# 2023 Realistic Verified Databricks-Certified-Professional-Data-Engineer exam dumps Q&As - Databricks-Certified-Professional-Data-Engineer Free Update [Q102-Q117

2023 Realistic Verified Databricks-Certified-Professional-Data-Engineer exam dumps Q&As - Databricks-Certified-Professional-Data-Engineer Free Update
Use Real Databricks-Certified-Professional-Data-Engineer Dumps - 100% Free Databricks-Certified-Professional-Data-Engineer Exam Dumps

**NEW QUESTION 102**

Create a sales database using the DBFS location &#8216;dbfs:/mnt/delta/databases/sales.db/&#8217;
- CREATE DATABASE sales FORMAT DELTA LOCATION &#8216;dbfs:/mnt/delta/databases/sales.db/&#8221;
- CREATE DATABASE sales USING LOCATION &#8216;dbfs:/mnt/delta/databases/sales.db/&#8217;
- CREATE DATABASE sales LOCATION &#8216;dbfs:/mnt/delta/databases/sales.db/&#8217;
- The sales database can only be created in Delta lake
- CREATE DELTA DATABASE sales LOCATION &#8216;dbfs:/mnt/delta/databases/sales.db/&#8217;

Explanation

The answer is

CREATE DATABASE sales LOCATION &#8216;dbfs:/mnt/delta/databases/sales.db/&#8217;

Note: with the introduction of the Unity catalog and three-layer namespace usage of SCHEMA and DATABASE is interchangeable

**NEW QUESTION 103**

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?
- Each query requires a separate task and setup 10 tasks under a single job to run at 8 AM to refresh the dashboard
- The entire dashboard with 10 queries can be refreshed at once, single schedule needs to be set up to refresh at 8 AM.
- Setup JOB with linear dependency to all load all 10 queries into a table so the dashboard can be refreshed at once.
- A dashboard can only refresh one query at a time, 10 schedules to set up the refresh.
- Use Incremental refresh to run at 8 AM every day.

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&#8217;s owner and users with the Can Edit permission can configure a dashboard to auto-matically 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&#8217;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.

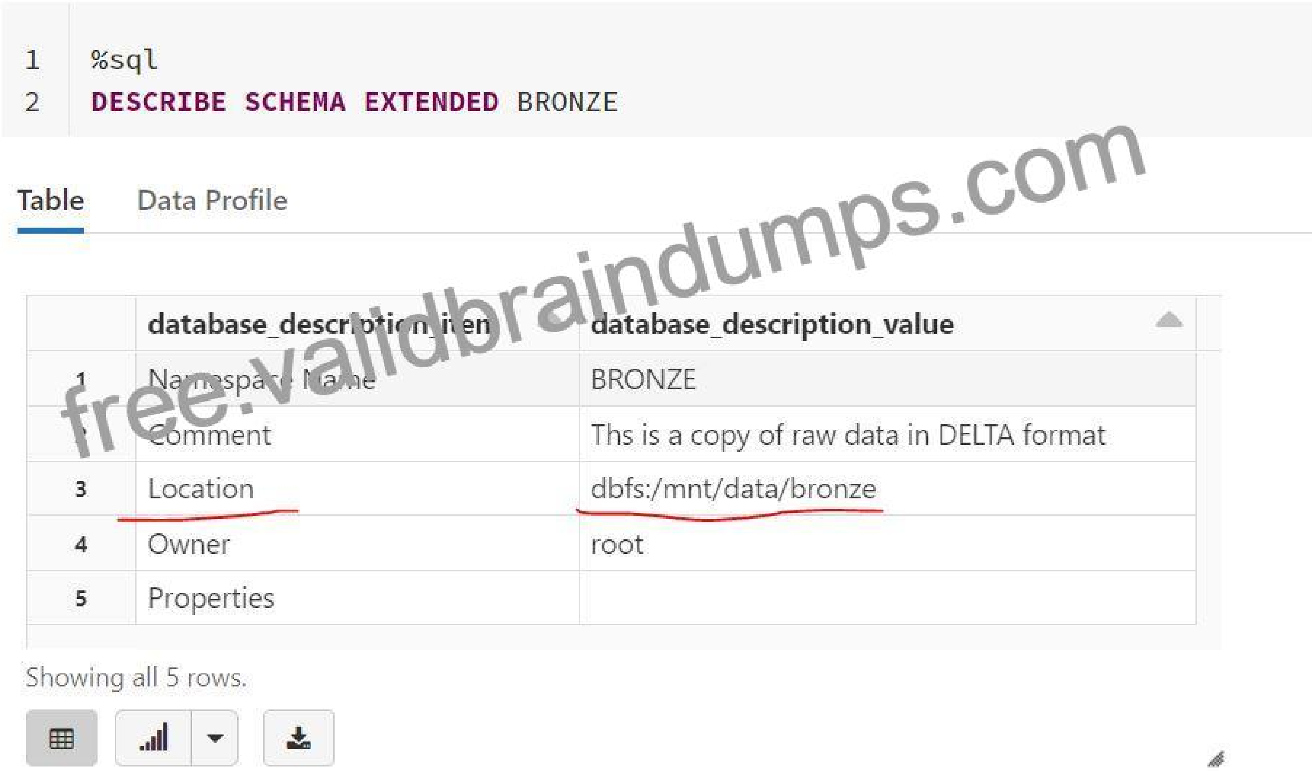
**NEW QUESTION 104**

How do you check the location of an existing schema in Delta Lake?
* Run SQL command SHOW LOCATION schema_name
* Check unity catalog UI
* Use Data explorer
* Run SQL command DESCRIBE SCHEMA EXTENDED schema_name

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

Here is an example of how it looks

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



```
%sql
DESCRIBE SCHEMA EXTENDED BRONZE
```

| | database_description_item | database_description_value |
|---|---|---|
| 1 | Namespace Name | BRONZE |
| 2 | Comment | Ths is a copy of raw data in DELTA format |
| 3 | Location | dbfs:/mnt/data/bronze |
| 4 | Owner | root |
| 5 | Properties | |

Showing all 5 rows.

**NEW QUESTION 105**

A data analyst has provided a data engineering team with the following Spark SQL query:

1.SELECT district,

2.avg(sales)

3.FROM store_sales_20220101

4.GROUP BY district;

The data analyst would like the data engineering team to run this query every day. The date at the end of the

table name (20220101) should automatically be replaced with the current date each time the query is run.

Which of the following approaches could be used by the data engineering team to efficiently auto-mate this

process?
* They could wrap the query using PySpark and use Python&#8217;s string variable system to automatically

update the table name
* They could request that the data analyst rewrites the query to be run less frequently
* They could pass the table into PySpark and develop a robustly tested module on the existing query
* They could replace the string-formatted date in the table with a timestamp-formatted date
* They could manually replace the date within the table name with the current day&#8217;s date

**NEW QUESTION 106**

How does a Delta Lake differ from a traditional data lake?
* Delta lake is Datawarehouse service on top of data lake that can provide reliability, se-curity, and performance
* Delta lake is a caching layer on top of data lake that can provide reliability, security, and performance
* Delta lake is an open storage format like parquet with additional capabilities that can provide reliability, security, and performance
* Delta lake is an open storage format designed to replace flat files with additional capa-bilities that can provide reliability, security, and performance
* Delta lake is proprietary software designed by Databricks that can provide reliability, security, and performance
Explanation

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

* Open source

* Builds up on standard data format

* Optimized for cloud object storage

* Built for scalable metadata handling

Delta lake is not

* Proprietary technology

* Storage format

* Storage medium

* Database service or data warehouse

**NEW QUESTION 107**

What is the probability that the total of two dice will be greater than 8, given that the first die is a 6?
* 1/3
* 2/3
* 1/6
* 2/6

**NEW QUESTION 108**

In order to use Unity catalog features, which of the following steps needs to be taken on man-aged/external tables in the Databricks workspace?
* Enable unity catalog feature in workspace settings
* Migrate/upgrade objects in workspace managed/external tables/view to unity catalog
* Upgrade to DBR version 15.0
* Copy data from workspace to unity catalog
* Upgrade workspace to Unity catalog
Explanation

Upgrade tables and views to Unity Catalog &#8211; Azure Databricks | Microsoft Docs Managed table: Upgrade a managed to Unity Catalog External table: Upgrade an external table to Unity Catalog

**NEW QUESTION 109**

Which of the following statements describes Delta Lake?
* Delta Lake is an open source platform to help manage the complete machine learning lifecycle
* Delta Lake is an open format storage layer that delivers reliability, security, and per-formance
* Delta Lake is an open source data storage format for distributed data
* Delta Lake is an open source analytics engine used for big data workloads
* Delta Lake is an open format storage layer that processes data
Explanation

Delta Lake

**NEW QUESTION 110**

What is the main difference between the silver layer and the gold layer in medalion architecture?
* Silver may contain aggregated data
* Gold may contain aggregated data

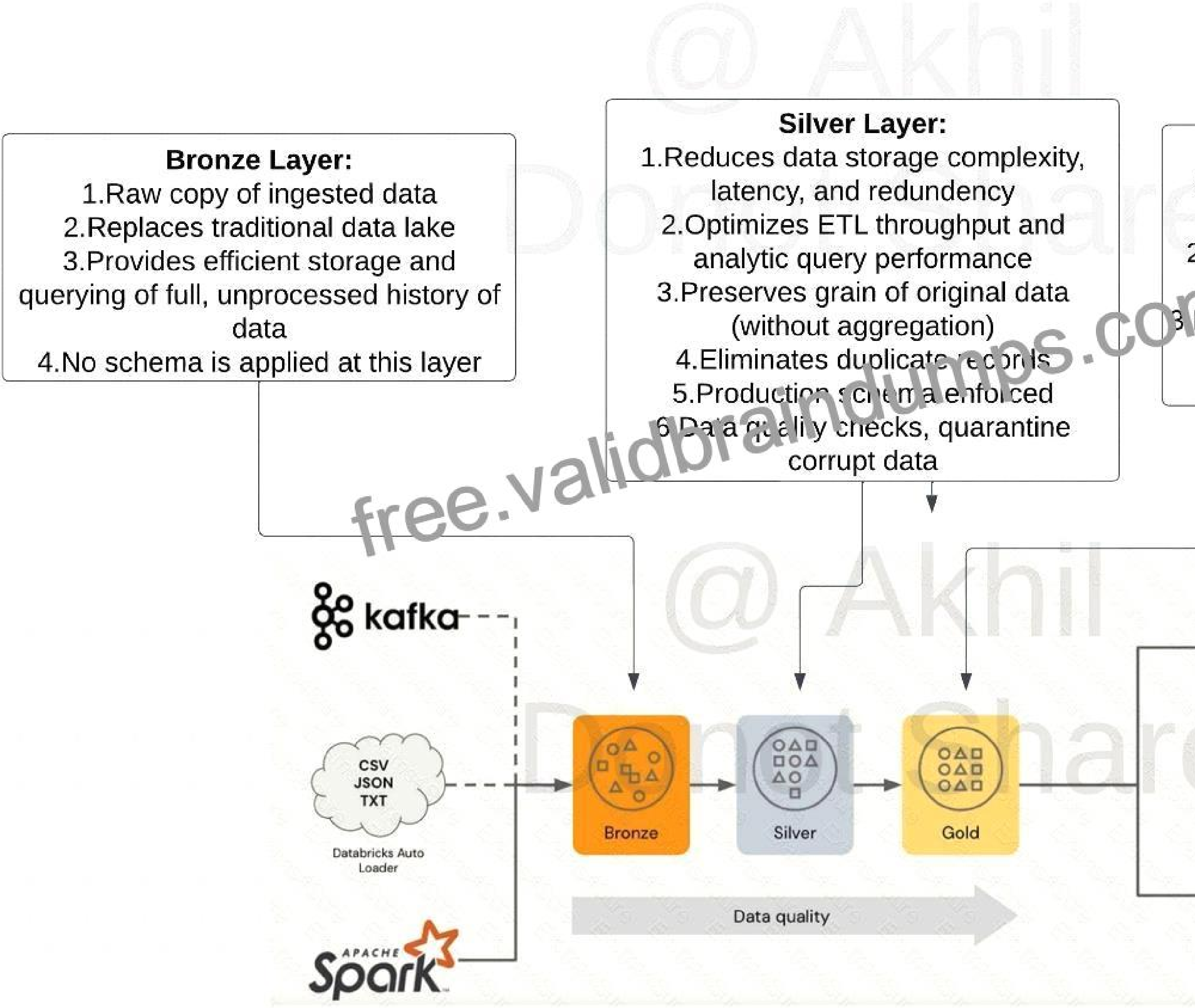
* Data quality checks are applied in gold
* Silver is a copy of bronze data
* God is a copy of silver data
Explanation

Medallion Architecture &#8211; 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



**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

**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

**NEW QUESTION 111**

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

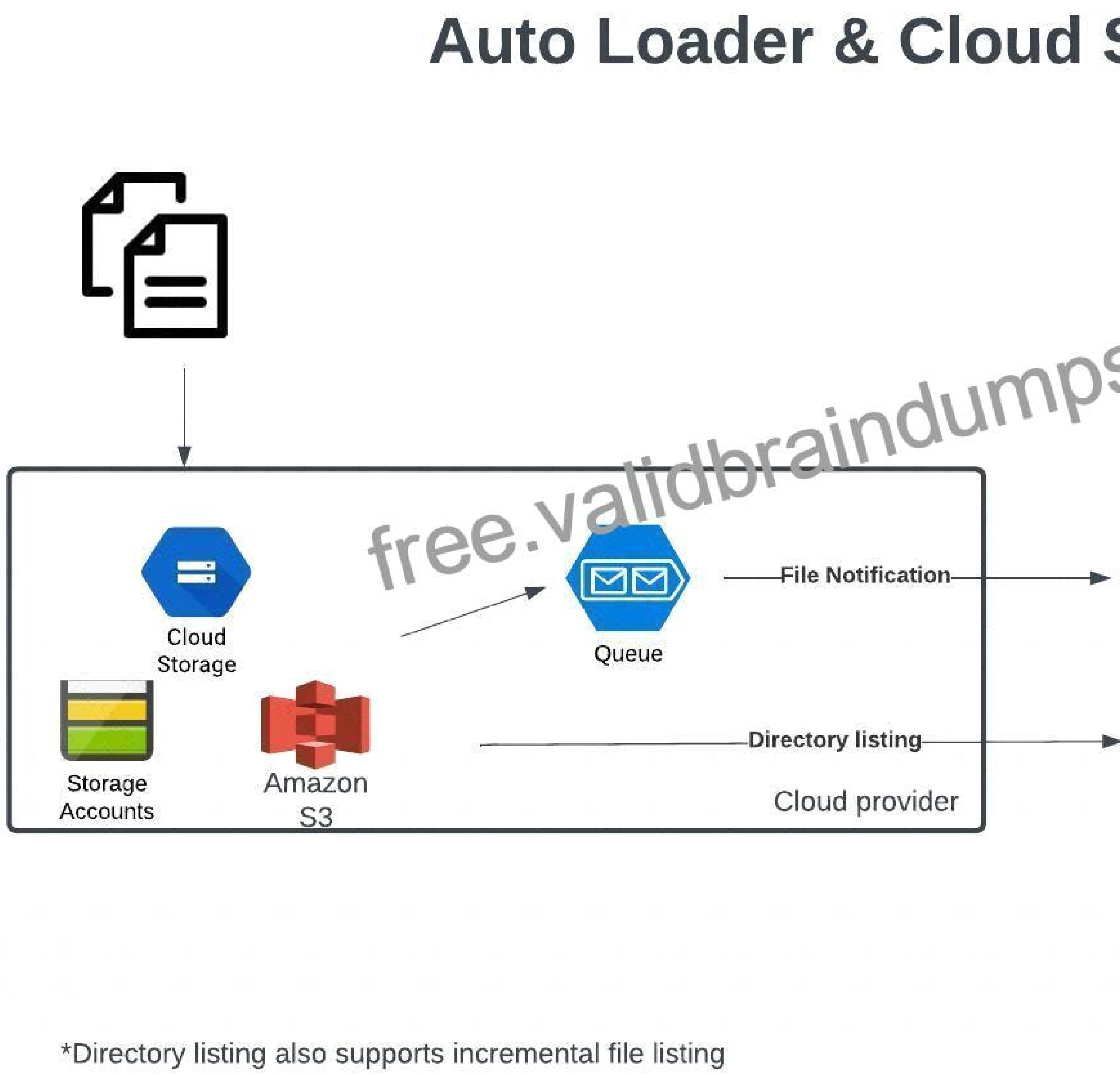
- AUTO LOADER is native to DELTA lake it cannot support external cloud object storage
- AUTO LOADER has to be triggered from an external process when the file arrives in the cloud storage
- AUTO LOADER needs to be converted to a Structured stream process
- AUTO LOADER can only process continuous data when stored in DELTA lake
- AUTO LOADER can support file notification method so it can process data as it arrives

Explanation

Auto Loader supports two modes when ingesting new files from cloud object storage Directory listing: Auto Loader identifies new files by listing the input directory, and uses a directory polling approach.

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

Diagram Description automatically generated



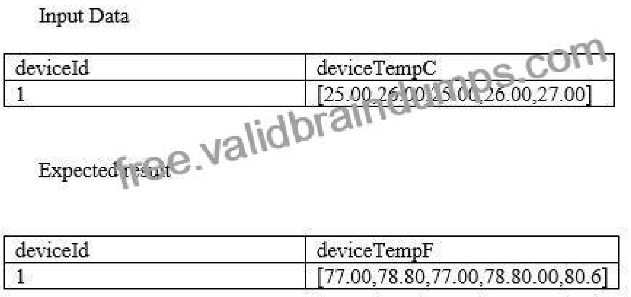
File notification is more efficient and can be used to process the data in real-time as data arrives in cloud object storage.

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

**NEW QUESTION 112**

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

Schema: deviceId INT, deviceTemp ARRAY<double>



Input Data

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

Expected result

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

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

FROM sensors
* APPLY
* MULTIPLY
* ARRAYEXPR
* TRANSFORM
* FORALL

Explanation

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

1.transform(expr, func)

**NEW QUESTION 113**

What is the purpose of the silver layer in a Multi hop architecture?
* Replaces a traditional data lake
* Efficient storage and querying of full, unprocessed history of data
* Eliminates duplicate data, quarantines bad data
* Refined views with aggregated data
* Optimized query performance for business-critical data

Explanation

Medallion Architecture &#8211; 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

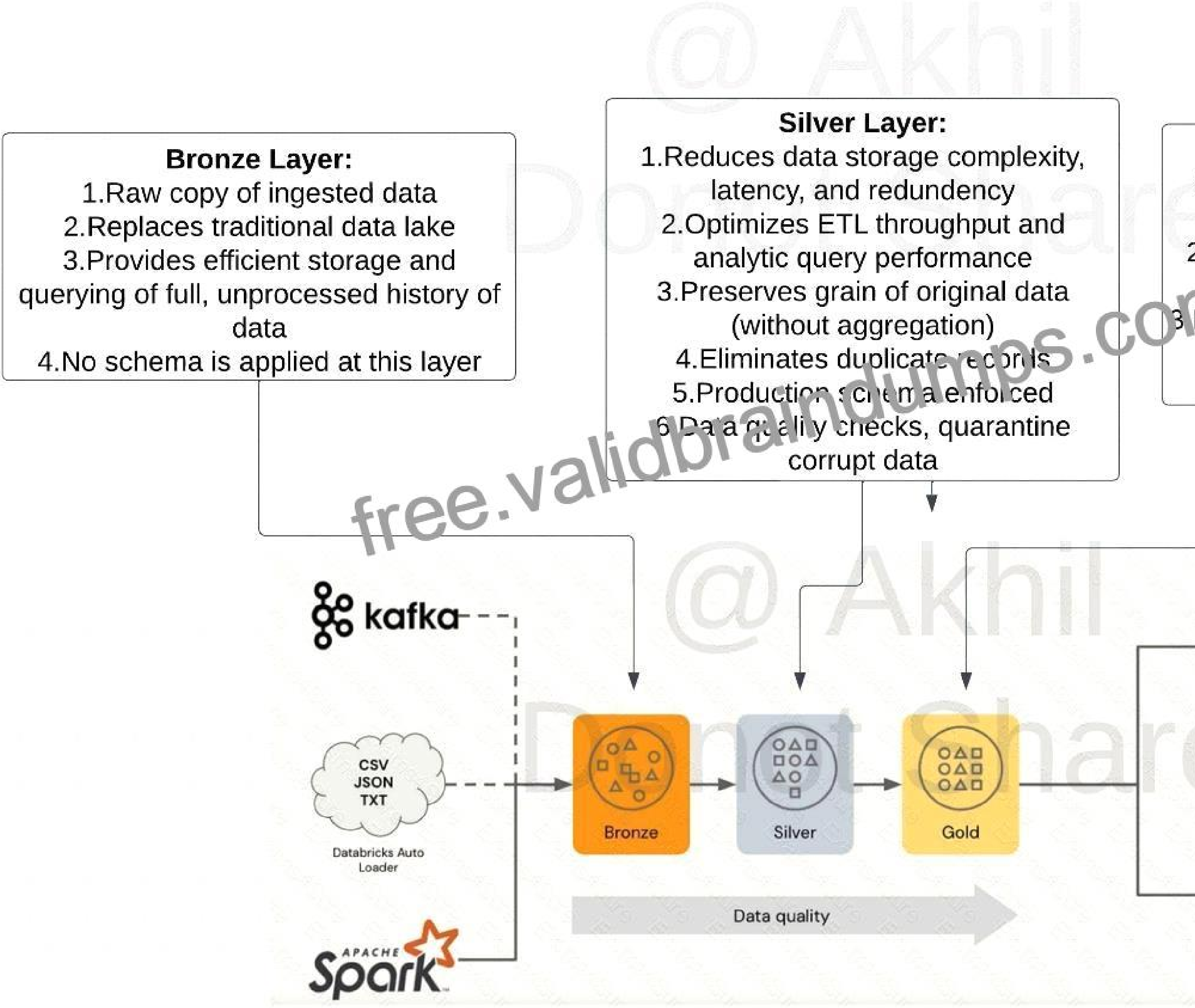
5. production schema enforced

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



**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

**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

**NEW QUESTION 114**

Which of the following approaches can the data engineer use to obtain a version-controllable con-figuration of the Job&#8217;s schedule and configuration?

- They can link the Job to notebooks that are a part of a Databricks Repo.
- They can submit the Job once on a Job cluster.
- They can download the JSON equivalent of the job from the Job&#8217;s page.
- They can submit the Job once on an all-purpose cluster.
- They can download the XML description of the Job from the Job&#8217;s page

**NEW QUESTION 115**

You currently working with the marketing team to setup a dashboard for ad campaign analysis, since the team is not sure how often the dashboard should be refreshed they have decided to do a manual refresh on an as needed basis. Which of the following steps can be taken to reduce the overall cost of the compute when the team is not using the compute?

*Please note that Databricks recently change the name of SQL Endpoint to SQL Warehouses.

- They can turn on the Serverless feature for the SQL endpoint(SQL Warehouse).
- They can decrease the maximum bound of the SQL endpoint(SQL Warehouse) scaling range.
- They can decrease the cluster size of the SQL endpoint(SQL Warehouse).
- They can turn on the Auto Stop feature for the SQL endpoint(SQL Warehouse).
- They can turn on the Serverless feature for the SQL endpoint(SQL Warehouse) and change the Spot Instance Policy from &#8220;Reliability Optimized&#8221; to &#8220;Cost optimized&#8221;

Explanation

The answer is, They can turn on the Auto Stop feature for the SQL endpoint(SQL Warehouse).

Use auto stop to automatically terminate the cluster when you are not using it.

**NEW QUESTION 116**

Which of the below commands can be used to drop a DELTA table?

- DROP DELTA table_name
- DROP TABLE table_name
- DROP TABLE table_name FORMAT DELTA
- DROP table_name

**NEW QUESTION 117**

Suppose there are three events then which formula must always be equal to P(E1|E2,E3)?

- P(E1,E2,E3)P(E1)/P(E2:E3)
- P(E1,E2;E3)/P(E2,E3)
- P(E1,E2|E3)P(E2|E3)P(E3)
- P(E1,E2|E3)P(E3)
- P(E1,E2,E3)P(E2)P(E3)

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 &#8220;the probability of A given B&#8221;

It is commonly denoted by P(A|B):or sometimes PB(A). In case that both &#8220;A&#8221; and &#8220;B&#8221; are categorical variables, conditional probability table is typically used to represent the conditional probability.

**Pass Databricks-Certified-Professional-Data-Engineer exam Updated 220 Questions:** https://www.validbraindumps.com/Databricks-Certified-Professional-Data-Engineer-exam-prep.html]