IOT data from cloud object storage every 15 mins, recently a change was made to the notebook code to update the processing logic but the team later realized that the notebook was failing for the last 24 hours, what steps team needs to take to reprocess the data that was not loaded after the notebook was corrected?
**A.** Move the files that were not processed to another location and manually copy the files into the ingestion path to reprocess them
**B.** Enable back_fill = TRUE to reprocess the data
**C.** Delete the checkpoint folder and run the autoloader again
**D.** Autoloader automatically re-processes data that was not loaded
**E.** Manually re-load the data
**Answer: D (LEAVE A REPLY)**
Explanation
The answer is,
Autoloader automatically re-processes data that was not loaded using the checkpoint.

**NEW QUESTION: 33**
When investigating a data issue you realized that a process accidentally updated the table, you want to query the same table with yesterday's version of the data so you can review what the prior version looks like, what is the best way to query historical data so you can do your analysis?
**A.** SELECT * FROM TIME_TRAVEL(table_name) WHERE time_stamp = 'timestamp'
**B.** TIME_TRAVEL FROM table_name WHERE time_stamp = date_sub(current_date(), 1)
**C.** SELECT * FROM table_name TIMESTAMP AS OF date_sub(current_date(), 1)
**D.** DISCRIBE HISTORY table_name AS OF date_sub(current_date(), 1)
**E.** SHOW HISTORY table_name AS OF date_sub(current_date(), 1)
**Answer: (SHOW ANSWER)**
Explanation
The answer is SELECT * FROM table_name TIMESTAMP as of date_sub(current_date(), 1) FYI, Time travel supports two ways one is using timestamp and the second way is using version number,
Timestamp:
1.SELECT count(*) FROM my_table TIMESTAMP AS OF "2019-01-01"
2.SELECT count(*) FROM my_table TIMESTAMP AS OF date_sub(current_date(), 1)
3.SELECT count(*) FROM my_table TIMESTAMP AS OF "2019-01-01 01:30:00.000" Version Number:
1.SELECT count(*) FROM my_table VERSION AS OF 5238

2.SELECT count(*) FROM my_table@v5238
3.SELECT count(*) FROM delta.`/path/to/my/table@v5238`
https://databricks.com/blog/2019/02/04/introducing-delta-time-travel-for-large-scale-data-lakes.html

**NEW QUESTION: 34**
Which of the following data workloads will utilize a Bronze table as its destination?
**A.** A job that aggregates cleaned data to create standard summary statistics
**B.** A job that queries aggregated data to publish key insights into a dashboard
**C.** A job that ingests raw data from a streaming source into the Lakehouse
**D.** A job that develops a feature set for a machine learning application
**E.** A job that enriches data by parsing its timestamps into a human-readable format
**Answer: C (LEAVE A REPLY)**
Explanation
The answer is A job that ingests raw data from a streaming source into the Lakehouse.
The ingested data from the raw streaming data source like Kafka is first stored in the Bronze layer as first destination before it is further optimized and stored in Silver.
Medallion Architecture - Databricks
Bronze Layer:
1. Raw copy of ingested data
2. Replaces traditional data lake
3. Provides efficient storage and querying of full, unprocessed history of data
4. No schema is applied at this layer
Exam focus: Please review the below image and understand the role of each layer(bronze, silver, gold) in medallion architecture, you will see varying questions targeting each layer and its purpose.
Sorry I had to add the watermark some people in Udemy are copying my content.
Purpose of each layer in medallion architecture




**NEW QUESTION: 35**
You are working on a email spam filtering assignment, while working on this you find there is new word e.g.
HadoopExam comes in email, and in your solutions you never come across this word before, hence probability
of this words is coming in either email could be zero. So which of the following algorithm can help you to avoid zero probability?
**A.** Naive Bayes
**B.** Laplace Smoothing
**C.** Logistic Regression
**D.** All of the above
**Answer: B (LEAVE A REPLY)**
Explanation
Laplace smoothing is a technique for parameter estimation which accounts for unobserved events. It is more
robust and will not fail completely when data that has never been observed in training shows up.

**NEW QUESTION: 36**
Your team member is trying to set up a delta pipeline and build a second gold table to the same pipeline with aggregated metrics based on an existing Delta Live table called sales_orders_cleaned but he is facing a problem in starting the pipeline, the pipeline is failing to state it cannot find the table sales_orders_cleaned, you are asked to identify and fix the problem.

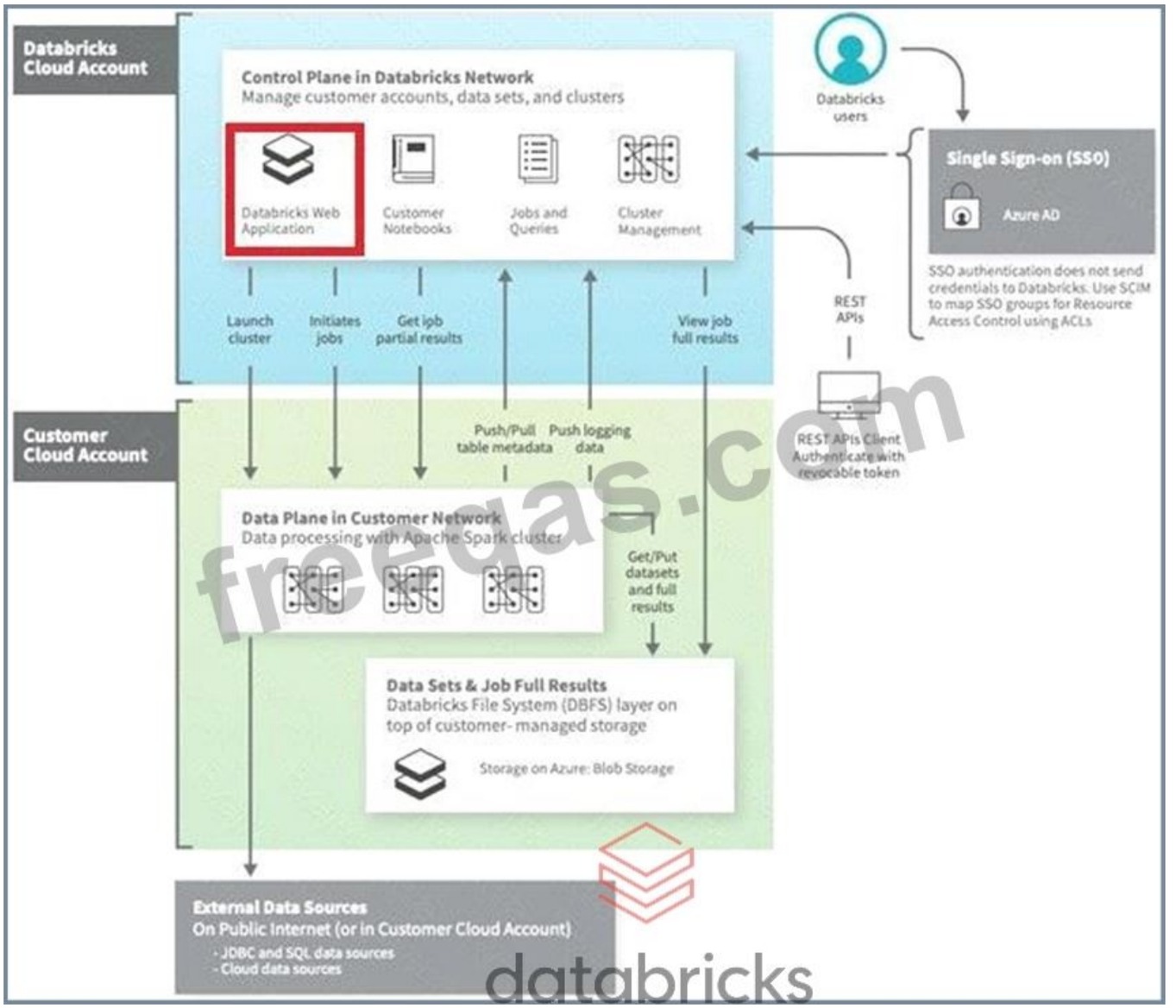
1.CREATE LIVE TABLE sales_order_in_chicago
2.AS
3.SELECT order_date, city, sum(price) as sales,
4.FROM sales_orders_cleaned
5.WHERE city = 'Chicago')
6.GROUP BY order_date, city
**A.** Use STREAMING LIVE instead of LIVE table
**B.** Delta live table can be used in a group by clause
**C.** Delta live tables pipeline can only have one table
**D.** Sales_orders_cleaned table is missing schema name LIVE
**E.** The pipeline needs to be deployed so the first table is created before we add a second table
**Answer: (SHOW ANSWER)**
Explanation
The answer is, Sales_orders_cleaned table is missing schema name LIVE
Every Delta live table should have schema LIVE
Here is the correct syntax,
1.CREATE LIVE TABLE sales_order_in_chicago
2.AS
3.SELECT order_date, city, sum(price) as sales,
4.FROM LIVE.sales_orders_cleaned
5.WHERE city = 'Chicago')
6.GROUP BY order_date, city

**NEW QUESTION: 37**
Drop the customers database and associated tables and data, all of the tables inside the database are managed tables. Which of the following SQL commands will help you accomplish this?
**A.** DROP DATABASE customers FORCE
**B.** DROP DATABASE customers CASCADE
**C.** DROP DATABASE customers INCLUDE
**D.** All the tables must be dropped first before dropping database
**E.** DROP DELTA DATABSE customers
**Answer: C (LEAVE A REPLY)**
Explanation
The answer is DROP DATABASE customers CASCADE
Drop database with cascade option drops all the tables, since all of the tables inside the database are managed tables we do not need to perform any additional steps to clean the data in the storage.

**NEW QUESTION: 38**
Which of the following describes a scenario in which a data engineer will want to use a Job cluster instead of
an all-purpose cluster?

**A.** An ad-hoc analytics report needs to be developed while minimizing compute costs
**B.** A Databricks SQL query needs to be scheduled for upward reporting
**C.** An automated workflow needs to be run every 30 minutes
**D.** A data engineer needs to manually investigate a production error
**E.** A data team needs to collaborate on the development of a machine learning model
**Answer: C (LEAVE A REPLY)**

**NEW QUESTION: 39**
Which of the following are stored in the control pane of Databricks Architecture?
**A.** Job Clusters
**B.** All Purpose Clusters
**C.** Databricks Filesystem
**D.** Databricks Web Application
**E.** Delta tables
**Answer: D (LEAVE A REPLY)**
Explanation
The answer is Databricks Web Application
Azure Databricks architecture overview - Azure Databricks | Microsoft Docs 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 pane.
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.
Timeline Description automatically generated





Bottom of Form

Top of Form

**NEW QUESTION: 40**

The team has decided to take advantage of table properties to identify a business owner for each table, which of the following table DDL syntax allows you to populate a table property identifying the business owner of a table CREATE TABLE inventory (id INT, units FLOAT)

**A.** SET TBLPROPERTIES business_owner = 'supply chain'
CREATE TABLE inventory (id INT, units FLOAT)

**B.** TBLPROPERTIES (business_owner = 'supply chain')

**C.** CREATE TABLE inventory (id INT, units FLOAT)
SET (business_owner = 'supply chain')

**D.** CREATE TABLE inventory (id INT, units FLOAT)
SET PROPERTY (business_owner = 'supply chain')

**E.** CREATE TABLE inventory (id INT, units FLOAT)
SET TAG (business_owner = 'supply chain')

**Answer: B (LEAVE A REPLY)**

Explanation

CREATE TABLE inventory (id INT, units FLOAT) TBLPROPERTIES (business_owner = 'supply chain') Table properties and table options (Databricks SQL) | Databricks on AWS Alter table command can used to update the TBLPROPERTIES ALTER TABLE inventory SET TBLPROPERTIES(business_owner , 'operations')

**NEW QUESTION: 41**

Which of the following SQL statement can be used to query a table by eliminating duplicate rows from the query results?

**A.** SELECT DISTINCT * FROM table_name

**B.** SELECT DISTINCT * FROM table_name HAVING COUNT(*) > 1

**C.** SELECT DISTINCT_ROWS (*) FROM table_name

**D.** SELECT * FROM table_name GROUP BY * HAVING COUNT(*) < 1

**E.** SELECT * FROM table_name GROUP BY * HAVING COUNT(*) > 1

**Answer: A (LEAVE A REPLY)**

Explanation

The answer is SELECT DISTINCT * FROM table_name

**NEW QUESTION: 42**

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 mes-sage?

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

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

**C.** All-purpose clusters are less expensive than the job clusters

**D.** All-purpose clusters are more expensive than the job clusters

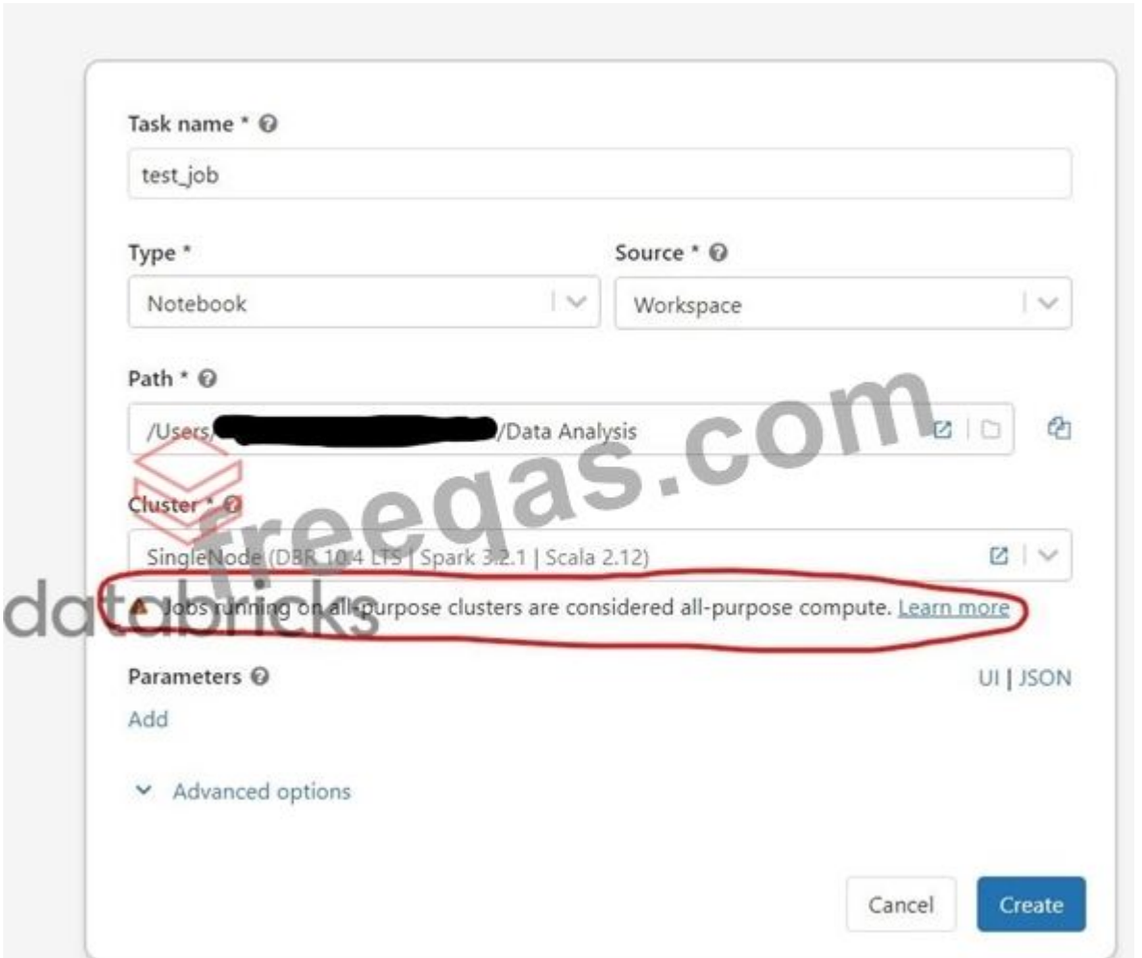
**E.** All-purpose cluster provide interactive messages that can not be viewed in a job

**Answer: D (LEAVE A REPLY)**

Explanation

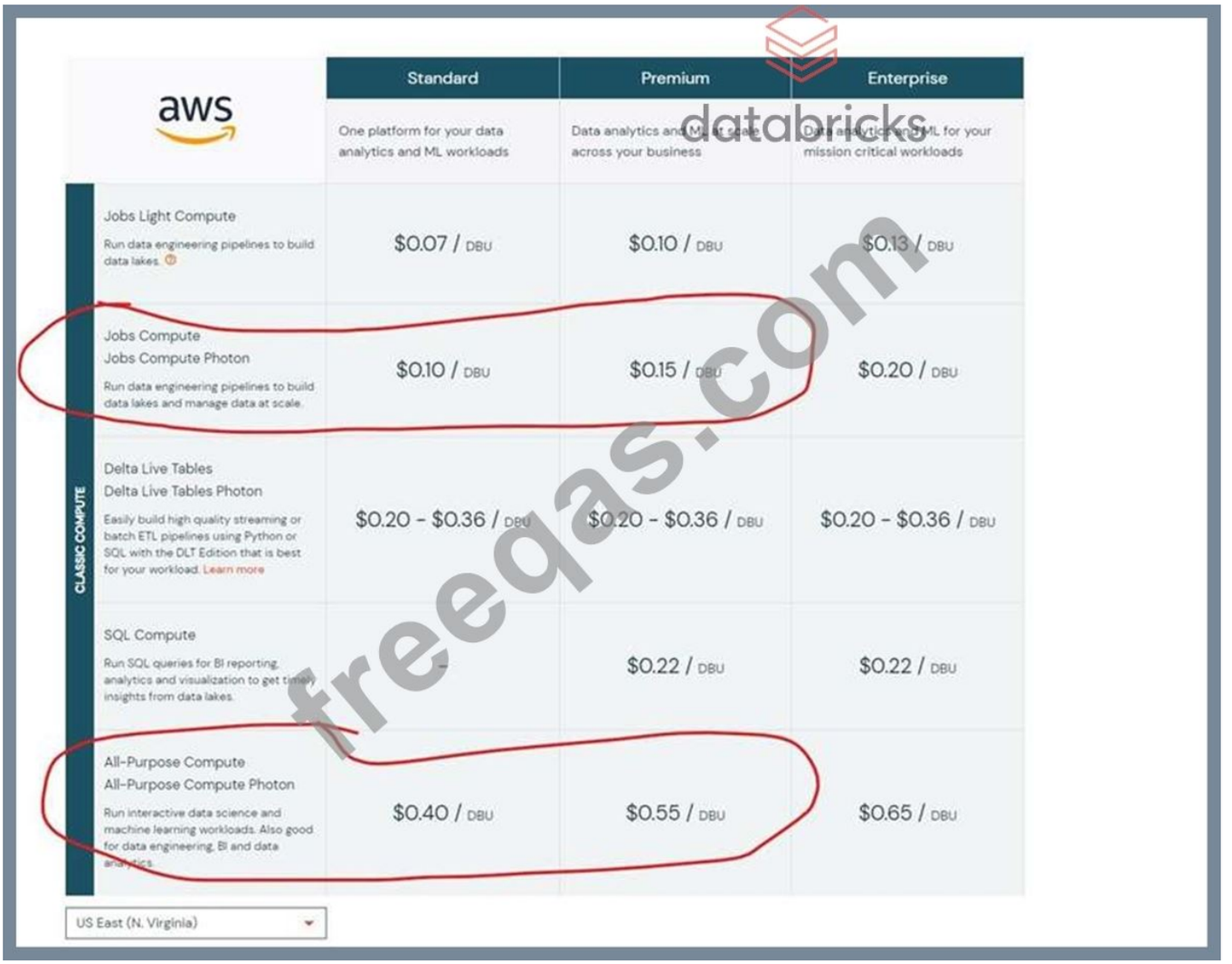
Warning message:

Graphical user interface, text, application, email Description automatically generated





Pricing for All-purpose clusters are more expensive than the job clusters AWS pricing(Aug 15th 2022)Graphical user interface Description automatically generated





Bottom of Form

Top of Form

**NEW QUESTION: 43**

How VACCUM and OPTIMIZE commands can be used to manage the DELTA lake?

**A.** VACCUM command can be used to compact small parquet files, and the OP-TIMZE command can be used to delete parquet files that are marked for dele-tion/unused.

**B.** VACCUM command can be used to delete empty/blank parquet files in a delta table. OPTIMIZE command can be used to update stale statistics on a delta table.

**C.** VACCUM command can be used to compress the parquet files to reduce the size of the table, OPTIMIZE command can be used to cache frequently delta tables for better performance.

**D.** VACCUM command can be used to delete empty/blank parquet files in a delta table, OPTIMIZE command can be used to cache frequently delta tables for better perfor-mance.

**E.** OPTIMIZE command can be used to compact small parquet files, and the VAC-CUM command can be used to delete parquet files that are marked for deletion/unused.

(Correct)

**Answer: E (LEAVE A REPLY)**

Explanation

VACCUM:
You can remove files no longer referenced by a Delta table and are older than the retention thresh-old by running the vacuum command on the table. vacuum is not triggered automatically. The de-fault retention threshold for the files is 7 days. To change this behavior, see Configure data reten-tion for time travel.
OPTIMIZE:
Using OPTIMIZE you can compact data files on Delta Lake, this can improve the speed of read queries on the table. Too many small files can significantly degrade the performance of the query.

**NEW QUESTION: 44**
You are currently asked to work on building a data pipeline, you have noticed that you are currently working with a data source that has a lot of data quality issues and you need to monitor data quality and enforce it as part of the data ingestion process, which of the following tools can be used to address this problem?
**A.** AUTO LOADER
**B.** DELTA LIVE TABLES
**C.** JOBS and TASKS
**D.** UNITY Catalog and Data Governance
**E.** STRUCTURED STREAMING with MULTI HOP
**Answer: (SHOW ANSWER)**
Explanation
The answer is, DELTA LIVE TABLES
Delta live tables expectations can be used to identify and quarantine bad data, all of the data quality metrics are stored in the event logs which can be used to later analyze and monitor.
DELTA LIVE Tables expectations
Below are three types of expectations, make sure to pay attention differences between these three.
Retain 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:
Python
1.@dlt.expect("valid timestamp", "col("timestamp") > '2012-01-01'")
SQL
1.CONSTRAINT valid_timestamp EXPECT (timestamp > '2012-01-01')
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:
Python
1.@dlt.expect_or_drop("valid_current_page", "current_page_id IS NOT NULL AND cur-rent_page_title IS NOT NULL") SQL
1.CONSTRAINT valid_current_page EXPECT (current_page_id IS NOT NULL and cur-rent_page_title IS NOT NULL) ON VIOLATION DROP ROW Fail on invalid records:

When invalid records are unacceptable, use the expect or fail operator to halt execution imme-diately when a record fails validation. If the operation is a table update, the system atomically rolls back the transaction:
Python
1.@dlt.expect_or_fail("valid_count", "count > 0")
SQL
1.CONSTRAINT valid_count EXPECT (count > 0) ON VIOLATION FAIL UPDATE

**NEW QUESTION: 45**
What is the main difference between the below two commands?
1.INSERT OVERWRITE table_name
2.SELECT * FROM table
1.CREATE OR REPLACE TABLE table_name
2.AS SELECT * FROM table
**A.** INSERT OVERWRITE replaces data by default, CREATE OR REPLACE replaces data and Schema by default
**B.** INSERT OVERWRITE replaces data and schema by default, CREATE OR REPLACEreplaces data by default
**C.** INSERT OVERWRITE maintains historical data versions by de-fault, CREATE OR REPLACEclears the historical data versions by default
**D.** INSERT OVERWRITE clears historical data versions by de-fault, CREATE OR REPLACE maintains the historical data versions by default
**E.** Both are same and results in identical outcomes
**Answer: A (LEAVE A REPLY)**
Explanation
The main difference between INSERT OVERWRITE and CREATE OR REPLACE TABLE(CRAS) is that CRAS can modify the schema of the table, i.e it can add new columns or change data types of existing columns. By default INSERT OVERWRITE only overwrites the data.
INSERT OVERWRITE can also be used to overwrite schema, only when spark.databricks.delta.schema.autoMerge.enabled is set true if this option is not enabled and if there is a schema mismatch command will fail.

**NEW QUESTION: 46**
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?
**A.** Move the workspace from Central US zone to East US Zone
**B.** Change the timestamp on the delta tables to America/New_York format
**C.** Change the spark configuration of SQL endpoint to format the timestamp to Ameri-ca/New_York
**D.** Under SQL Admin Console, set the SQL configuration parameter time zone to Ameri-ca/New_York
**E.** Add SET Timezone = America/New_York on every of the SQL queries in the dashboard.

**Answer: (SHOW ANSWER)**

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:

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

2.Click the SQL Warehouse Settings tab.

3.In the SQL Configuration Parameters textbox, specify one key-value pair per line. Sepa-rate the name of the parameter from its value using a space. For example, to ena-ble ANSI_MODE:

Graphical user interface, text, application Description automatically generated


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

timezone America/New_York

SQL configuration parameters | Databricks on AWS

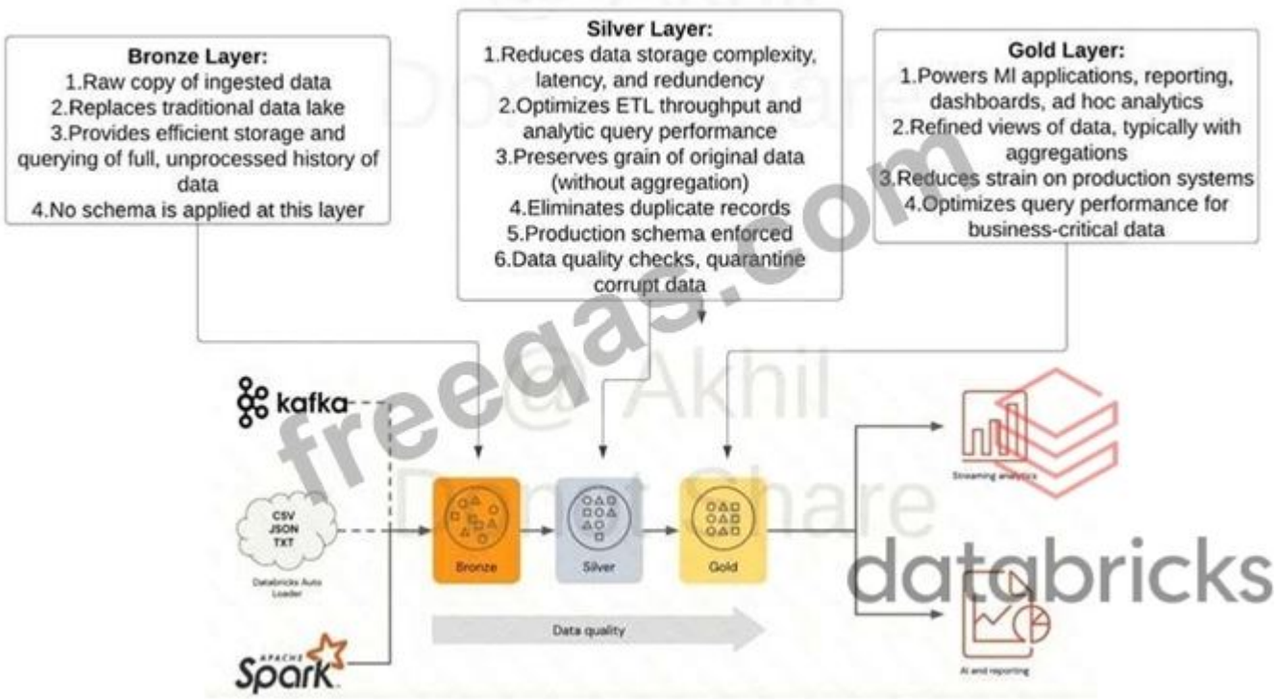


**NEW QUESTION: 47**

A new data engineer has started at a company. The data engineer has recently been added to the company's

Databricks workspace as new.engineer@company.com. The data engineer needs to be able to query the table
sales in the database retail. The new data engineer already has been granted USAGE on the database retail.
Which of the following commands can be used to grant the appropriate permissions to the new data engineer?
**A.** GRANT USAGE ON TABLE new.engineer@company.com TO sales;
**B.** GRANT SELECT ON TABLE sales TO new.engineer@company.com;
**C.** GRANT CREATE ON TABLE sales TO new.engineer@company.com;
**D.** GRANT USAGE ON TABLE sales TO new.engineer@company.com;
**E.** GRANT SELECT ON TABLE new.engineer@company.com TO sales;
**Answer: B (LEAVE A REPLY)**

**NEW QUESTION: 48**
What is the purpose of the bronze layer in a Multi-hop Medallion architecture?
**A.** Copy of raw data, easy to query and ingest data for downstream processes.
**B.** Powers ML applications
**C.** Data quality checks, corrupt data quarantined
**D.** Contain aggregated data that is to be consumed into Silver
**E.** Reduces data storage by compressing the data
**Answer: A (LEAVE A REPLY)**
Explanation
The answer is, copy of raw data, easy to query and ingest data for downstream processes, Medallion Architecture - Databricks Here are the typical role of Bronze Layer in a medallion architecture.
Bronze Layer:
1. Raw copy of ingested data
2. Replaces traditional data lake
3. Provides efficient storage and querying of full, unprocessed history of data
4. No schema is applied at this layer
Exam focus: Please review the below image and understand the role of each layer(bronze, silver, gold) in medallion architecture, you will see varying questions targeting each layer and its purpose.
Sorry I had to add the watermark some people in Udemy are copying my content.




**NEW QUESTION: 49**

Consider flipping a coin for which the probability of heads is p, where p is unknown, and our goa is to estimate p. The obvious approach is to count how many times the coin came up heads and divide by the total

number of coin flips. If we flip the coin 1000 times and it comes up heads 367 times, it is very reasonable to

estimate p as approximately 0.367. However, suppose we flip the coin only twice and we get heads both times.

Is it reasonable to estimate p as 1.0? Intuitively, given that we only flipped the coin twice, it seems a bit rash to conclude that the coin will always come up heads, and_____is a way of avoiding such rash

conclusions.

**A.** Naive Bayes

**B.** Laplace Smoothing

**C.** Logistic Regression

**D.** Linear Regression

**Answer: B (LEAVE A REPLY)**

Explanation

Smooth the estimates:consider flipping a coin for which the probability of heads is p, where p is unknown, and

our goal is to estimate p. The obvious approach is to count how many times the coin came up heads and divide

by the total number of coin flips. If we flip the coin 1000 times and it comes up heads 367 times, it is very
reasonable to estimate p as approximately 0.367. However, suppose we flip the coin only twice and we get
heads both times. Is it reasonable to estimate p as 1.0? Intuitively, given that we only flipped the coin twice, it
seems a bit rash to conclude that the coin will always come up heads, and smoothing is a way of avoiding such
rash conclusions. A simple smoothing method, called Laplace smoothing (or Laplace's law of succession or
add-one smoothing in R&N), is to estimate p by (one plus the number of heads) / (two plus the total number of
flips). Said differently, if we are keeping count of the number of heads and the number of tails, this rule is
equivalent to starting each of our counts at one, rather than zero. Another advantage of Laplace smoothing is
that it avoids estimating any probabilities to be zero, even for events never observed in the data. Laplace
add-one smoothing now assigns too much probability to unseen words

**NEW QUESTION: 50**
Which of the following two options are supported in identifying the arrival of new files, and incre-mental data from Cloud object storage using Auto Loader?
**A.** Directory listing, File notification
**B.** Checking pointing, watermarking
**C.** Writing ahead logging, read head logging
**D.** File hashing, Dynamic file lookup
**E.** Checkpointing and Write ahead logging
**Answer: A (LEAVE A REPLY)**
Explanation
The answer is A, Directory listing, File notifications
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

**NEW QUESTION: 51**
How do you check the location of an existing schema in Delta Lake?
**A.** Run SQL command SHOW LOCATION schema_name
**B.** Check unity catalog UI
**C.** Use Data explorer

**D.** Run SQL command DESCRIBE SCHEMA EXTENDED schema_name

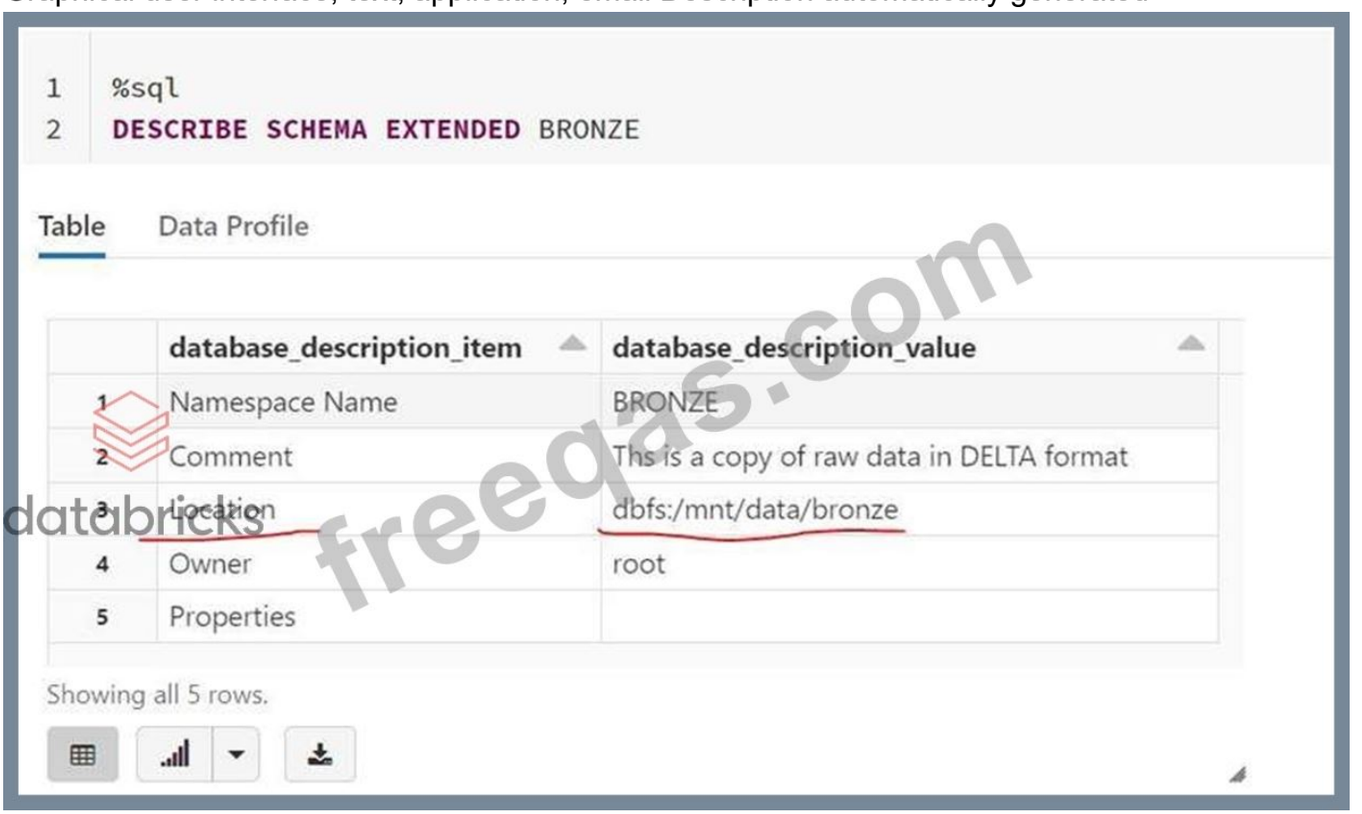
E Schemas are internally in-store external hive meta stores like MySQL or SQL Server

**Answer: D (LEAVE A REPLY)**

Explanation

Here is an example of how it looks

Graphical user interface, text, application, email Description automatically generated






**NEW QUESTION: 52**

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

**A.** Optimizes ETL throughput and analytic query performance

**B.** Eliminate duplicate records

**C.** Preserves grain of original data, without any aggregations

**D.** Data quality checks and schema enforcement

**E.** Powers ML applications, reporting, dashboards and adhoc reports.

**Answer: E (LEAVE A REPLY)**

Explanation

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

Review the below link for more info,

Medallion Architecture - Databricks

Gold Layer:

1.Powers Ml applications, reporting, dashboards, ad hoc analytics

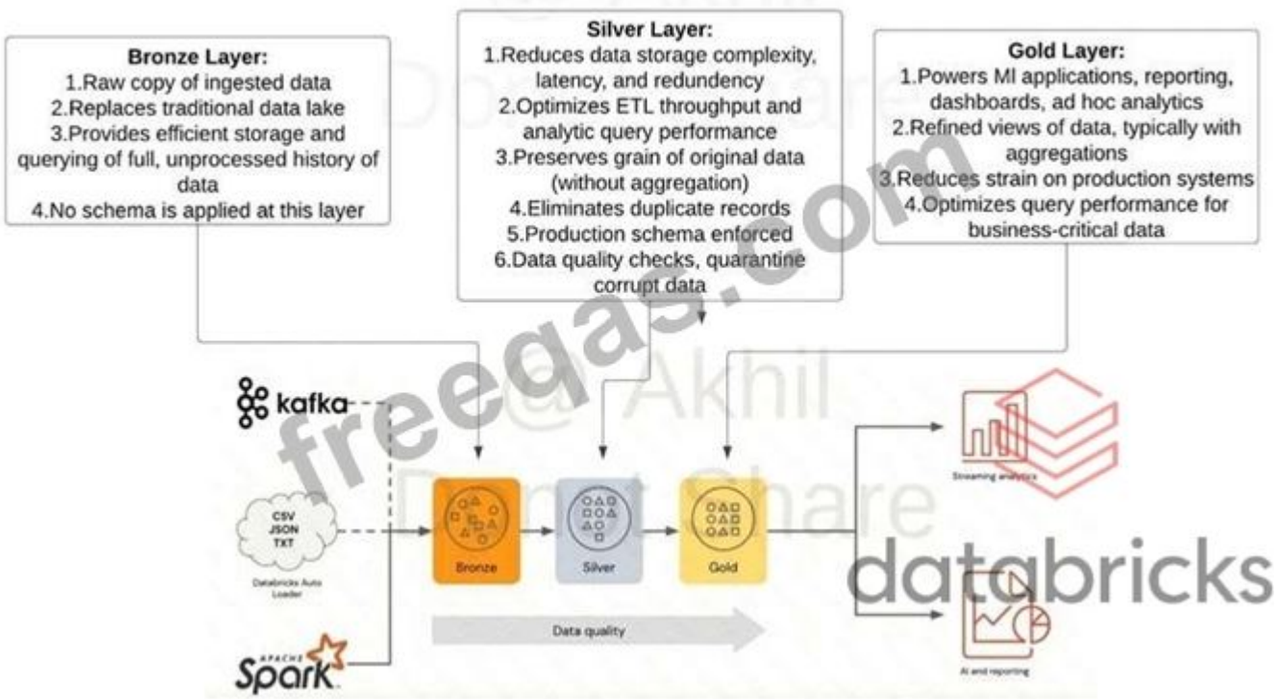
2.Refined views of data, typically with aggregations

3.Reduces strain on production systems

4.Optimizes query performance for business-critical data

Exam focus: Please review the below image and understand the role of each layer(bronze, silver, gold) in medallion architecture, you will see varying questions targeting each layer and its purpose.

Sorry I had to add the watermark some people in Udemy are copying my content.




**NEW QUESTION: 53**

You are currently working on a project that requires the use of SQL and Python in a given note-book, what would be your approach

**A.** Create two separate notebooks, one for SQL and the second for Python

**B.** A single notebook can support multiple languages, use the magic command to switch between the two.

**C.** Use an All-purpose cluster for python, SQL endpoint for SQL

**D.** Use job cluster to run python and SQL Endpoint for SQL

**Answer: (SHOW ANSWER)**

Explanation

The answer is, A single notebook can support multiple languages, use the magic command to switch between the two.

Use %sql and %python magic commands within the same notebook

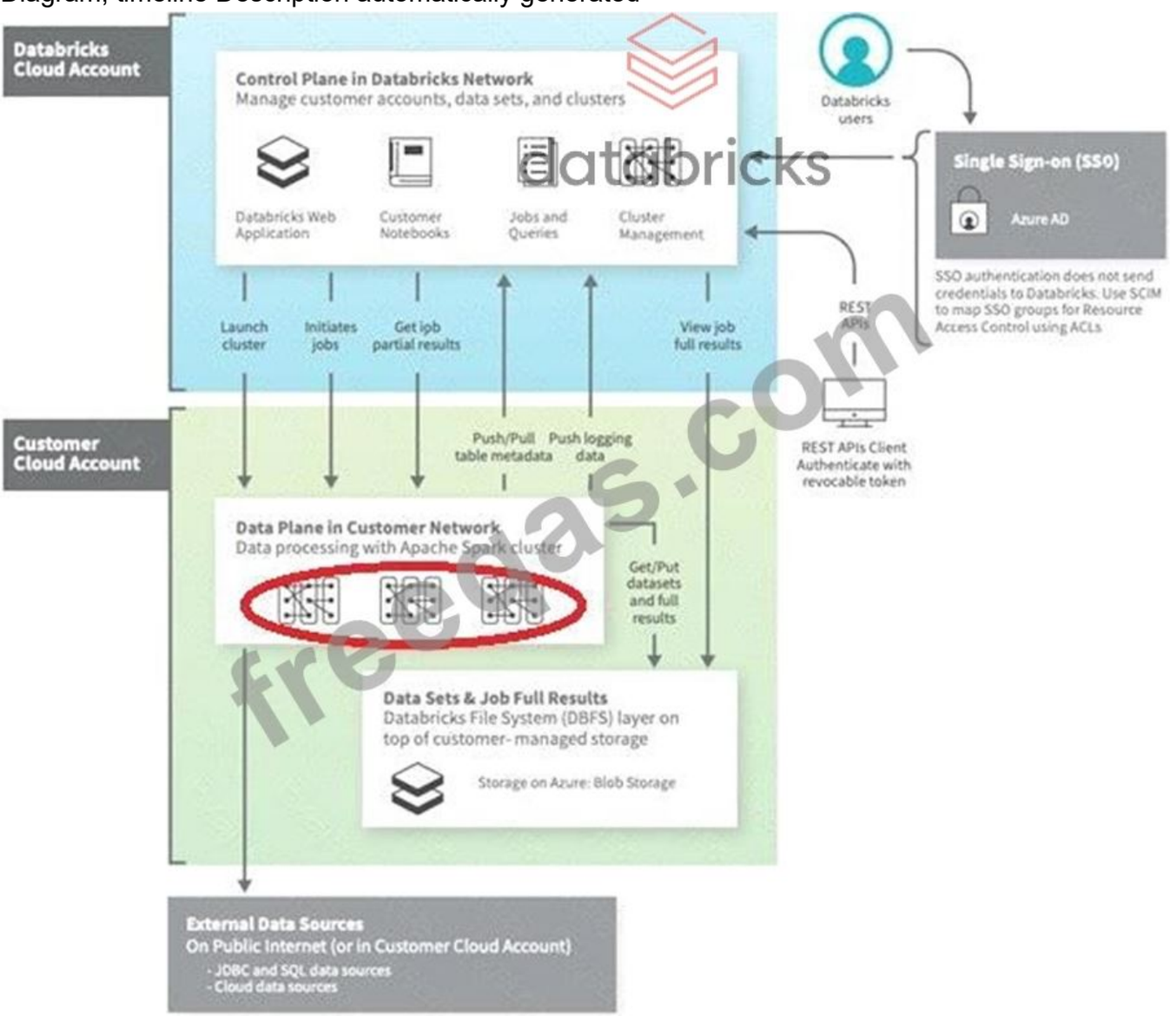
**NEW QUESTION: 54**

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

**A.** Data plane

**B.** Control plane

**C.** Databricks Filesystem
**D.** JDBC data source
**E.** Databricks web application
**Answer: A (LEAVE A REPLY)**
Explanation
The answer is Data Plane, which is where compute(all-purpose, Job Cluster, DLT) are stored this is generally a customer cloud account, there is one exception SQL Warehouses, currently there are 3 types of SQL Warehouse compute available(classic, pro, serverless), in classic and pro compute is located in customer cloud account but serverless computed is located in Databricks cloud account.
Diagram, timeline Description automatically generated




**NEW QUESTION: 55**
While investigating a data issue, you wanted to review yesterday's version of the table using below command, while querying the previous version of the table using time travel you realized that you are no longer able to view the historical data in the table and you could see it the table was updated yesterday

based on the table history(DESCRIBE HISTORY table_name) command what could be the reason why you can not access this data?
SELECT * FROM table_name TIMESTAMP AS OF date_sub(current_date(), 1)
**A.** You currently do not have access to view historical data
**B.** By default, historical data is cleaned every 180 days in DELTA
**C.** A command VACUUM table_name RETAIN 0 was ran on the table
**D.** Time travel is disabled
**E.** Time travel must be enabled before you query previous data
**Answer: C (LEAVE A REPLY)**
Explanation
The answer is, VACUUM table_name RETAIN 0 was ran
The VACUUM command recursively vacuums directories associated with the Delta table and re-moves data files that are no longer in the latest state of the transaction log for the table and are older than a retention threshold. The default is 7 Days.
When VACUUM table_name RETAIN 0 is ran all of the historical versions of data are lost time travel can only provide the current state.

**NEW QUESTION: 56**
Question-26. There are 5000 different color balls, out of which 1200 are pink color. What is the maximum
likelihood estimate for the proportion of "pink" items in the test set of color balls?
**A.** 2.4
**B.** 24 0
**C.** .24
**D.** .48
**E.** 4.8
**Answer: (SHOW ANSWER)**
Explanation
Given no additional information, the MLE for the probability of an item in the test set is exactly its frequency
in the training set. The method of maximum likelihood corresponds to many well-known estimation methods
in statistics. For example, one may be interested in the heights of adult female penguins, but be unable to
measure the height of every single penguin in a population due to cost or time constraints. Assuming that the
heights are normally (Gaussian) distributed with some unknown mean and variance, the mean and variance
can be estimated with MLE while only knowing the heights of some sample of the overall population.
MLE

would accomplish this by taking the mean and variance as parameters and finding particular parametric values
that make the observed results the most probable (given the model).
In general, for a fixed set of data and underlying statistical model the method of maximum likelihood selects
the set of values of the model parameters that maximizes the likelihood function. Intuitively, this maximizes
the "agreement" of the selected model with the observed data, and for discrete random variables it indeed
maximizes the probability of the observed data under the resulting distribution. Maximum-likelihood estimation gives a unified approach to estimation, which is well-defined in the case of the normal distribution
and many other problems. However in some complicated problems, difficulties do occur: in such problems,
maximum-likelihood estimators are unsuitable or do not exist.

**NEW QUESTION: 57**
What is the top-level object in unity catalog?
**A.** Catalog
**B.** Table
**C.** Workspace
**D.** Database
**E.** Metastore
**Answer: E (LEAVE A REPLY)**
Explanation
Key concepts - Azure Databricks | Microsoft Docs

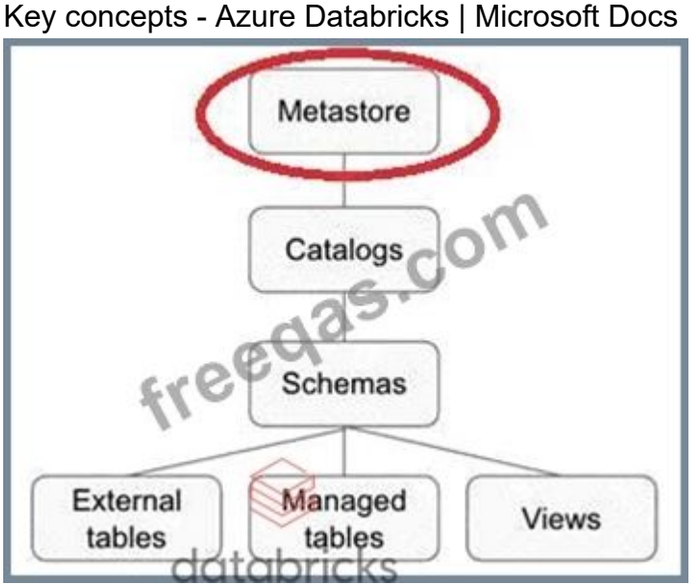



Diagram Description automatically generated

**NEW QUESTION: 58**

How do you upgrade an existing workspace managed table to a unity catalog table?
**A.** ALTER TABLE table_name SET UNITY_CATALOG = TRUE
**B.** Create table catalog_name.schema_name.table_name
as select * from hive_metastore.old_schema.old_table
**C.** Create table table_name as select * from hive_metastore.old_schema.old_table
**D.** Create table table_name format = UNITY as select * from old_table_name
**E.** Create or replace table_name format = UNITY using deep clone old_table_name
**Answer: B (LEAVE A REPLY)**
Explanation
The answer is Create table catalog_name.schema_name.table_name as select * from hive_metastore.old_schema.old_table Basically, we are moving the data from an internal hive metastore to a metastore and catalog that is registered in the Unity catalog.
note: if it is a managed table the data is copied to a different storage account, for a large tables this can take a lot of time. For an external table the process is different.
Managed table: Upgrade a managed to Unity Catalog
External table: Upgrade an external table to Unity Catalog

**NEW QUESTION: 59**
Which of the statements is correct when choosing between lakehouse and Datawarehouse?
**A.** Traditional Data warehouses have special indexes which are optimized for Machine learning
**B.** Traditional Data warehouses can serve low query latency with high reliability for BI workloads
**C.** SQL support is only available for Traditional Datawarehouse's, Lakehouses support Python and Scala
**D.** Traditional Data warehouses are the preferred choice if we need to support ACID, Lakehouse does not support ACID.
**E.** Lakehouse replaces the current dependency on data lakes and data warehouses uses an open standard storage format and supports low latency BI workloads.
**Answer: E (LEAVE A REPLY)**
Explanation
The lakehouse replaces the current dependency on data lakes and data warehouses for modern data companies that desire:
* Open, direct access to data stored in standard data formats.
* Indexing protocols optimized for machine learning and data science.
* Low query latency and high reliability for BI and advanced analytics.

**NEW QUESTION: 60**
Which of the following is correct for the global temporary view?
**A.** global temporary views cannot be accessed once the notebook is detached and attached
**B.** global temporary views can be accessed across many clusters
**C.** global temporary views can be still accessed even if the notebook is detached and at-tached
**D.** global temporary views can be still accessed even if the cluster is restarted

**E.** global temporary views are created in a database called temp database

**Answer: C (LEAVE A REPLY)**

Explanation

The answer is global temporary views can be still accessed even if the notebook is detached and attached There are two types of temporary views that 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 reattached local temporary view is lost.

* A global temporary view is available to all the notebooks in the cluster, even if the notebook is detached and reattached it can still be accessible but if a cluster is restarted the global temporary view is lost.

**NEW QUESTION: 61**

Which of the following commands results in the successful creation of a view on top of the delta stream(stream on delta table)?

**A.** Spark.read.format("delta").table("sales").createOrReplaceTempView("streaming_vw")

**B.** Spark.readStream.format("delta").table("sales").createOrReplaceTempView("streaming_vw")

**C.**

Spark.read.format("delta").table("sales").mode("stream").createOrReplaceTempView("streaming_vw")

**D.**

Spark.read.format("delta").table("sales").trigger("stream").createOrReplaceTempView("streaming_vw")

**E.** Spark.read.format("delta").stream("sales").createOrReplaceTempView("streaming_vw")

**F.** You can not create a view on streaming data source.

**Answer: B (LEAVE A REPLY)**

Explanation

The answer is

Spark.readStream.table("sales").createOrReplaceTempView("streaming_vw") When you load a Delta table as a stream source and use it in a streaming query, the query processes all of the data present in the table as well as any new data that arrives after the stream is started.

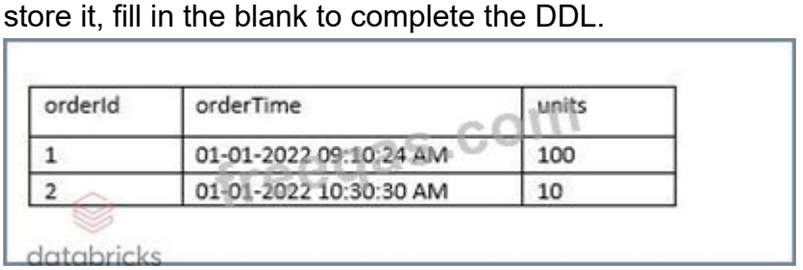
You can load both paths and tables as a stream, you also have the ability to ignore deletes and changes(updates, Merge, overwrites) on the delta table.

Here is more information,

https://docs.databricks.com/delta/delta-streaming.html#delta-table-as-a-source

**NEW QUESTION: 62**
Which of the following tool provides Data Access control, Access Audit, Data Lineage, and Data discovery?
**A.** Lakehouse
**B.** DELTA lake
**C.** Unity Catalog
**D.** Data Governance
**E.** DELTA LIVE Pipelines
**Answer: C (LEAVE A REPLY)**

**NEW QUESTION: 63**
You were asked to create a table that can store the below data, orderTime is a timestamp but the finance team when they query this data normally prefer the orderTime in date format, you would like to create a calculated column that can convert the orderTime column timestamp datatype to date and store it, fill in the blank to complete the DDL.




**A.** AS DEFAULT (CAST(orderTime as DATE))
**B.** GENERATED ALWAYS AS (CAST(orderTime as DATE))
Correct)
**C.** GENERATED DEFAULT AS (CAST(orderTime as DATE))
**D.** AS (CAST(orderTime as DATE))
**E.** Delta lake does not support calculated columns, value should be inserted into the table as part of the ingestion process
**Answer: B (LEAVE A REPLY)**
Explanation
The answer is, GENERATED ALWAYS AS (CAST(orderTime as DATE))
https://docs.microsoft.com/en-us/azure/databricks/delta/delta-batch#--use-generated-columns Delta Lake supports generated columns which are a special type of columns whose values are au-tomatically generated based on a user-specified function over other columns in the Delta table. When you write to a table with generated columns and you do not explicitly provide values for them, Delta Lake automatically computes the values.
Note: Databricks also supports partitioning using generated column

**NEW QUESTION: 64**

The Delta Live Tables Pipeline is configured to run in Development mode using the Triggered Pipeline Mode.
what is the expected outcome after clicking Start to update the pipeline?
**A.** All datasets will be updated once and the pipeline will shut down. The compute resources will be terminated
**B.** All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will be deployed for the update and terminated when the pipeline is stopped
**C.** All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist after the pipeline is stopped to allow for additional development and testing
**D.** All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional development and testing
**E.** All datasets will be updated continuously and the pipeline will not shut down. The compute resources will persist with the pipeline
**Answer: E (LEAVE A REPLY)**
Explanation
The answer is All datasets will be updated once and the pipeline will shut down. The compute re-sources will persist to allow for additional testing.
DLT pipeline supports two modes Development and Production, you can switch between the two based on the stage of your development and deployment lifecycle.
Development and production modes
When you run your pipeline in development mode, the Delta Live Tables system:
*Reuses a cluster to avoid the overhead of restarts.
*Disables pipeline retries so you can immediately detect and fix errors.
In production mode, the Delta Live Tables system:
*Restarts the cluster for specific recoverable errors, including memory leaks and stale credentials.
*Retries execution in the event of specific errors, for example, a failure to start a cluster.
Use the buttons in the Pipelines UI to switch between develop-ment and production modes. By default, pipelines run in development mode.
Switching between development and production modes only controls cluster and pipeline execution behavior.
Storage locations must be configured as part of pipeline settings and are not affected when switching between modes.
Please review additional DLT concepts using below link
https://docs.databricks.com/data-engineering/delta-live-tables/delta-live-tables-concepts.html#delta-live-tables-c

**NEW QUESTION: 65**
You have noticed the Data scientist team is using the notebook versioning feature with git integra-tion, you have recommended them to switch to using Databricks Repos, which of the below reasons could be the reason the why the team needs to switch to Databricks Repos.
**A.** Databricks Repos allows multiple users to make changes

**B.** Databricks Repos allows merge and conflict resolution
**C.** Databricks Repos has a built-in version control system
**D.** Databricks Repos automatically saves changes
**E.** Databricks Repos allow you to add comments and select the changes you want to commit.
**Answer: E (LEAVE A REPLY)**
Explanation
The answer is Databricks Repos allow you to add comments and select the changes you want to commit.

**NEW QUESTION: 66**
A data architect is designing a data model that works for both video-based machine learning work-loads and
highly audited batch ETL/ELT workloads.
Which of the following describes how using a data lakehouse can help the data architect meet the needs of
both workloads?
**A.** A data lakehouse combines compute and storage for simple governance
**B.** A data lakehouse fully exists in the cloud
**C.** A data lakehouse provides autoscaling for compute clusters
**D.** A data lakehouse stores unstructured data and is ACID-compliant
**E.** A data lakehouse requires very little data modeling
**Answer: D (LEAVE A REPLY)**

**NEW QUESTION: 67**
A data engineer has set up two Jobs that each run nightly. The first Job starts at 12:00 AM, and it usually
completes in about 20 minutes. The second Job depends on the first Job, and it starts at 12:30 AM. Sometimes,
the second Job fails when the first Job does not complete by 12:30 AM.
Which of the following approaches can the data engineer use to avoid this problem?
**A.** They can set up the data to stream from the first Job to the second Job
**B.** They can use cluster pools to help the Jobs run more efficiently
**C.** They can limit the size of the output in the second Job so that it will not fail as easily
**D.** They can set up a retry policy on the first Job to help it run more quickly
**E.** They can utilize multiple tasks in a single job with a linear dependency
**Answer: E (LEAVE A REPLY)**

**NEW QUESTION: 68**
Kevin is the owner of the schema sales, Steve wanted to create new table in sales schema called regional_sales so Kevin grants the create table permissions to Steve. Steve creates the new table called regional_sales in sales schema, who is the owner of the table regional_sales

**A.** Kevin is the owner of sales schema, all the tables in the schema will be owned by Kevin
**B.** Steve is the owner of the table
**C.** By default ownership is assigned DBO
**D.** By default ownership is assigned to DEFAULT_OWNER
**E.** Kevin and Smith both are owners of table
**Answer: B (LEAVE A REPLY)**
Explanation
A user who creates the object becomes its owner, does not matter who is the owner of the parent object.

**NEW QUESTION: 69**
A data engineer wants to horizontally combine two tables as a part of a query. They want to use a shared
column as a key column, and they only want the query result to contain rows whose value in the key column is
present in both tables.
Which of the following SQL commands can they use to accomplish this task?
**A.** OUTER JOIN
**B.** INNER JOIN
**C.** MERGE
**D.** UNION
**E.** LEFT JOIN
**Answer: (SHOW ANSWER)**

**NEW QUESTION: 70**
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
**A.** if department == "supply chain" or process:
**B.** if department == "supply chain" or process = TRUE:
**C.** if department == "supply chain" | process == TRUE:
**D.** if department == "supply chain" | if process == TRUE:
**E.** if department = "supply chain" | process:
**Answer: (SHOW ANSWER)**

**NEW QUESTION: 71**
What is the purpose of a silver layer in Multi hop architecture?
**A.** Replaces a traditional data lake
**B.** Efficient storage and querying of full and unprocessed history of data
**C.** A schema is enforced, with data quality checks.
**D.** Refined views with aggregated data
**E.** Optimized query performance for business-critical data

**Answer: C (LEAVE A REPLY)**

Explanation

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

Medallion Architecture - Databricks

Silver Layer:

1.Reduces data storage complexity, latency, and redundency

2.Optimizes ETL throughput and analytic query performance

3.Preserves grain of original data (without aggregation)

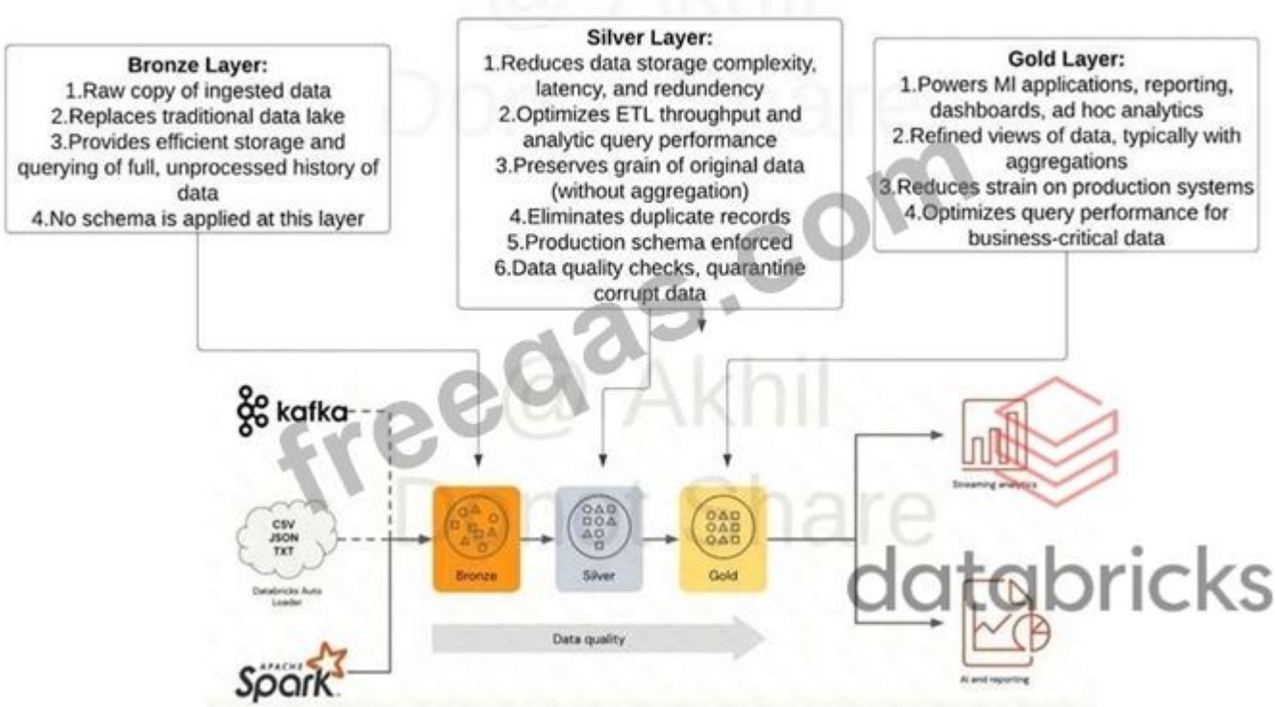
4.Eliminates duplicate records

5.production schema enforced

6.Data quality checks, quarantine corrupt data

Exam focus: Please review the below image and understand the role of each layer(bronze, silver, gold) in medallion architecture, you will see varying questions targeting each layer and its purpose.

Sorry I had to add the watermark some people in Udemy are copying my content.




**NEW QUESTION: 72**

You were asked to create or overwrite an existing delta table to store the below transaction data.

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

**A.** 1.CREATE OR REPLACE DELTA TABLE transactions (

2.transactionId int,

3.transactionDate timestamp,

4.unitsSold int)
**B.** 1.CREATE OR REPLACE TABLE IF EXISTS transactions (
2.transactionId int,
3.transactionDate timestamp,
4.unitsSold int)
5.FORMAT DELTA
**C.** 1.CREATE IF EXSITS REPLACE TABLE transactions (
2.transactionId int,
3.transactionDate timestamp,
4.unitsSold int)
**D.** 1.CREATE OR REPLACE TABLE transactions (
2.transactionId int,
3.transactionDate timestamp,
4.unitsSold int)
**Answer: D (LEAVE A REPLY)**
Explanation
The answer is
1.CREATE OR REPLACE TABLE transactions (
2.transactionId int,
3.transactionDate timestamp,
4.unitsSold int)
When creating a table in Databricks by default the table is stored in DELTA format.

**NEW QUESTION: 73**
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 _____
**A.** Python, Job, SQL Cluster, email address
**B.** SQL, Alert, Refresh, email address
**C.** SQL, Job, SQL Cluster, email address
**D.** SQL, Job, Refresh, email address
**E.** Python, Job, Refresh, email address
**Answer: B (LEAVE A REPLY)**
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.

1.Do one of the following:

*Click Create in the sidebar and select Alert.

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

2.Search for a target query.

Graphical user interface, text, application Description automatically generated






To alert on multiple columns, you need to modify your query. See Alert on multiple col-umns.

3.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.






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.

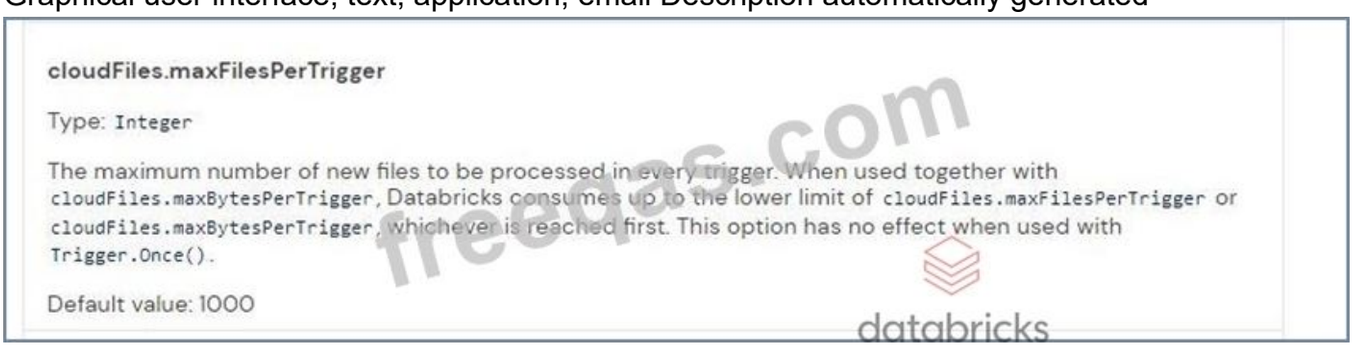
4.In the When triggered, send notification field, select how many notifications are sent when your alert is triggered:
*Just once: Send a notification when the alert status changes from OK to TRIGGERED.
*Each time alert is evaluated: Send a notification whenever the alert status is TRIGGERED regardless of its status at the previous evaluation.
*At most every: Send a notification whenever the alert status is TRIGGERED at a spe-cific interval. This choice lets you avoid notification spam for alerts that trigger of-ten.
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.
5.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.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}}.
b.Click the Preview toggle button to preview the rendered result.
Important
The preview is useful for verifying that template variables are rendered cor-rectly. It is not an accurate representation of the eventual notification content, as each alert destination can display notifications differently.
c.Click the Save Changes button.
6.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 cre-dential on the alert's refresh schedule.
*If the query is a Run as viewer query, the query runs using the alert creator's cre-dential on the alert's refresh schedule.

7.Click Create Alert.
8.Choose an alert destination.
Important
If you skip this step you will not be notified when the alert is triggered.




**NEW QUESTION: 74**
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 used to start the cluster in a timely fashion so your job can run immediatley?
**A.** Setup an additional job to run ahead of the actual job so the cluster is running second job starts
**B.** Use the Databricks cluster pools feature to reduce the startup time
**C.** Use Databricks Premium edition instead of Databricks standard edition
**D.** Pin the cluster in the cluster UI page so it is always available to the jobs
**E.** Disable auto termination so the cluster is always running
**Answer: B (LEAVE A REPLY)**
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 a pool and follow some best practices,
Graphical user interface, text Description automatically generated


**NEW QUESTION: 75**
You had AUTO LOADER to process millions of files a day and noticed slowness in load process, so you scaled up the Databricks cluster but realized the performance of the Auto loader is still not improving, what is the best way to resolve this.
**A.** AUTO LOADER is not suitable to process millions of files a day
**B.** Setup a second AUTO LOADER process to process the data
**C.** Increase the maxFilesPerTrigger option to a sufficiently high number
**D.** Copy the data from cloud storage to local disk on the cluster for faster access
**E.** Merge files to one large file
**Answer: (SHOW ANSWER)**
Explanation
The default value of maxFilesPerTrigger is 1000 it can be increased to a much higher number but will require a much larger compute to process.
Graphical user interface, text, application, email Description automatically generated




https://docs.databricks.com/ingestion/auto-loader/options.html

**NEW QUESTION: 76**

A data engineering manager has noticed that each of the queries in a Databricks SQL dashboard takes a few
minutes to update when they manually click the "Refresh" button. They are curious why this might be occurring, so a team member provides a variety of reasons on why the delay might be occurring.
Which of the following reasons fails to explain why the dashboard might be taking a few minutes to update?
**A.** The queries attached to the dashboard might take a few minutes to run under normal circumstances
**B.** The queries attached to the dashboard might all be connected to their own, unstarted Databricks clusters
**C.** The SQL endpoint being used by each of the queries might need a few minutes to start up
**D.** The Job associated with updating the dashboard might be using a non-pooled endpoint
**E.** The queries attached to the dashboard might first be checking to determine if new data is available
**Answer: D (LEAVE A REPLY)**



**NEW QUESTION: 77**
You are trying to calculate total sales made by all the employees by parsing a complex struct data type that stores employee and sales data, how would you approach this in SQL Table definition, batchId INT, performance ARRAY<STRUCT<employeeId: BIGINT, sales: INT>>, in-sertDate TIMESTAMP Sample data of performance column
1.[
2.{ "employeeId":1234
3."sales" : 10000},
4.
5.{ "employeeId":3232
6."sales" : 30000}
7.]
Calculate total sales made by all the employees?
Sample data with create table syntax for the data:
1.create or replace table sales as
2.select 1 as batchId ,

3.from_json('[{ "employeeId":1234,"sales" : 10000 },{ "employeeId":3232,"sales" : 30000 }]',
4. 'ARRAY<STRUCT<employeeId: BIGINT, sales: INT>>') as performance,
5. current_timestamp() as insertDate
6.union all
7.select 2 as batchId ,
8. from_json('[{ "employeeId":1235,"sales" : 10500 },{ "employeeId":3233,"sales" : 32000 }]',
9. 'ARRAY<STRUCT<employeeId: BIGINT, sales: INT>>') as performance,
10. current_timestamp() as insertDate

**A.** 1.WITH CTE as (SELECT EXPLODE (performance) FROM table_name)
2.SELECT SUM (performance.sales) FROM CTE

**B.** 1.WITH CTE as (SELECT FLATTEN (performance) FROM table_name)
2.SELECT SUM (sales) FROM CTE

**C.** 1.select aggregate(flatten(collect_list(performance.sales)), 0, (x, y) -> x + y)
2.as total_sales from sales

**D.** SELECT SUM(SLICE (performance, sales)) FROM employee

**E.** 1.select reduce(flatten(collect_list(performance:sales)), 0, (x, y) -> x + y)
2.as total_sales from sales

**Answer: C (LEAVE A REPLY)**

Explanation

The answer is

1.select aggregate(flatten(collect_list(performance.sales)), 0, (x, y) -> x + y)
2.as total_sales from sales

Nested Struct can be queried using the . notation performance.sales will give you access to all the sales values in the performance column.

Note: option D is wrong because it uses performance:sales not performance.sales. ":" this is only used when referring to JSON data but here we are dealing with a struct data type. for the exam please make sure to understand if you are dealing with JSON data or Struct data.

```
select performance as  total_sales from sales
```

```
row 1 -    [{"employeeId":1235,"sales":10500},{"employeeId":3233,"sales":32000}]
row 2 -    [{"employeeId":1234,"sales":10000},{"employeeId":3232,"sales":30000}]
```

```
select performance.sales as  total_sales from sales -- selects sales values
```

```
row 1 -   [10500,32000]
row 2 -   [10000,30000]
```

```
select collect_list(performance.sales) as  total_sales from sales
```

```
[[10500,32000],[10000,30000]]
```

```
select flatten(collect_list(performance.sales)) as  total_sales from sales
```

```
[10500,32000,10000,30000]
```

```
select aggregate(
                flatten(collect_list(performance.sales))
                , 0 -- starting value
                , (x, y) -> x + y -- add every two elements in the array until it results a single valu
              ) as  total_sales from sales
```

```
82500
```



Other solutions:

we can also use reduce instead of aggregate

select reduce(flatten(collect_list(performance.sales)), 0, (x, y) -> x + y) as total_sales from sales we can also use explode and sum instead of using any higher-order funtions.

1.with cte as (

2. select

3. explode(flatten(collect_list(performance.sales))) sales from sales

4.)

5.select

6. sum(sales) from cte

Sample data with create table syntax for the data:

1.create or replace table sales as

2.select 1 as batchId ,

3.from_json('[{ "employeeId":1234,"sales" : 10000 },{ "employeeId":3232,"sales" : 30000 }]',

4. 'ARRAY<STRUCT<employeeId: BIGINT, sales: INT>>') as performance,

5. current_timestamp() as insertDate

6.union all

7.select 2 as batchId ,

8. from_json('[{ "employeeId":1235,"sales" : 10500 },{ "employeeId":3233,"sales" : 32000 }]',
9. 'ARRAY<STRUCT<employeeId: BIGINT, sales: INT>>') as performance,
10. current_timestamp() as insertDate

**NEW QUESTION: 78**
Which of the following SQL command can be used to insert or update or delete rows based on a condition to check if a row(s) exists?
**A.** MERGE INTO table_name
**B.** COPY INTO table_name
**C.** UPDATE table_name
**D.** INSERT INTO OVERWRITE table_name
**E.** INSERT IF EXISTS table_name
**Answer: A (LEAVE A REPLY)**
Explanation
here is the additional documentation for your review.
https://docs.databricks.com/spark/latest/spark-sql/language-manual/delta-merge-into.html
1.MERGE INTO target_table_name [target_alias]
2. USING source_table_reference [source_alias]
3. ON merge_condition
4. [ WHEN MATCHED [ AND condition ] THEN matched_action ] [...]
5. [ WHEN NOT MATCHED [ AND condition ] THEN not_matched_action ] [...]
6.
7.matched_action
8. { DELETE |
9. UPDATE SET * |
10. UPDATE SET { column1 = value1 } [, ...] }
11.
12.not_matched_action
13. { INSERT * |
14. INSERT (column1 [, ...] ) VALUES (value1 [, ...])

**NEW QUESTION: 79**
The data engineering team is using a SQL query to review data completeness every day to monitor the ETL job, and query output is being used in multiple dashboards which of the following ap-proaches can be used to set up a schedule and automate this process?
**A.** They can schedule the query to run every day from the Jobs UI.
**B.** They can schedule the query to refresh every day from the query's page in Databricks SQL
**C.** They can schedule the query to run every 12 hours from the Jobs UI.
**D.** They can schedule the query to refresh every day from the SQL endpoint's page in Databricks SQL.
**E.** They can schedule the query to refresh every 12 hours from the SQL endpoint's page in Databricks SQL

**Answer: B (LEAVE A REPLY)**

Explanation

The answer is They can schedule the query to refresh every 12 hours from the SQL endpoint's page in Databricks SQL, The query pane view in Databricks SQL workspace provides the ability to add or edit and schedule individual queries to run.

You can use scheduled query executions to keep your dashboards updated or to enable routine alerts. By default, your queries do not have a schedule.

Note

If your query is used by an alert, the alert runs on its own refresh schedule and does not use the query schedule.

To set the schedule:

* Click the query info tab.
* Graphical user interface, text, application, email Description automatically generated
* Click the link to the right of Refresh Schedule to open a picker with schedule intervals.
* Graphical user interface, application Description automatically generated
* 3.Set the schedule.
* The picker scrolls and allows you to choose:
* *An interval: 1-30 minutes, 1-12 hours, 1 or 30 days, 1 or 2 weeks
* *A time. The time selector displays in the picker only when the interval is greater than 1 day and the day selection is greater than 1 week. When you schedule a specific time, Databricks SQL takes input in your computer's timezone and converts it to UTC. If you want a query to run at a certain time in UTC, you must adjust the picker by your local offset. For example, if you want a query to execute at 00:00 UTC each day, but your current timezone is PDT (UTC-7), you should select 17:00 in the picker:
* Graphical user interface Description automatically generated

**NEW QUESTION: 80**

A data engineer has set up a notebook to automatically process using a Job. The data engineer's manager wants
to version control the schedule due to its complexity.

Which of the following approaches can the data engineer use to obtain a version-controllable configuration of
the Job's schedule?

**A.** They can link the Job to notebooks that are a part of a Databricks Repo
**B.** They can submit the Job once on a Job cluster
**C.** They can submit the Job once on an all-purpose cluster
**D.** They can download the JSON description of the Job from the Job's page
**E.** They can download the XML description of the Job from the Job's page

**Answer: D (LEAVE A REPLY)**

**NEW QUESTION: 81**

Projecting a multi-dimensional dataset onto which vector has the greatest variance?

**A.** first principal component
**B.** first eigenvector
**C.** not enough information given to answer
**D.** second eigenvector
**E.** second principal component
**Answer: (SHOW ANSWER)**
Explanation
The method based on principal component analysis (PCA) evaluates the features according to the projection of
the largest eigenvector of the correlation matrix on the initial dimensions, the method based on Fisher's linear
discriminant analysis evaluates. Them according to the magnitude of the components of the discriminant
vector.
The first principal component corresponds to the greatest variance in the data, by definition. If we project the
data onto the first principal component line, the data is more spread out (higher variance) than if projected onto
any other line, including other principal components.

**NEW QUESTION: 82**
Once a cluster is deleted, below additional actions need to performed by the administrator
**A.** Remove virtual machines but storage and networking are automatically dropped
**B.** Drop storage disks but Virtual machines and networking are automatically dropped
**C.** Remove networking but Virtual machines and storage disks are automatically dropped
**D.** Remove logs
**E.** No action needs to be performed. All resources are automatically removed.
**Answer: E (LEAVE A REPLY)**
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

**NEW QUESTION: 83**
How do you handle failures gracefully when writing code in Pyspark, fill in the blanks to complete the below statement
1._____
2.
3.
Spark.read.table("table_name").select("column").write.mode("append").SaveAsTable("new_table_name")
4.
5._____
6.
7. print(f"query failed")
**A.** try: failure:
**B.** try: catch:
**C.** try: except:
**D.** try: fail:
**E.** try: error:
**Answer: C (LEAVE A REPLY)**
Explanation
The answer is try: and except:

**NEW QUESTION: 84**
A data engineer needs to dynamically create a table name string using three Python varia-bles: region, store,
and year. An example of a table name is below when region = "nyc", store = "100", and year = "2021": nyc100_sales_2021
Which of the following commands should the data engineer use to construct the table name in Py-thon?
**A.** f"{region}+{store}+_sales_+{year}"
**B.** "{region}{store}_sales_{year}"
**C.** "{region}+{store}+"_sales_"+{year}"
**D.** "{region}+{store}+_sales_+{year}"
**E.** f"{region}{store}_sales_{year}"
**Answer: E (LEAVE A REPLY)**

**NEW QUESTION: 85**
You are asked to set up an alert to notify in an email every time a KPI indicater increases beyond a threshold value, team also asked you to include the actual value in the alert email notification.
**A.** Use notebook and python code to run every minute, using python variables to capture send the information in an email
**B.** Setup an alert but use the default template to notify the message in email's subject
**C.** Setup an alert but use the custom template to notify the message in email's subject
**D.** Use the webhook destination instead so alert message can be customized

**E.** Use custom email hook to customize the message

**Answer: (SHOW ANSWER)**

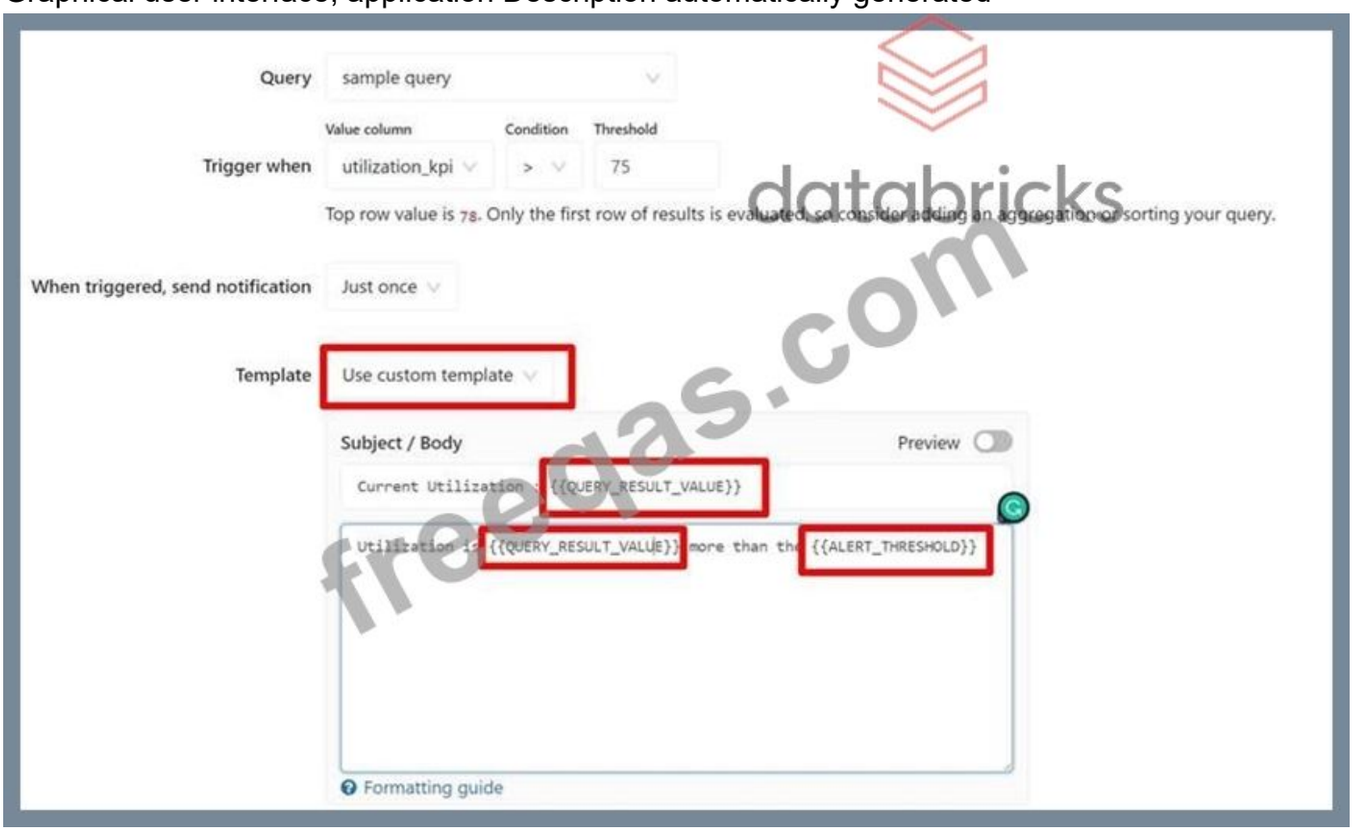
Explanation

Alerts support custom template supports using variables to customize the default message, set up the query to compare the KPI current value to the threshold and use the variable QUERY_RESULT_VALUE to display the value in the email notification.

here is a simple alert, that uses variables in the custom template to present these values in the email notification message, when the alert is fired these variables get replaced with actual values.

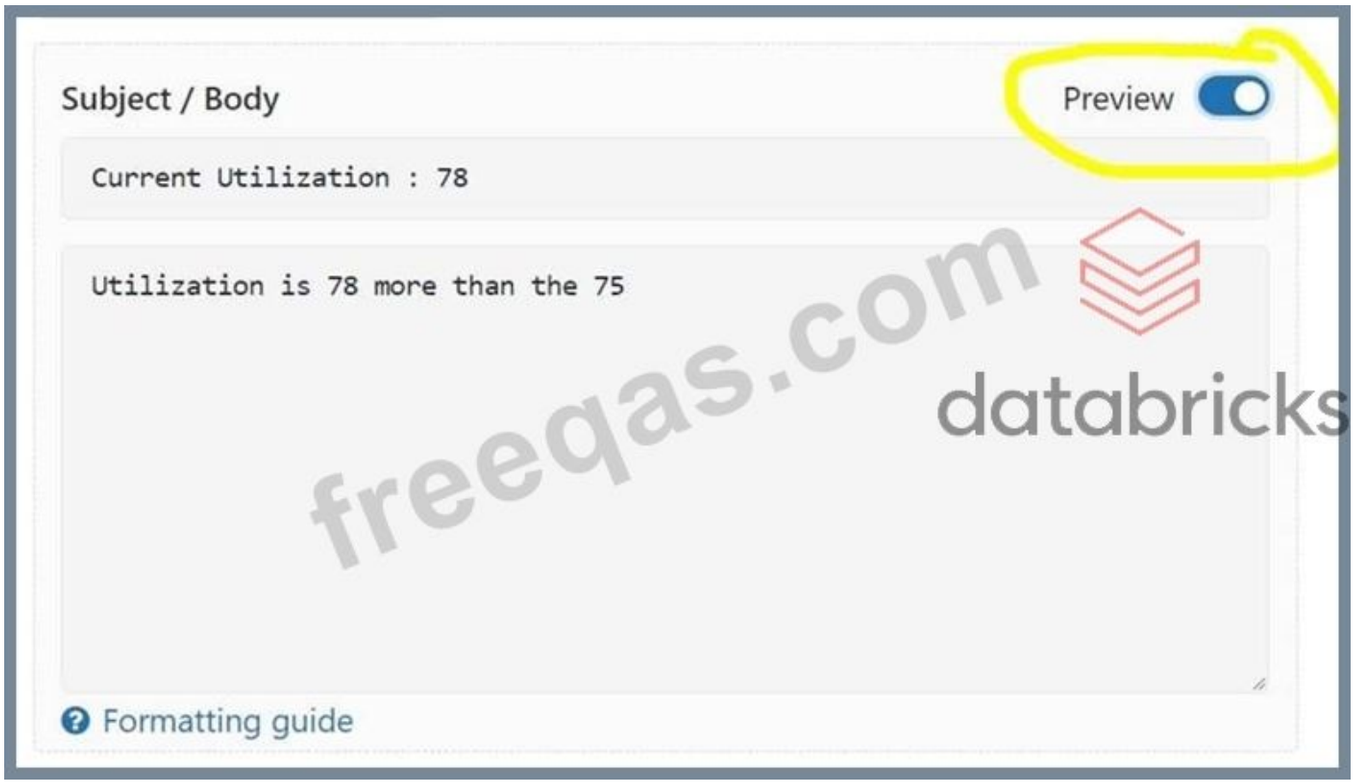
Alert with custom template

Graphical user interface, application Description automatically generated




When you enable preview you can see how the alert looks when you substitute the variables.

Graphical user interface, text, application, email Description automatically generated




Below are additional template variables available to you with the custom template.

Alerts | Databricks on AWS

Graphical user interface, text, application, email Description automatically generated

5. 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. 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}}`.

  b. Click the **Preview** toggle button to preview the rendered result.



**NEW QUESTION: 86**

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 |

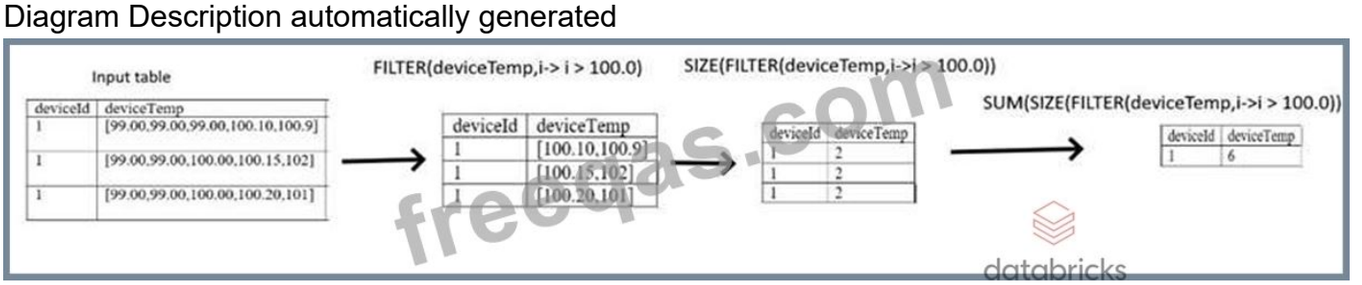
SELECT deviceId, __ (__ (__(deviceTemp], i -> i > 100.00)))
FROM devices
GROUP BY deviceId

**A.** SUM, COUNT, SIZE
**B.** SUM, SIZE, SLICE
**C.** SUM, SIZE, ARRAY_CONTAINS
**D.** SUM, SIZE, ARRAY_FILTER
**E.** SUM, SIZE, FILTER

**Answer: E (LEAVE A REPLY)**

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
Diagram Description automatically generated




**NEW QUESTION: 87**

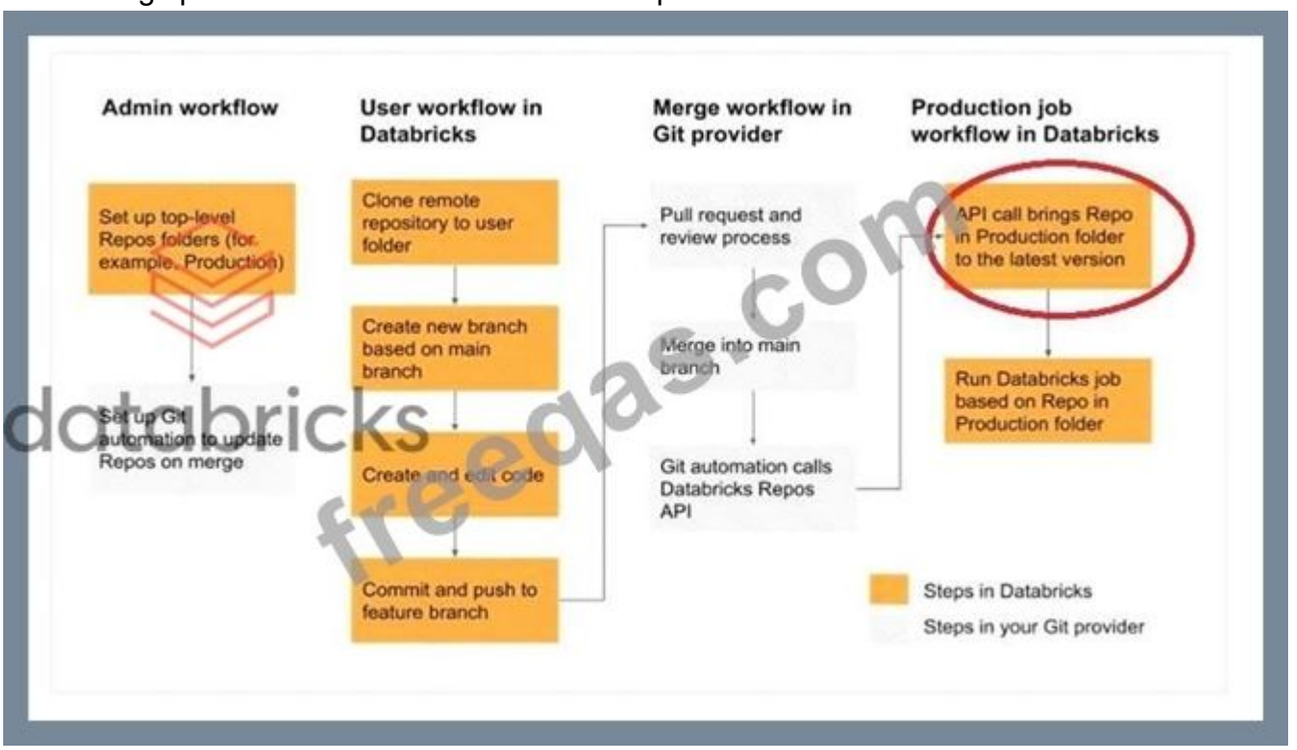
Which of the following developer operations in CI/CD flow can be implemented in Databricks Re-pos?

**A.** Merge when code is committed
**B.** Pull request and review process
**C.** Trigger Databricks Repos API to pull the latest version of code into production folder
**D.** Resolve merge conflicts
**E.** Delete a branch

**Answer: C (LEAVE A REPLY)**

Explanation

See the 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




**NEW QUESTION: 88**
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 |

**A.** 3
**B.** 2
(Correct)
**C.** 1
**D.** 0
**E.** NULL
**Answer: (SHOW ANSWER)**
Explanation
The answer is 2,
Count(DISTINCT *) removes rows with any column with a NULL value

**NEW QUESTION: 89**

You are working to set up two notebooks to run on a schedule, the second notebook is dependent on the first notebook but both notebooks need different types of compute to run in an optimal fashion, what is the best way to set up these notebooks as jobs?

**A.** Use DELTA LIVE PIPELINES instead of notebook tasks

**B.** A Job can only use single cluster, setup job for each notebook and use job dependency to link both jobs together

**C.** Each task can use different cluster, add these two notebooks as two tasks in a single job with linear dependency and modify the cluster as needed for each of the tasks

**D.** Use a single job to setup both notebooks as individual tasks, but use the cluster API to setup the second cluster before the start of second task

**E.** Use a very large cluster to run both the tasks in a single job

**Answer: C (LEAVE A REPLY)**

Explanation

Tasks in Jobs support different clusters for each task in the same job.

**NEW QUESTION: 90**

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

**A.** Efficiently process new data incrementally from cloud object storage

**B.** Efficiently move data incrementally from one delta table to another delta table

**C.** Incrementally process new data from streaming data sources like Kafka into delta lake

**D.** Incrementally process new data from relational databases like MySQL

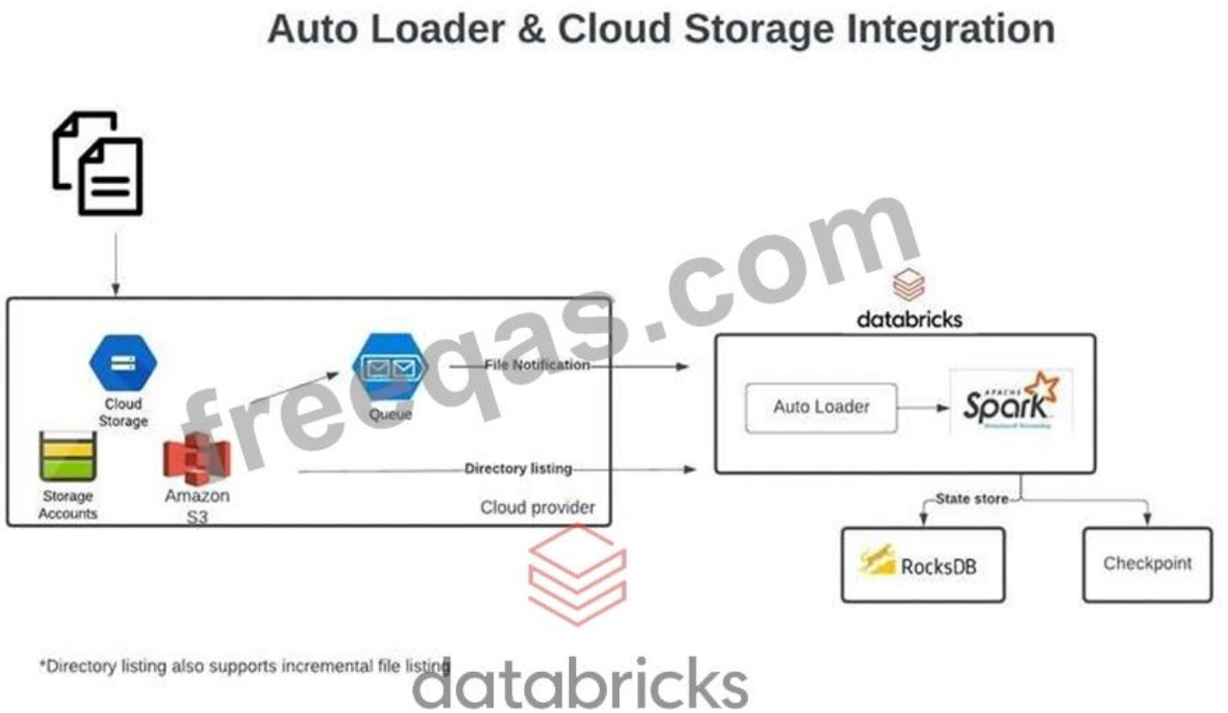
**E.** Efficiently copy data from one data lake location to another data lake location

**Answer: A (LEAVE A REPLY)**

Explanation

The answer is, Efficiently process new data incrementally from cloud object storage, AU-TO LOADER only supports ingesting files stored in a cloud object storage. Auto Loader cannot process streaming data sources like Kafka or Delta streams, use Structured streaming for these data sources.

Diagram Description automatically generated




Auto Loader and Cloud Storage Integration

Auto Loader supports a couple of ways to ingest data incrementally

1.Directory listing - List Directory and maintain the state in RocksDB, supports incremental file listing

2.File notification - Uses a trigger+queue to store the file notification which can be later used to retrieve the file, unlike Directory listing File notification can scale up to millions of files per day.

[OPTIONAL]

Auto Loader vs COPY INTO?

Auto Loader

Auto Loader incrementally and efficiently processes new data files as they arrive in cloud storage without any additional setup. Auto Loader provides a new 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.

When to use Auto Loader instead of the COPY INTO?

*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.

*You do not plan to load subsets of previously uploaded files. With Auto Loader, it can be more difficult to reprocess subsets of files. However, you can use the COPY INTO SQL command to reload subsets of files while an Auto Loader stream is simultaneously running.

**NEW QUESTION: 91**

You are working on a marketing team request to identify customers with the same information between two tables CUSTOMERS_2021 and CUSTOMERS_2020 each table contains 25 columns with the

same schema, You are looking to identify rows that match between two tables across all columns, which of the following can be used to perform in SQL

**A.** 1.SELECT * FROM CUSTOMERS_2021
2. UNION
3.SELECT * FROM CUSTOMERS_2020

**B.** 1.SELECT * FROM CUSTOMERS_2021
2. UNION ALL
3.SELECT * FROM CUSTOMERS_2020

**C.** 1.SELECT * FROM CUSTOMERS_2021 C1
2.INNER JOIN CUSTOMERS_2020 C2
3.ON C1.CUSTOMER_ID = C2.CUSTOMER_ID

**D.** 1.SELECT * FROM CUSTOMERS_2021
2. INTERSECT
3.SELECT * FROM CUSTOMERS_2020

**E.** 1.SELECT * FROM CUSTOMERS_2021
2.EXCEPT
3.SELECT * FROM CUSTOMERS_2020

**Answer: D (LEAVE A REPLY)**

Explanation

Answer is,

1.SELECT * FROM CUSTOMERS_2021
2. INTERSECT
3.SELECT * FROM CUSTOMERS_2020

To compare all the rows between both the tables across all the columns using intersect will help us achieve that, an inner join is only going to check if the same column value exists across both the tables on a single column.

INTERSECT [ALL | DISTINCT]

*Returns the set of rows which are in both subqueries.

If ALL is specified a row that appears multiple times in the subquery1 as well as in subquery will be returned multiple times.

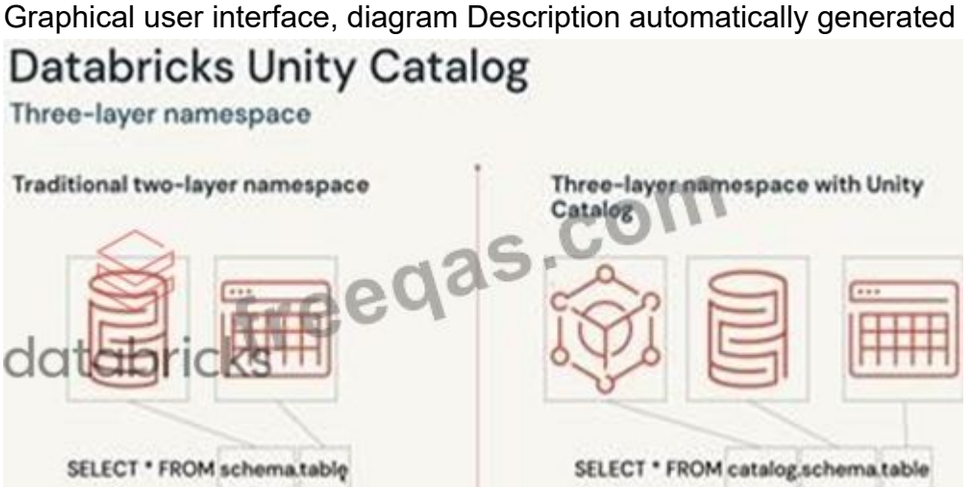
If DISTINCT is specified the result does not contain duplicate rows. This is the default.

**NEW QUESTION: 92**
Delete records from the transactions Delta table where transactionDate is greater than current timestamp?
**A.** DELETE FROM transactions where transactionDate > current_timestamp()
**B.** DELETE FROM transactions FORMAT DELTA where transactionDate > cur-renct_timestmap()
**C.** DELETE FROM transactions where transactionDate > current_timestamp() KEEP_HISTORY
**D.** DELETE FROM transactions if transctionDate > current_timestamp()
**E.** DELET FROM transactions where transactionDate GE current_timestamp()
**Answer: (SHOW ANSWER)**

**NEW QUESTION: 93**
Question-3: In machine learning, feature hashing, also known as the hashing trick (by analogy to the kernel
trick), is a fast and space-efficient way of vectorizing features (such as the words in a language), i.e., turning
arbitrary features into indices in a vector or matrix. It works by applying a hash function to the features and
using their hash values modulo the number of features as indices directly, rather than looking the indices up in
an associative array. So what is the primary reason of the hashing trick for building classifiers?
**A.** It creates the smaller models
**B.** It requires the lesser memory to store the coefficients for the model
**C.** It reduces the non-significant features e.g. punctuations
**D.** Noisy features are removed
**Answer: B (LEAVE A REPLY)**
Explanation
This hashed feature approach has the distinct advantage of requiring less memory and one less pass through
the training data, but it can make it much harder to reverse engineer vectors to determine which original feature mapped to a vector location. This is because multiple features may hash to the same location. With
large vectors or with multiple locations per feature, this isn't a problem for accuracy but it can make it hard to
understand what a classifier is doing.
Models always have a coefficient per feature, which are stored in memory during model building. The hashing
trick collapses a high number of features to a small number which reduces the number of coefficients and thus

memory requirements. Noisy features are not removed; they are combined with other features and so still have
an impact.
The validity of this approach depends a lot on the nature of the features and problem domain; knowledge of
the domain is important to understand whether it is applicable or will likely produce poor results. While hashing features may produce a smaller model, it will be one built from odd combinations of real-world features, and so will be harder to interpret.
An additional benefit of feature hashing is that the unknown and unbounded vocabularies typical of word-like
variables aren't a problem.

**NEW QUESTION: 94**
How do you access or use tables in the unity catalog?
**A.** schema_name.table_name
**B.** schema_name.catalog_name.table_name
**C.** catalog_name.table_name
**D.** catalog_name.database_name.schema_name.table_name
**E.** catalog_name.schema_name.table_name
**Answer: (SHOW ANSWER)**
Explanation
The answer is catalog_name.schema_name.table_name
Graphical user interface, diagram Description automatically generated




Note: Database and Schema are analogous they are interchangeably used in the Unity catalog.
FYI, A catalog is registered under a metastore, by default every workspace has a default metastore called hive_metastore, with a unity catalog you have the ability to create meatstores and share that across multiple workspaces.

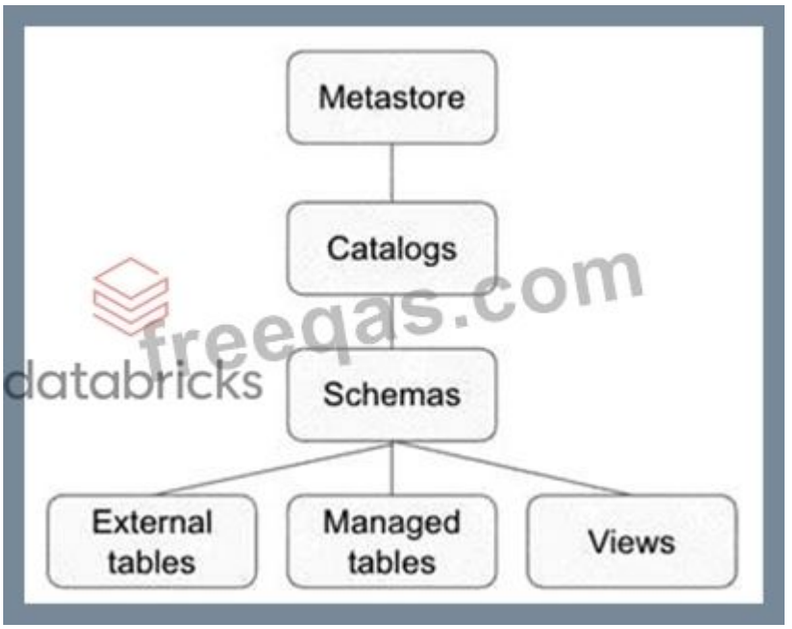


Diagram Description automatically generated

**NEW QUESTION: 95**
Data engineering team has provided 10 queries and asked Data Analyst team to build a dashboard and refresh the data every day at 8 AM, identify the best approach to set up data refresh for this dashaboard?
**A.** Each query requires a separate task and setup 10 tasks under a single job to run at 8 AM to refresh the dashboard
**B.** The entire dashboard with 10 queries can be refreshed at once, single schedule needs to be set up to refresh at 8 AM.
**C.** Setup JOB with linear dependency to all load all 10 queries into a table so the dashboard can be refreshed at once.
**D.** A dashboard can only refresh one query at a time, 10 schedules to set up the refresh.
**E.** Use Incremental refresh to run at 8 AM every day.
**Answer: (SHOW ANSWER)**
Explanation
The answer is,
The entire dashboard with 10 queries can be refreshed at once, single schedule needs to be set up to refresh at
8 AM.
Automatically refresh a dashboard
A dashboard's owner and users with the Can Edit permission can configure a dashboard to automatically refresh on a schedule. To automatically refresh a dashboard:
* Click the Schedule button at the top right of the dashboard. The scheduling dialog appears.
* Graphical user interface, text, application, email, Teams Description automatically generated
* 2.In the Refresh every drop-down, select a period.
* 3.In the SQL Warehouse drop-down, optionally select a SQL warehouse to use for all the queries.

If you don't select a warehouse, the queries execute on the last used SQL ware-house.
* 4.Next to Subscribers, optionally enter a list of email addresses to notify when the dashboard is automatically updated.
* Each email address you enter must be associated with a Azure Databricks account or con-figured as an alert destination.
* 5.Click Save. The Schedule button label changes to Scheduled.

**NEW QUESTION: 96**
Which of the following is true, when building a Databricks SQL dashboard?
**A.** A dashboard can only use results from one query
**B.** Only one visualization can be developed with one query result
**C.** A dashboard can only connect to one schema/Database
**D.** More than one visualization can be developed using a single query result
**E.** A dashboard can only have one refresh schedule
**Answer: D (LEAVE A REPLY)**
Explanation
the answer is, More than one visualization can be developed using a single query result.
In the query editor pane + Add visualization tab can be used for many visualizations for a single query result.
Graphical user interface, text, application Description automatically generated



**NEW QUESTION: 97**
A data architect has determined that a table of the following format is necessary:

Which of the following code blocks uses SQL DDL commands to create an empty Delta table in the above
format regardless of whether a table already exists with this name?
**A.** 1. CREATE TABLE table_name AS
2. SELECT id STRING, birthDate DATE, avgRating FLOAT
**B.** 1. CREATE OR REPLACE TABLE table_name ( id STRING, birthDate DATE, avgRating FLOAT )
**C.** 1. CREATE OR REPLACE TABLE table_name AS
2. SELECT id STRING, birthDate DATE, avgRating FLOAT USING DELTA
**D.** 1. CREATE TABLE IF NOT EXISTS table_name ( id STRING, birthDate DATE, avgRating FLOAT )
**E.** 1. CREATE OR REPLACE TABLE table_name
2. WITH COLUMNS ( id STRING, birthDate DATE, avgRating FLOAT ) USING DELTA
**Answer: B (LEAVE A REPLY)**

**NEW QUESTION: 98**
Which of the following benefits does Delta Live Tables provide for ELT pipelines over standard data pipelines
that utilize Spark and Delta Lake on Databricks?
**A.** The ability to access previous versions of data tables
**B.** The ability to perform batch and streaming queries
**C.** The ability to write pipelines in Python and/or SQL
**D.** The ability to automatically scale compute resources
**E.** The ability to declare and maintain data table dependencies
**Answer: E (LEAVE A REPLY)**

**NEW QUESTION: 99**
When defining external tables using formats CSV, JSON, TEXT, BINARY any query on the exter-nal tables caches the data and location for performance reasons, so within a given spark session any new files that may have arrived will not be available after the initial query. How can we address this limitation?
**A.** UNCACHE TABLE table_name
**B.** CACHE TABLE table_name
**C.** REFRESH TABLE table_name
**D.** BROADCAST TABLE table_name
**E.** CLEAR CACH table_name
**Answer: C (LEAVE A REPLY)**
Explanation
The answer is REFRESH TABLE table_name
REFRESH TABLE table_name will force Spark to refresh the availability of external files and any changes.
When spark queries an external table it caches the files associated with it, so that way if the table is queried again it can use the cached files so it does not have to retrieve them again from cloud object

storage, but the drawback here is that if new files are available Spark does not know until the Refresh command is ran.

**NEW QUESTION: 100**

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

**A.** Default location, DBFS:/user/

**B.** Default location, /user/db/

**C.** Default Storage account

**D.** Statement fails "Unable to create database without location"

**E.** Default Location, dbfs:/user/hive/warehouse

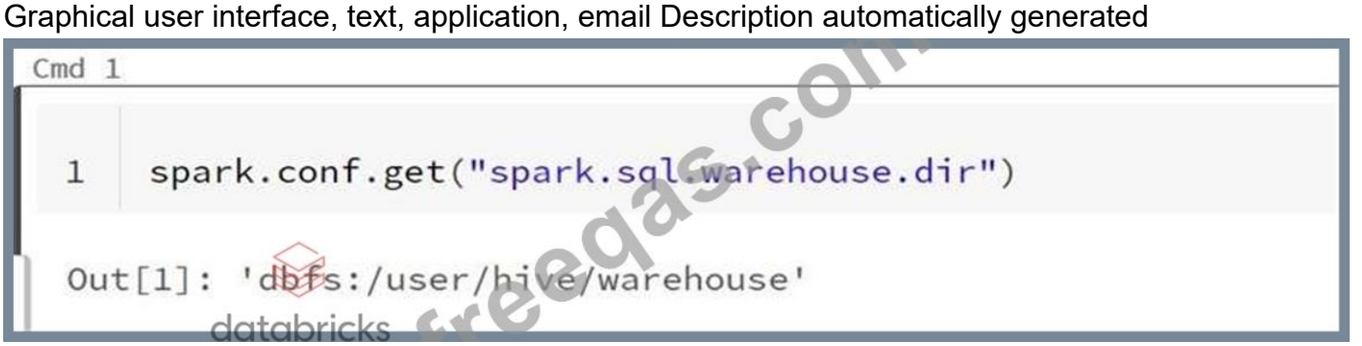
**Answer: E (LEAVE A REPLY)**

Explanation

The Answer is dbfs:/user/hive/warehouse this is the default location where spark stores user databases, the default can be changed using spark.sql.warehouse.dir a parameter. You can also provide a custom location using the LOCATION keyword.

Here is how this works,

Graphical user interface, text, application, email Description automatically generated

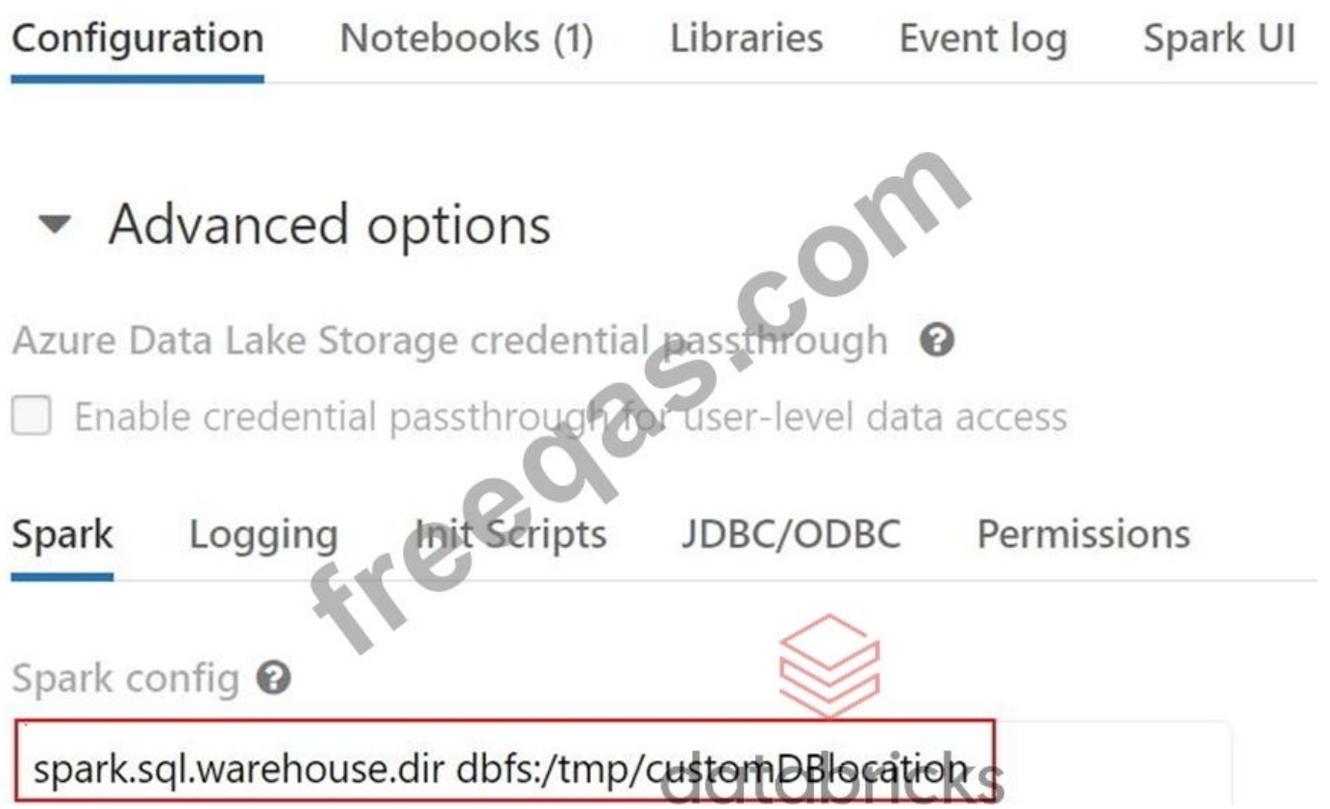


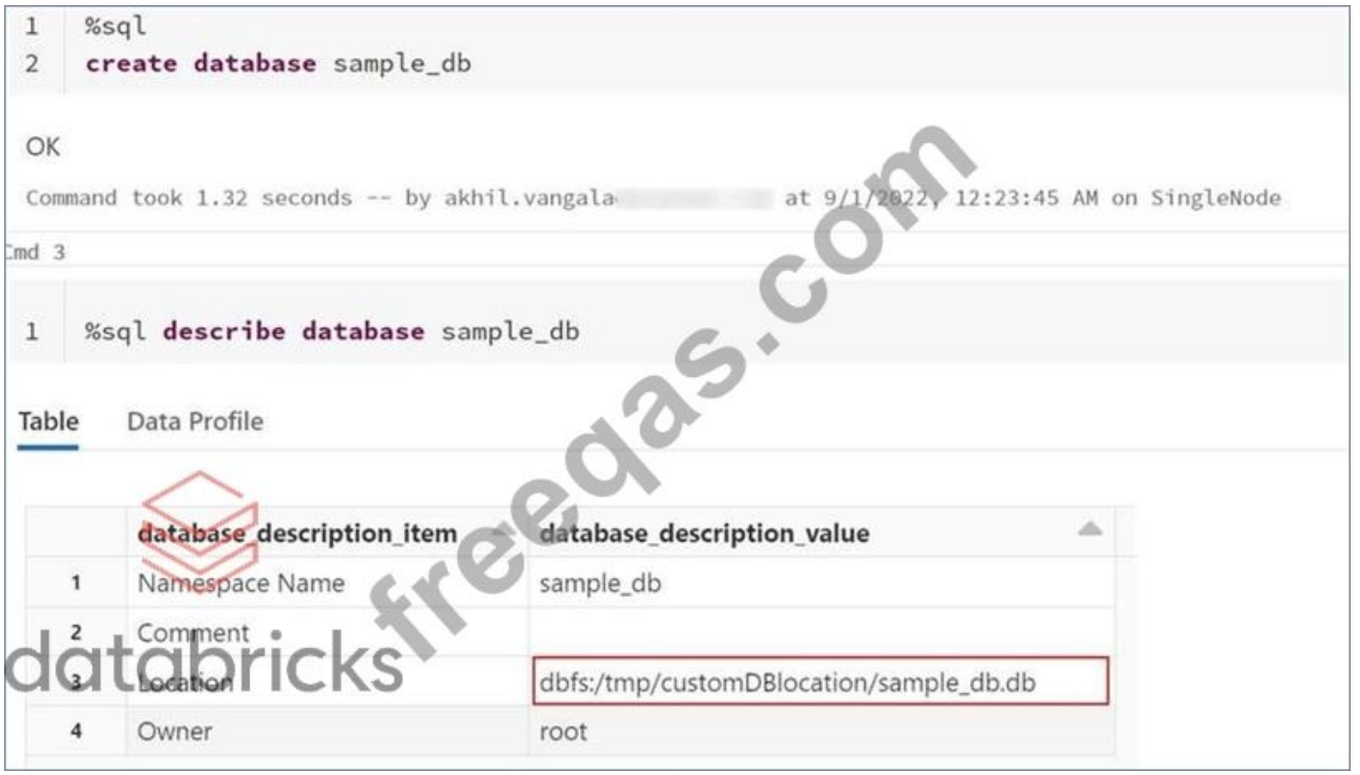

Default location


FYI, This can be changed used using cluster spark config or session config.

Modify spark.sql.warehouse.dir location to change the default location

Graphical user interface, text, application Description automatically generated








**NEW QUESTION: 101**
You are currently working on a notebook that will populate a reporting table for downstream process consumption, this process needs to run on a schedule every hour, what type of cluster are you going to use to set up this job?
**A.** Since it's just a single job and we need to run every hour, we can use an all-purpose cluster
**B.** The job cluster is best suited for this purpose.
**C.** Use Azure VM to read and write delta tables in Python
**D.** Use delta live table pipeline to run in continuous mode
**Answer: B (LEAVE A REPLY)**
Explanation
The answer is, The Job cluster is best suited for this purpose.
Since you don't need to interact with the notebook during the execution especially when it's a scheduled job, job cluster makes sense. Using an all-purpose cluster can be twice as expensive as a job cluster.
FYI,
When you run a job scheduler with option of creating a new cluster when the job is complete it terminates the cluster. You cannot restart a job cluster.

**NEW QUESTION: 102**
What is the main difference between the silver layer and the gold layer in medalion architecture?
**A.** Silver may contain aggregated data
**B.** Gold may contain aggregated data
**C.** Data quality checks are applied in gold
**D.** Silver is a copy of bronze data

**E.** God is a copy of silver data

**Answer: B (LEAVE A REPLY)**

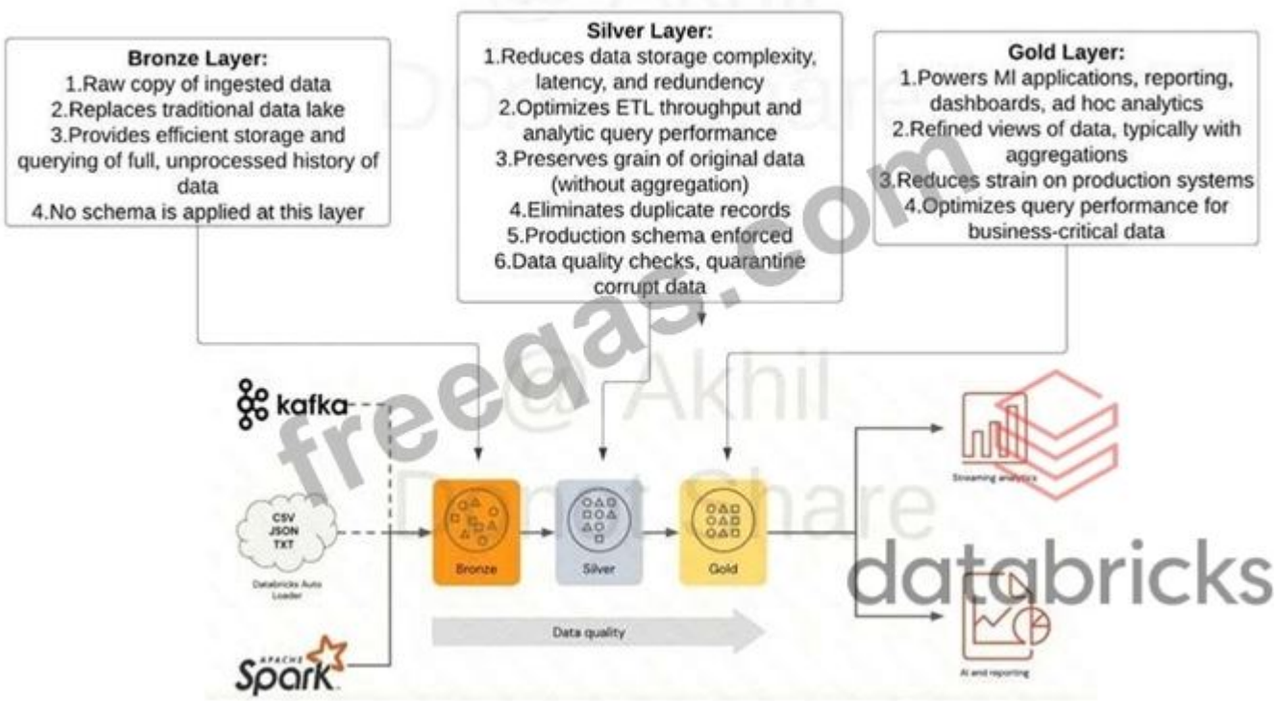
Explanation

Medallion Architecture - Databricks

Exam focus: Please review the below image and understand the role of each layer(bronze, silver, gold) in medallion architecture, you will see varying questions targeting each layer and its purpose.

Sorry I had to add the watermark some people in Udemy are copying my content.

A diagram of a house Description automatically generated with low confidence




**NEW QUESTION: 103**

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 an average of 8 minutes to start the cluster, what feature can be used to start the cluster in a timely fashion?

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

**B.** Use the Databricks cluster pools feature to reduce the startup time

**C.** Use Databricks Premium edition instead of Databricks standard edition

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

**E.** Disable auto termination so the cluster is always running

**Answer: B (LEAVE A REPLY)**

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

**NEW QUESTION: 104**
Suppose there are three events then which formula must always be equal to P(E1|E2,E3)?
**A.** P(E1,E2,E3)P(E1)/P(E2:E3)
**B.** P(E1,E2;E3)/P(E2,E3)
**C.** P(E1,E2|E3)P(E2|E3)P(E3)
**D.** P(E1,E2|E3)P(E3)
**E.** P(E1,E2,E3)P(E2)P(E3)
**Answer: (SHOW ANSWER)**
Explanation
This is an application of conditional probability: P(E1,E2)=P(E1|E2)P(E2). so
P(E1|E2) = P(E1.E2)/P(E2)
P(E1,E2,E3)/P(E2,E3)
If the events are A and B respectively, this is said to be "the probability of A given B"
It is commonly denoted by P(A|B):or sometimes PB(A). In case that both "A" and "B" are categorical variables, conditional probability table is typically used to represent the conditional probability.