Below are the some questions which came during the certification exam

===========================================================================

1)A junior data engineer from your team wants to insert 5 records in the **employees** table. They have come up with the following set of SQL queries.

1. INSERT INTO employees VALUES (234, ‘Erich Heard’);
2. INSERT INTO employees VALUES (209, ‘Paul Fosbury’);
3. INSERT INTO employees VALUES (141, ‘Ricky Matt’);
4. INSERT INTO employees VALUES (940, ‘Jeff Sims’);
5. INSERT INTO employees VALUES (744, ‘Chriss Holmes’);

Each of the statements is processed as a separate transaction and you need to modify the statement to be able to insert all 5 records in one go. Which of the following SQL statements can be used to insert these 5 records in a single transaction?

* 
  1. INSERT INTO TABLE employees VALUES
  2. (234, ‘Erich Heard’),
  3. (209, ‘Paul Fosbury’),
  4. (141, ‘Ricky Matt’),
  5. (940, ‘Jeff Sims’),
  6. (744, ‘Chriss Holmes’);
* 
  1. INSERT INTO employees MULTIPLE VALUES
  2. (234, ‘Erich Heard’),
  3. (209, ‘Paul Fosbury’),
  4. (141, ‘Ricky Matt’),
  5. (940, ‘Jeff Sims’),
  6. (744, ‘Chriss Holmes’);
* 
  1. INSERT INTO employees ‘5’ VALUES
  2. (234, ‘Erich Heard’),
  3. (209, ‘Paul Fosbury’),
  4. (141, ‘Ricky Matt’),
  5. (940, ‘Jeff Sims’),
  6. (744, ‘Chriss Holmes’);
* 
  1. INSERT INTO employees VALUES
  2. (234, ‘Erich Heard’),
  3. (209, ‘Paul Fosbury’),
  4. (141, ‘Ricky Matt’),
  5. (940, ‘Jeff Sims’),
  6. (744, ‘Chriss Holmes’);

**(Correct)**

* 
  1. INSERT INTO employees
  2. (234, ‘Erich Heard’),
  3. (209, ‘Paul Fosbury’),
  4. (141, ‘Ricky Matt’),
  5. (940, ‘Jeff Sims’),
  6. (744, ‘Chriss Holmes’);

#### Explanation

The correct code block is:

1. INSERT INTO employees VALUES
2. (234, ‘Erich Heard’),
3. (209, ‘Paul Fosbury’),
4. (141, ‘Ricky Matt’),
5. (940, ‘Jeff Sims’),
6. (744, ‘Chriss Holmes’);

Rest of the options are not syntactically correct. The **INSERT** query works the same way for adding multiple records as it works for adding a single record. You can provide comma separated rows to add them in one go.

More Info: [Multi-row insert into a table](https://docs.databricks.com/sql/language-manual/sql-ref-syntax-dml-insert-into.html#:~:text=INSERT%20INTO%20students%20VALUES%0A%20%20%20%20(%27Bob%20Brown%27%2C%20%27456%20Taylor%20St%2C%20Cupertino%27%2C%20222222)%2C%0A%20%20%20%20(%27Cathy%20Johnson%27%2C%20%27789%20Race%20Ave%2C%20Palo%20Alto%27%2C%20333333)%3B)

Question 2: **Correct**

You have created a new managed table **members** in the **company** database using the following set of SQL statements:

1. CREATE DATABASE IF NOT EXISTS company;
2. USE company;
3. CREATE OR REPLACE TABLE members(id int, name string);

What will be the location of the newly created table?

* 

**dbfs:/user/warehouse/company.db/**

* 

**dbfs:/hive/warehouse/company/**

* 

**dbfs:/user/hive/company/**

* 

**dbfs:/user/hive/warehouse/company.db/**

**(Correct)**

* 

**dbfs:/user/lakehouse/company.db/**

#### Explanation

This is a trick question. First of all, you should know that the database has been created without any **LOCATION** parameter which means that the database will be created at the **default location**. The default location for a database is always **dbfs:/user/hive/warehouse/**

Now, as the name of the database is **company**, a directory named **company.db** will be created in the above location and thus the data for the **members** table will be stored in **company.db** directory. So, correct path for the **members** data storage is **dbfs:/user/hive/warehouse/company.db/**

**SCHEMA and DATABASE can be used interchangeably i.e. CREATE SCHEMA is equivalent to CREATE DATABASE**

If you do not specify the location while creating the database, the database will always be created at the default location. A **trick to remember** the default location of the database is **UHW** i.e. **U**ser **H**ive **W**arehouse.

More Info: [Default location for CREATE DATABASE command](https://docs.databricks.com/spark/latest/spark-sql/language-manual/sql-ref-syntax-ddl-create-schema.html#:~:text=If%20the%20location%20is%20not%20specified%2C%20the%20schema%20is%20created%20in%20the%20default%20warehouse%20directory%2C%20whose%20path%20is%20configured%20by%20the%20static%20configuration%20spark.sql.warehouse.dir)

Question 3: **Correct**

You, as a data engineer, want to use a SQL query in a Python function. What approach can you follow?

* 

**spark.sql() function should be used to run the SQL query**

**(Correct)**

* 

**Change the cell’s language to SQL as the SQL cell allows the usage of Python code as well**

* 

**SQL query cannot be accessed inside Python code**

* 

**pyspark.sql() function should be used to run the SQL query**

* 

**Install Spark SQL driver to run the query**

#### Explanation

If you are using a Databricks notebook, **spark.sql()** is an essential part of your code as it can be used to run SQL queries inside a Python cell. **sql()** is a function on a **SparkSession** object. In Databricks, **spark** is an object of **SparkSession** and that is the reason you can use **spark.sql(query)** to run a SQL query on a view or a table.

More Info: [Using spark.sql() to run SQL queries in Python](https://spark.apache.org/docs/2.2.0/sql-programming-guide.html#:~:text=sqlDF%20%3D%20spark.sql(%22SELECT%20*%20FROM%20people%22))

The UPDATE portion of the MERGE will return a duplicate row error in cases where the ON condition matches more than one row. Typically this happens when your USING clause returns duplicate ids

Does data bricks Vendor Lock-in ?

No,with Databricks you can simply leave your data whenever you want. You can then use Databricks to connect to it and process it for virtually any use case.

Data will not be stored in databricks,data would be in the cloud storage location only.

In which format Databricks store the data –

Answer-parquet:

\_delta\_log Directory

▪ contains a number of Parquet data files

▪ Records in Delta Lake tables are stored as data in Parquet files.

▪ Transactions to Delta Lake tables are recorded

▪ Each transaction results in a new JSON file being written to the Delta Lake transaction log

▪ Rather than overwriting or immediately deleting files containing changed data, Delta Lake uses the transaction log to indicate whether or not files are valid in a current version of the table. display(dbutils.fs.ls(f"{DA.paths.user\_db}/students/\_delta\_log"))

Which of the following users can use the Gold table as a source?

* **A user that needs to feed the raw data to a table**
* **A user that needs to design a dashboard using aggregated data**
* **A user that needs to add a column to the table**
* **The Gold table is the end of multi-hop architecture and is not used by any user**
* **A user that needs to join static data with streaming data to make the data richer**

#### Explanation

**Gold Table** is the **last** layer in multi-hop architecture. It has all the **aggregated** data and can be consumed by the **dashboards** or **reporting tools**.

More Info: [Uses of Gold table in medallion architecture](https://databricks.com/glossary/medallion-architecture#:~:text=The%20Gold%20layer%20is%20for%20reporting)

Question 31: **Correct**

Which of the following SQL statements counts the number of unique rows from the Silver table **routes**?

* **SELECT count(\*) FROM routes;**
* **SELECT count(DISTINCT \*) FROM routes;**
* **SELECT count\_if(\* is DISTINCT) FROM routes;**
* **SELECT count(\*) FROM routes WHERE \* is DISTINCT;**
* **SELECT count(UNIQUE (\*)) FROM routes;**

#### Explanation

Let us look at all the options one by one

**SELECT count(\*) FROM routes**

**INCORRECT**. It will count the total number of rows from the table **routes** and not the unique (distinct) rows.

**SELECT count(DISTINCT \*) FROM routes**

**CORRECT**. This will print the count of distinct rows from the table **routes**. Also note, **DISTINCT (\*)** can also be used instead of **DISTINCT \***

**SELECT count\_if(\* is DISTINCT) FROM routes**

**INCORRECT**. This is not a valid statement and will return an error.

**SELECT count(\*) FROM routes WHERE \* is DISTINCT**

**INCORRECT**. This is also an invalid statement and will return an error.

**SELECT count(UNIQUE (\*)) FROM routes**

**INCORRECT**. **UNIQUE** is not a valid keyword, **DISTINCT** should be used instead.

More Info: [count() function in SQL](https://docs.databricks.com/sql/language-manual/functions/count.html)

A junior data engineer has joined a team and needs to create a DLT pipeline. Which of the following describes the flow of actions to create a new pipeline?

Graphical user interface, text, application

Description automatically generated

How to set job failure notification:

Graphical user interface, text, application

Description automatically generated

Question 41:

Which of the following is optional while creating a Delta Live Table pipeline?

* Target Database(Correct)
* Pipeline name
* Notebook Library
* Minimum and maximum workers
* Pipeline mode

#### Explanation

Every option except the **Target Database** name is mandatory. **Target Database name is optional** and can be skipped while creating a pipeline.

More Info: [Use of target database name in a DLT pipeline](https://databricks.com/discover/pages/getting-started-with-delta-live-tables#:~:text=A%20Target%20is%20optional%20but%20recommended%20since%20the%20target%20is%20the%20target%20database%20where%20other%20authorized%20members%20can%20access%20the%20resulting%20data%20from%20the%20pipeline.)

Please note all the mandatory parameters and optional parameters.

**Notebook libraries is mandatory:**

Graphical user interface, text, application

Description automatically generated

Chart

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generated

Question 39 Operations team is using a centralized data quality monitoring system, a user can publish data quality metrics through a webhook, you were asked to develop a process to send messages using webhook if there is atleast one duplicate record, which of following approaches can be taken to integrate with current data quality monitoring system

* Use notebook and Jobs to use python to publish DQ metrics
* Setup an alert to send an email, use python to parse email, and publish a webhook message
* Setup an alert with custom template Setup an alert with custom Webhook destination
* Setup an alert with dynamic template

Explanation Alerts supports multiple destinations, email is the default destinations.

[Alert destinations | Databricks on AWS](https://docs.databricks.com/sql/admin/alert-destinations.html)

Graphical user interface, application

Description automatically generated

Grant All Privilge Syntax:

[GRANT | Databricks on AWS](https://docs.databricks.com/spark/latest/spark-sql/language-manual/security-grant.html#:~:text=GRANT%20privilege_types%20ON%20securable_object%20TO%20principal)

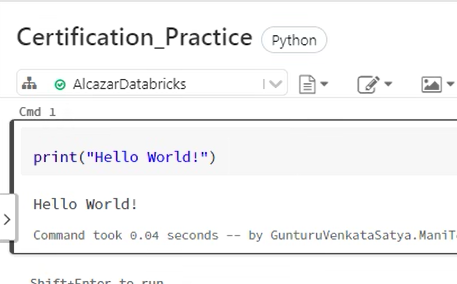
> GRANT CREATE ON SCHEMA <schema-name> TO `alf@melmak.et`;

> GRANT ALL PRIVILEGES ON TABLE forecasts TO finance;

> GRANT SELECT ON TABLE sample\_data TO USERS;

In a notebook, all the cells contain code in Python language and you want to add another cell with a SQL statement in it. You changed the default language of the notebook to accomplish this task. What changes (if any) can be seen in the already existing Python cells?

In this case, the magic command %python will be added at the beginning of all the cells that contain Python code. If you try to change the default language for a notebook, all the cells containing code in previously selected default language will be updated and %previous\_default\_language (like %sql, %scala) will be added as the first line in the cells. This is done to assure that the default language changes will not affect the working of the notebook.



Now added another cell which execute SQL commands and changing the default language to SQL

Graphical user interface, application

Description automatically generated

All of the cells which has python statements will prefix %python

Graphical user interface, text, application

Description automatically generated

Question 3: **Correct**

Which of the following commands can be used to combine small files to a bigger file to achieve better performance?

* 

**VACUUM**

* 

**DELETE**

* 

**OPTIMIZE**

**(Correct)**

* 

**COMBINE**

* 

**RESTORE**

#### Explanation

Some of the above commands can be asked in the actual exam. So, it is preferred that you know each one of them. Let us go through all the options and learn more about their usage to help you train for the actual exam.

**VACUUM** – **Cleans** up files that are older than X **number of hours**.

**DELETE**– Used to delete data from a table

**OPTIMIZE** – Used for **optimizing** the files by combining **small files to a bigger file**.

**COMBINE** – Invalid command

**RESTORE** – Used for restoring a table to a **previous** **version** or **timestamp**

Looking at all the options and their usage you now know why **OPTIMIZE** is the correct answer.

More Info: [VACUUM](https://docs.databricks.com/sql/language-manual/delta-vacuum.html) | [DELETE FROM](https://docs.databricks.com/sql/language-manual/delta-delete-from.html) | [OPTIMIZE](https://docs.databricks.com/sql/language-manual/delta-optimize.html) | [RESTORE](https://docs.databricks.com/sql/language-manual/delta-restore.html)

Which of the following commands can be used to rotate a table on one of its axes?

• ROTATE

**• PIVOT (Correct)**

• FILTER

• EXIST

• REDUCE

Note: In exam question mention as long table to wide table

Explanation PIVOT command helps in viewing different data perspectives of a table including slicing, dicing and rotating. Using PIVOT you can easily convert **a long table(more rows) to a wide table(more columns).**

Question 21 A data analyst needs to count the number of NULL values in column secondary\_mobile from the data present in the personal\_details table. Which of the following SQL statements, when executed, will fetch the required result?

• SELECT count(NULL secondary\_mobile) FROM personal\_details

• SELECT count\_if(secondary\_mobile NULL) FROM personal\_details

**• SELECT count\_if(secondary\_mobile IS NULL) FROM personal\_details (Correct)**

• SELECT count\_null(secondary\_mobile) FROM personal\_details

• SELECT count\_if(NULL secondary\_mobile) FROM personal\_details

Explanation To count the number of NULL values in a column, various methods can be adopted but the one which is used the most is count\_if() function. count\_if() function accepts conditions like email IS NOT NULL or length(phone\_number) != 10 etc.

FILTER is a higher-order function which can filter the contents of an array column using a lambda function. FILTER function takes two arguments, the array column name and the condition.

Here, details is the name of the column and i -> i.salary > 10000 is the condition. If the condition is TRUE, the value is added to the column filtered\_workers. Also note, i is the variable which iterates over the array which means i.salary is the salary of each worker in the array.

FILTER (details, i -> i.salary > 10000) AS filtered\_workers

Question related to syntax of delta live table:Please see DLT syntax:

[Process streaming data with Delta Live Tables - Azure Databricks | Microsoft Docs](https://docs.microsoft.com/en-us/azure/databricks/workflows/delta-live-tables/delta-live-tables-incremental-data)

CREATE OR REFRESH STREAMING LIVE TABLE customers\_silver

AS SELECT \* FROM STREAM(LIVE.customers\_bronze)

CREATE LIVE TABLE sales\_order\_in\_chicago AS

SELECT order\_date, city, sum(price) as sales,

FROM LIVE.sales\_orders\_cleaned

WHERE city = 'Chicago') GROUP BY order\_date, city

Data quality constraints in Delta Live Tables

CONSTARINT constraint\_name EXPECT (condition) ON VILOATION FAIL UPDATE – As soon as the condition is not met, the pipeline will fail

Cron scheduler supports time zone selection

For scheduling a job, starting time, time interval and time zone can be selected.

More Info: Cron job scheduling in Databricks

Graphical user interface, text, application

Description automatically generated

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

A. UPDATE

B. COPY

C. INSERT INTO

D. DELETE

E. UNION

A data engineer has configured a Structured Streaming job to read from a table, manipulate

the data, and then perform a streaming write into a new table.

The code block used by the data engineer is below:

(spark.table("sales")

.withColumn("avg\_price", col("sales") / col("units"))

.writeStream

.option("checkpointLocation", checkpointPath)

.outputMode("complete")

.\_\_\_\_\_

.table("new\_sales")

)

If the data engineer only wants the query to execute a single micro-batch to process all of

the available data, which of the following lines of code should the data engineer use to fill in

the blank?

A. trigger(once=True)

B. trigger(continuous="once")

C. processingTime("once")

D. trigger(processingTime="once")

E. processingTime(1)

Question 36

A Delta Live Table pipeline includes two datasets defined using STREAMING LIVE TABLE.

Three datasets are defined against Delta Lake table sources using LIVE TABLE.

The table is configured to run in Development mode using the Triggered Pipeline Mode.

Assuming previously unprocessed data exists and all definitions are valid, 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

testing.

D. All datasets will be updated once and the pipeline will shut down. The compute

resources will persist to allow for additional testing.

E. All datasets will be updated continuously and the pipeline will not shut down. The

compute resources will persist with the pipeline

Question 40

A data analyst has noticed that their Databricks SQL queries are running too slowly. They

claim that this issue is affecting all of their sequentially run queries. They ask the data

engineering team for help. The data engineering team notices that each of the queries uses

the same SQL endpoint, but the SQL endpoint is not used by any other user.

Which of the following approaches can the data engineering team use to improve the

latency of the data analyst’s queries?

A. They can turn on the Serverless feature for the SQL endpoint.

B. They can increase the maximum bound of the SQL endpoint’s scaling range.

C. They can increase the cluster size of the SQL endpoint. --Answer

D. They can turn on the Auto Stop feature for the SQL endpoint.

E. They can turn on the Serverless feature for the SQL endpoint and change the Spot

Instance Policy to “Reliability Optimized.”

Single node cluster: for smaller datasets/workloads

Single-node clusters are a cost-efficient option for single machine workloads. Single-node clusters support Spark and Spark data sources including Delta, as well as libraries including scikit-learn and tensorflow included in the Runtime for Machine Learning.

Graphical user interface, application

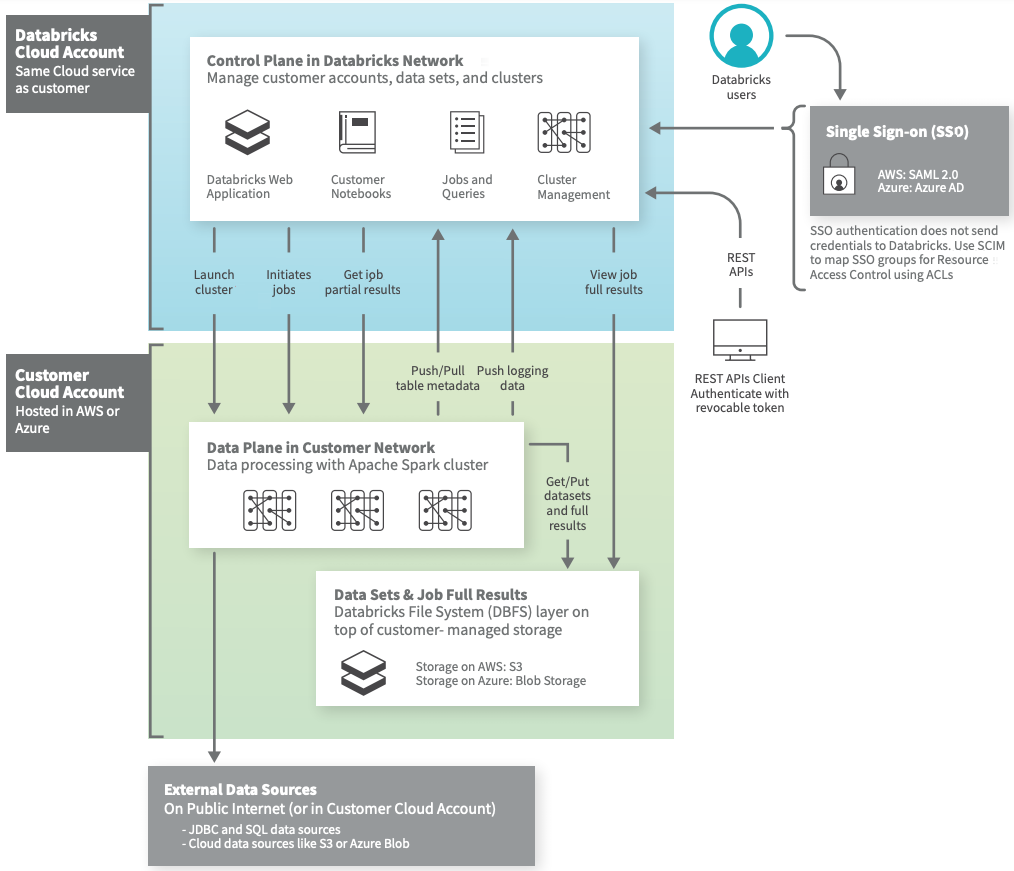
Description automatically generated

Advantage of Data Lakehouse

Which of the following are stored in the control pane of Databricks Architecture?

Databricks Web application

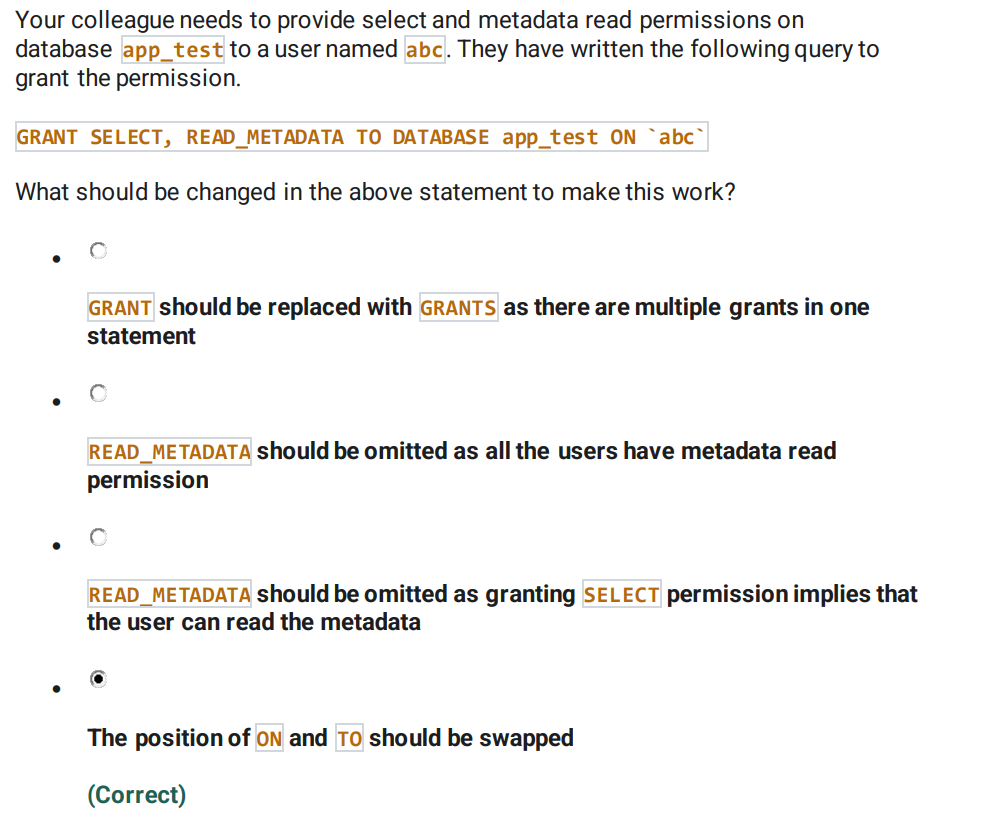
Refer below diagram:



While querying the previous version of the table using time travel you realized you are no longer able to query the historical data?

* You currently do not have access to view historical data
* By default, historical data is cleaned every 180 days in DELTA
* A command VACUUM table\_name RETAIN 0 was ran
* Time travel is disabled
* Time travel must be enabled before you query previous data

Explanation The answer is, VACUUM table\_name RETAIN 0 was ran Recursively vacuum directories associated with the Delta table and remove data files that are no longer in the latest state of the transaction log for the table and are older than a retention threshold. 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



GRANT SELECT, READ\_METADATA ON DATABASE app\_test TO `abc` The query seems to be correct but the positions of TO and ON should be swapped. ON DATABASE comes first with the database name which is followed by TO while specifying the name of the user or group to which the grant needs to be issued. Also note, GRANTS is also a valid keyword but is used to view the grants on a specific data object like database, table etc.

[GRANT | Databricks on AWS](https://docs.databricks.com/spark/latest/spark-sql/language-manual/security-grant.html#:~:text=GRANT%20privilege_types%20ON%20securable_object%20TO%20principal)

Drop table customers ( Metadata deleted but data is available - what type of table?)

External

You are working on a process to load external CSV files into a Delta by leveraging the COPY INTO command, but after running the command for the second time no data was loaded into the table name, why is that?

COPY INTO table\_name FROM 'dbfs:/mnt/raw/\*.csv' FILEFORMAT = CSV

COPY INTO did not detect new files after the last load

The answer is COPY INTO did not detect new files after the last load, COPY INTO keeps track of files that were successfully loaded into the table, the next time when the COPY INTO runs it skips them. FYI, you can change this behavior by using COPY\_OPTIONS 'force'= 'true', when this option is enabled all files in the path/pattern are loaded. COPY INTO table\_identifier FROM [ file\_location | (SELECT identifier\_list FROM file\_location) ] FILEFORMAT = data\_source [FILES = [file\_name, ... | PATTERN = 'regex\_pattern'] [FORMAT\_OPTIONS ('data\_source\_reader\_option' = 'value', ...)] [COPY\_OPTIONS 'force' = ('false'|'true')]

Structured streaming uses to create an end-to-end fault tolerance?

Structured Streaming ensures end-to-end exactly-once fault-tolerance guarantees through checkpointing and Write Ahead Logs

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

How does Auto loader achieve incremental loading?

Auto Loader is configured to incrementally process files from a directory in cloud object storage into a Delta Lake table. Auto Loader with automatic schema inference and evolution, the 4 argument

Graphical user interface, text, application, email

Description automatically generated

spark.readStream .format("cloudFiles") .option("cloudFiles.format", source\_format) .option("cloudFiles.schemaLocation", checkpoint\_directory) .load(data\_source) .writeStream .option("checkpointLocation", checkpoint\_directory) .option("mergeSchema", "true") .table(table\_name))

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

* Data Sharing
* Unity Catalog
* DELTA lake
* Use single storage location
* DELTA LIVE Pipeline

The answer is Unity catalog

Difference between raw and bronze data

Bronze tables contain raw data ingested from various sources (JSON files, RDBMS data, IoT data, to name a few examples).

• Silver tables provide a more refined view of our data. We can join fields from various bronze tables to enrich streaming records, or update account statuses based on recent activity.

• Gold tables provide business-level aggregates often used for reporting and dashboarding. This would include aggregations such as daily active website users, weekly sales per store, or gross revenue per quarter by the department.

You are currently asked to work on building a data pipeline, you have noticed that you are currently working with data

source with 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? Ans :DELTA LIVE TABLES

Difference between CREATE LIVE STREAMING TABLE vs CREATE LIVE STREAMING

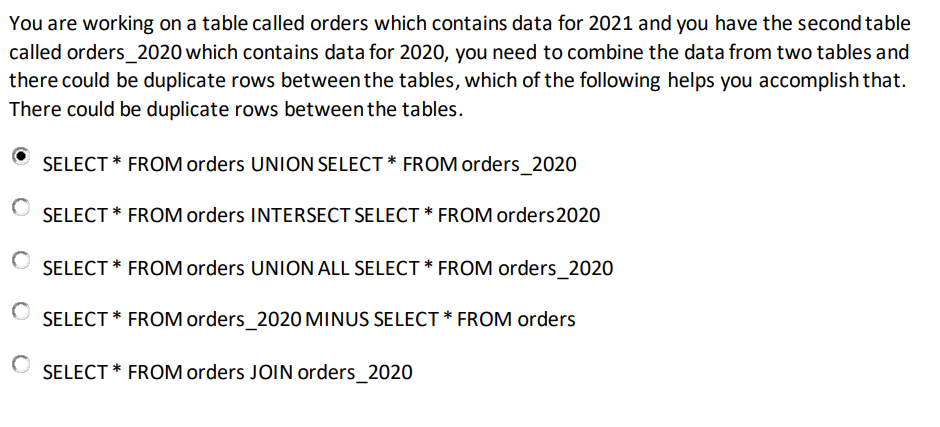
Graphical user interface, text, application, email

Description automatically generated

Explanation The answer is, They can increase the maximum bound of the SQL endpoint’s scaling range. SQL endpoint scales horizontally(scale-out) and vertical (scale-up), you have to understand when to use what. Scale-out -> to add more clusters for a SQL endpoint, change max number of clusters If you are trying to improve the throughput, being able to run as many queries as possible then having an additional cluster(s) will improve the performance. Scale-up-> Increase the size of the SQL endpoint, change cluster size from x-small to small, to medium, X Large.... If you are trying to improve the performance of a single query having additional memory, additional nodes and cpu in the cluster will improve the performance

18) GRANT USAGE on Database

19. Grant on Delta Live tables (User left the organization and permission on delta tables to be given to new Lead) – Ans - Administrator



Explanation Answer is SELECT \* FROM orders UNION SELECT \* FROM orders\_2020 UNION and UNION ALL are set operators, UNIONcombines the output from both queries but also eliminates the duplicates. UNION ALL combines the output from both queries.

Graphical user interface, text, application

Description automatically generated

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=DatabricksAcadem>

23. Insert data avoiding duplicating data (MERGE Command)

24. Database Location ( DESCRIBE DETAILS DATABASE)

Define an external SQL table by connecting to a local instance of an SQLite database using JDBC (org.apach.spark.sql.jdbc) Graphical user interface, text, application, email

Description automatically generated

Table Properties:

[Table properties and table options (Databricks SQL) | Databricks on AWS](https://docs.databricks.com/sql/language-manual/sql-ref-syntax-ddl-tblproperties.html#common-tblproperties-and-options-keys)

29. If column == "value" and process;

Questions from Abhishek

1. Question on Grant Usage privilege. New team has joined the group and want to see all the tables in a given database, what privileges are required?
2. Question on Grant ALL privileges. New team has joined the group and want complete access on the given database, what privileges are required?
3. Data engineer having access on a DB leaves organization. New engineer joins and wants the access. Assuming original owner does not have any more access to the DB what are the options?
   1. Workspace admin can grant access
   2. New owner can grant access
   3. Original owner can grant access
   4. Access transfer not possible
   5. Databricks account representative can grant access
4. Question on create table using jdbc syntax
5. Question on proceesingTime option, want to trigger at every few minutes.
6. There is a table with PII information. Audit team wants property to notify the type of data in the table. What are the options. Ans : Comment
7. Question on restarting mechanism for Autoloader. Ans : Checkpointing and Write-ahead log
8. Delta features for incremental process. Ans. Checkpointing
9. How is Array functions are valuable? Ans: To provide easy processing of JSON data
10. Taks exists, engineer creates a new preprocessing task and want to set its dependency on original task. What are the options? Ans : Create new task and add it as a dependency of original task
11. Notebooks current run is taking long time, how we can debug. Ans : Got to Jobs and check notebook under active runs
12. Need to send alert notification on a webhook, what is the way? Ans. : attache alert to webhook
13. Organization want to use SQL endpoint and want to cut down the cost. How to achieve it? Ans : SQL endpoint Auto Stop Feature
14. Dashboard performance is degrading due to multiple users accessing the same. How to improve the performance? Ans: Increase max range for the SQL endpoint
15. Single folder receiving files that are used by multiple downstream process. What is the option to process the file once incrementally? Ans: Autoloader
16. Statement to delete records with age greater than 25.
17. What resides in Databricks control plane. Ans : Databricks Workspace
18. When to use STREAMING LIVE TABLE? Ans : To incrementally load data
19. Engineer needs to debug each stage of the pipeline. Which tools to use. Ans : Delta Live table
20. Engineer wants to check the quality of each table produced in Delta Live table. How do you do it? Ans : Click on table and check Quality statistics.
21. Question on CHECK ON VIOLATION DROP ROW. Requirement is to drop bad records.
22. What is the behaviour of a pipeline when it is in production mode and continuous. Ans : Process data until pipeline is shut down. Clusters are deleted when pipeline is shut down.
23. Performance issue due to tie it take to start the cluster. What is the option. Ans : Use cluster pools.
24. Performance issue while running query, need to run immediately. What is the option. Ans : Use cluster pools.
25. Create table syntax to combine data of two identical tables without duplicates. Ans : UNION
26. Table is dropped, SHOW table does not show it any more but the data is still available. Reason? Ans : It was an external/unmanaged table.
27. Why data like house is preferred for data quality assurance? Ans : Due to ACID compliant characteristics
28. Option to load non-duplicate records. Ans . : MERGE statement
29. Engineer loads previous day data using COPY INTO command. He runs the command but count does not change. Reason ? Ans. : Previous day data is already loaded.
30. Difference between Bronze and Silver layer
31. Difference between Silver and Gold layer
32. Statement for Silver to Gold hop. Ans : Look for aggr function
33. What is the option to stop the query alerts after a week. Ans : Select the duration while setting up the query refresh.
34. Syntax for Python if statement with and condition
35. Few questions on advantages of using Lakehouse.

Questions from Golimi Suresh Babu

**\*\* DE Exam Draft**  
VACCUM Command use/syntax

VERSION Command usage/syntax  
Optimize Command use  
Job Cluster Vs All-Purpose Cluster  
Managed table Vs External table  
Multi-hop Architecture (Properties of 3 diff tables)  
MERGE INTO Vs COPY INTO Vs INSERT INTO  
View Vs TempView Vs GlobalTempView  
Deep clone Vs Shallow Clone  
Syntax for creating,updating,deleting and droping tables/databases  
Auto Loader - use/syntax  
Delta Live TABLES - use/syntax  
CONSTRAINT-EXPECT Commands  
spark.sql() query usage  
Grant/Revoke access on database/table/view - syntax

**Questions from Vijay Air**

1)How does Delta lake differ from traditional data lake?

2) When we execute the command DROP TABLE my\_table then all the metadata & data got deleted from filesystem what is the table type?

3)Which of below SQL function we can use the transform table from long to wide - PIVOT

4) What is difference between Bronze & Silver layer in multihop arechitecture.

5) which of following use case we can use Gold table as source.

6) which command can be used to read delta table in SQL (in pyspark and SQL mode)

7) Schema enforcement in Auto loader

8) How to optimize the perfromance of SQL Endpoint

9) Steps the crerate Delta live table

10) Use of Unity catalog - Account level

11) Multihop architecure details ( Bronze/Silver/Gold)

**Questions from Aniket Patki on 12/09**

1. What data will be in silver layer
2. What data will in gold layer
3. Why you will you Databricks repo over Git
4. Question on syntax of structure streaming
5. Questions on Trigger (once= true)
6. DLT constraints
7. Grant permission on All privilege , modify(2 Qns)
8. Question on union all.
9. Create table syntax
10. Advantage of delta lake over data lake
11. External table – table dropped and data exists
12. To avoid duplicate inserts – use Merge
13. Where to Check status of running Notebook.
14. Dashboard is taking time to start- cluster pool
15. Number od users trying to use same end point query slowed down- increase scale
16. Single query dashboard taking 30 min to reload- Increase cluster size.
17. Cant use time travel- Vacuum was used.
18. Operation happens in Git- Merge
19. Incremental data ingest - Use auto loader

**Questions from Rajesh K:**

1. Which of the following are stored in the control panel data bricks architecture under the database cloud account?
2. Job clusters
3. All-purpose clusters
4. Databricks web application
5. Databricks filesystem
6. Delta Tables

2. Developer options in CI/CD which can only be implemented through GIT provider when using Databricks Repos.

1. Commit & Push
2. Pull
3. Create and edit
4. new branch
5. Trigger DB repos

3. Which SQL command can be used in Databricks to implement incremental data ingestion?

1. Append into Delta table\_name
2. Append into Table\_name
3. Copy into table\_name
4. Insert into Delta table
5. Update table\_name

My other questions were mostly from Python or Pyspark.

1. We should know how to run/call SQL query from PYSPARK with command and Syntax. I faced these questions.
2. Usage of DBUTILS commands.
3. How to create a delta table with constraints, exact syntax.
4. Usage of STREAM command, with exact syntax.

**Questions from Santhi :**

1.sql fuction used to transform table from long to wide

2.which scenario use gold table as source

3.select the statement that use hop form RAW to bronze

4.autoloader option for trigger

5.Insert into query example

6.datalakehouse best describes---simplified governance

7.optimize command

8.View related question

9.what is the default format of databricks data---parquet

10.why parquet file prefer merge rather than create when creating managed table

11.where we can create alert to send mails

12.where we can see the owner name ---permissions tab

13.grant all permissions to table query

14.what is the behaviour of pipeline when it is production and continuos mode

15. change the code block from python to sql----%sql keyword at start of cell

16.query to use python variable in sql ---spark.sql

17.sql end points question

18.CTAS statement query

19. when drop the table data and metadata deleted---managed table

20. to sync the data from git central repository—git pull

21.what is there in customer cluster account

a. notebooks

b. web application

c. repos

d.data

e. cluter management

22.left join query---need to identify the result of the query when given two inputs

23.what is the advantage of parquet format

24.filter higher order function example

25.count\_if query to find null values

26. question on streming query syntax

27. load the data from CSV syntax

28. delta live table query example question (why we used live.tablename in join condition)

29. why we declare dependent job in scduling(need of that -in which scenario will do that)

30.question on check constraint with on violation

31. job creation mandatory option( ans: need to select atleast one notebook library)

Questions from Varun GoWda:

1. New team has joined the team and want to see the team tables in a given database. What privileges are required.

**Answer : GRANT** **SELECT** **ON** **TABLE** <**table**-name> **TO** users;

1. New team has joined the team and want complete access on the database. What privileges are required.

**Answer : GRANT** **ALL** **PRIVILEGES** **ON** **TABLE** forecasts **TO** finance;

1. Data Engineer having access to database leaves the organization, A new team member has joined the team. He is the team leader for the team. Assuming original owner doesn’t have access to the database, who can provide access to the new team member.?

Answer : Workspace admin can grant access

1. Create table using JDBC Syntax

Answer : Create table <table\_name>

Using org.apache.sqpark.sql.jdbc

1. Want to trigger every 5 minutes

Answer: .processingTime(“5 Minutes”)

1. There is a table with PII Information. Audit team wants property to notify the type of data in the table.. what are the options.

CREATE OR REPLACE TABLE users\_pii

COMMENT "Contains PII"

LOCATION "${da.paths.working\_dir}/tmp/users\_pii"

1. Restarting mechanism for autoloader – Checkpointing and Write Ahead log
2. Question on autoloader using for incremental data load – checkpointing
3. How array functions are useful – JSON Data
4. An engineer wants to create a new processing task and wants to set its dependency on the original task. How can he achieve it.? – Create a new task and add it as a dependency of original task.
5. The current notebook taking log time to run. How do you debug in the UI.? – Go to Jobs, in the runs tab check for active runs
6. Need to send alert notification on a webhook. What option.? – Attach alert to the new webhook
7. Update an existing table where age > 25 records needs to be removed—Delete from <table> where age>25;
8. What resides in Databricks Control Plane? – Databricks Web application
9. When to use STREAMING LIVE TABLE vs LIVE Table - ? To incrementally load the data
10. An engineer need to debug each line of the pipeline. Which tools to use.? – Delta Live Table
11. Insert data avoiding duplicate data– Use Merge
12. Table dropped, metadata deleted but still the data exists – External Table
13. Which of the following operations happens only in GIT over Databricks Repo – Merge
14. Why Time travel was not possible to retrieve the previous version – Vacuum with 0 Retention was ran
15. Use of unity Catalog – Account Level
16. In which case we can use gold table as source
17. What kind of data present at Bronze table.?
18. While using pyspark, which of the following command is required to query a table – spark.table(“”)
19. You have to access a notebook data outside of the session and also that should be having physical location as well. – Use Table
20. Python If syntax with and condition – if (name == “Databricks” and present):
21. Query statement for Silver to Gold Hop – Look for aggregate function In the query
22. An Engineer loads previous day data using COPY INTO command. He runs the command again tomorrow but there is no change in the data. What could be the reason.? – Previous day data is already loaded.
23. Why delta lake house preferred over traditional data warehouse for data quality assurance.? – Due to ACID compliant characteristics.
24. Create table syntax of 2 identical tables with no duplicate records. – Use Union
25. Performance issue while running a query and need to run immediately and also taking time at the start of the time – Use cluster pools
26. Engineer wants to check quality of each table produced in DLT – Click on table and Check quality statistics.
27. Question in CHECK ON VIOLATION DROP ROW
28. Behaviour of pipeline when its in production and continuous– all datasets will be updated at set intervals until the pipeline is shutdown. The compute resources will be deployed for the update and terminated when the pipeline is stopped.

Questions from Sayesha:

1.Sql Endpoint question - if many users are one sql endpoint at a time for small queries then how to improve its performance. Ans - Scale out  
2.to grant privilege to new user for databse to just view the tables present in it Ans Grant usage ....  
3.Any question on data quality related Ans - Delta Live tables  
4.what outcome of - CONSTRAINT valid\_count EXPECT (count > 0) ON VIOLATION FAIL UPDATE  
5.to set table property while creation Ans - CREATE TABLE inventory (id INT, units FLOAT) TBLPROPERTIES (business\_owner = ‘supply chain’)  
6.If you have to triger once in single micro batch  
spark.readStream  
.format("delta")  
.table("events\_log")  
.groupBy("customerId")  
.count()  
.writeStream  
.format("delta")  
.outputMode("complete")  
.option("checkpointLocation", "/tmp/delta/eventsByCustomer/\_checkpoints/")  
.trigger(\_\_\_\_\_)  
.table("target\_table")

Ans - .trigger(once = True)

7.To select from option to move data from bronze to silver  
8.To select from option to move data from silver to gold  
9.Change the ownership of a database to another user Ans use alter command.  
10.Question on pipeline in production mode trigger = continous , how it will work ?

1. which of the following is hop from raw data to broze table

ANS. option should not have any aggregare function, Also it should not have any withColumn statement, no dattype casting etc.

2. which keyword is used to convert table from long table to wide table

ANS pivot

3.Count null values in the column

SELECT count\_if(secondary\_mobile is null) from personal details

4.correct syntax of filter (Below 3 are correct syntaxes

df.filter('state === "OH").show(false)

df.filter($state === "OH").show(false)

df.filter(col("state") === "OH").show(false)

5.which of the following sql keyword can be used to append new rows to sql table

ANS. INSERT INTO

7. Data engineer wants the query to execute in multiple micro batches to process all the available data, Which of the following lines of code should the data engineer use to fill in the blanks

ANS- 1 trigger(once=true)

2. trigger(continious="Once")

3. processingTime("once")

4. trigger( ProcessingTime="ONCE")

5. ProcessingTime(1)

I dont know correct answer

8. Delta lake is open source what benefit it offers to customers

ANS. No vendor lock in

9. When database is created without providing location where it is stored

AND UHW \user\hive\warehouse

10. you were asked to develop a process to send messages using webhook

ANS setup an webhook alert

Delta Live Table : constraint ON VIOLATION FAIL what ouptut you will see : Pipeline failure  
Data Explorer: In order to Check owner of table which option you would choose  
Privilege : Read access on table required which privilege to be assigned Answer: Select  
Git: Your associate has pushed code in main branch you need to updated your repo with that change name action to perfrom this : Pull  
Incremental Multi Hop:  
Data is moving from Bronze to Silver Identify valid query for the scenario : Query include calulation i.e col1\* col2 not aggregation  
You want to create dashboard which type of data you will use : Gold Layer with aggregated data.  
Triggers: You want trigger to execute when file arrive which option will be chosen : .trigger(available=True)  
Cleaning Data  
You want data from long format to wide format chose opeartion : Pivot  
Higher Order Function: You have column in table if type array select employee who has worked more than 5 year select appropriate query  
query type question correct answer : filter(column\_name,k->k.column\_name>5)

Creating Delta Tables:  
You want column to be used as generated column select appropriate query: GENEARATE ALWAYS AS keyword to be selected

ETL:  
In order to load CSV data chose appropriate datsource type  
CREATE TABLE  
Using \_\_\_\_\_  
Options(Header=True)  
Location='path/to/file'

You want data from source without loading for single session which type of operation to use Temporary View  
You want table to be optimize table Optimize query format

1. ======================================================================================
2. Two questions on Databricks SQL warehouse/endpoint – Serverless feature. What could be happening for slow query?
3. Question on Data plane/Control plane, what comes under that?
4. DDL for CREATE OR REPLACE Delta table
5. SQL Keyword to append new rows – INSERT INTO
6. OPTIMIZE for combining smaller files
7. Usage of TEMP VIEW scenario based. Different scenario comes for different relational object usage.
8. Manage vs External table
9. Multi hop architecture, which scenario fits for gold table as source?
10. Autoloader creating all string type schema for all the fields for JSON file why?
11. High Order Function – SQL query for FILTER logic
12. Using sql in python notebook - %sql
13. Data Engineer want to write Analyst provided sql query in the same python notebook for testing - spark.sql
14. Using python variable in spark.sql – string interpolation in python F-string
15. Structured Streaming Trigger options, one time multiple batch, one time single batch different scenarios
16. Mandatory fields in creating DLT pipeline
17. Task dependency while creating jobs, What will be the behaviour on dependency. Like dependent will only run when previous will be successful
18. DLT Constraint syntax and working of the statement, ON VIOLATION FAIL UPDATE
19. Behaviour of DLT pipeline Development Continues Pipeline mode
20. Setting Alert to configured Job
21. Setting permissions to the table using GRANT, complete syntax required
22. Query to count nulls for a field - count\_if
23. Setting webhook alert for the query
24. Databricks repo, pull to get latest code change from git to repo

1) syntax to load csv data in delta table: Ans=> USING CSV clause

2) to verify the result of two tables after using Left outer join: Ans=> it will return all matching records from the left table and missing records in the right table will return null values

3) syntax of to insert data into a table : Ans=> INSERT INTO <table\_name> VALUES (<list of columns>)

4) How to check the owner of the table from data explorer?

5) which privileges are required to give full access on table : Ans=> GRANT ALL PREVILIGES

6) user of Merge statement : Ans=> to avoid duplicate records in the target table

7) difference between Bronze and Silver table? Ans=> Bronze stored raw data and Silver store more refined filtered data

8) Bronze table data load script : Ans => select the script which doesn't have any filter and Aggregation functions

9) use sql command in notebook (with python option selected) : Ans => use %sql magic keyword

10) spark command to run sql queries : Ans => spark.sql

11) Which option in UI to grants access : Ans => Data Explorer

12) difference between Auto loader vs loading from CSV file : Ans: can not use CREATE TABEL AS while reading from CSV file

13) some questions related to Endpoints

14) options to schedule jons : Ans=> Cron

15) why STREAM keyword used before the table : Ans => if it's using Streaming

16) when we dropped the tables both metadata and data files are deleted : Ans: The type of the table is Managed

17) syntax of Filter transformation

18) transform table data from Long to wide : Ans => use Pivot

19) How to setup an alert from UI?

20) convert the existing code to Delta live tables : Ans => convert scripts using SQL syntax

21) python syntax for creating a function

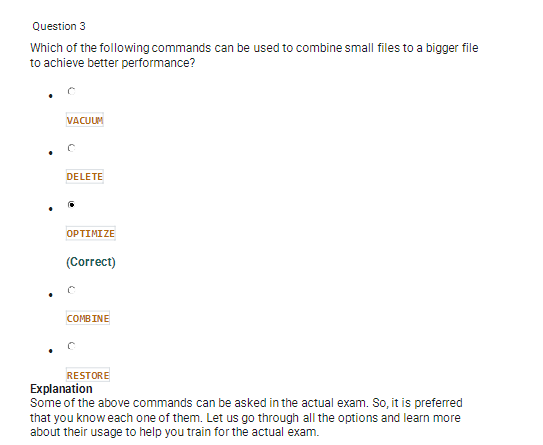
22) syntax for creating streaming spark API : Ana => readstream

Note: It is recommended to use your personal laptop to appear for exam as it asks to disable firewall setting and Antivirus which is not a best practice with workstation and also Admin rights are needed in order to disable Antivirus software so in order to avoid any such issue which is time consuming prefer using personal laptop.

Below given questions are what I remember from exam, these are not exact same question but are related with different wording and different examples so going through this will definitely help.

Questions:

1. Related to Datalake , Datalakehouse and Datawarehouse difference.



Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

A picture containing text

Description automatically generated

Graphical user interface, text, application, email

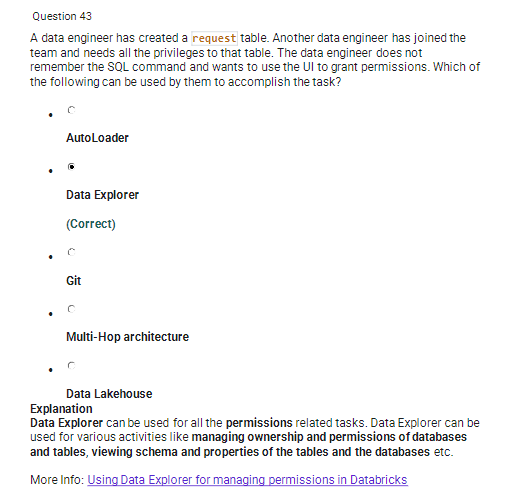
Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated



Note : Other questions were related to Streaming process , Multihop architecture and incremental data processing which were more tricky and need to study properly.

1) Which if the following two options are supported in identifying the arrival of new files, and incremental data from cloud object storage using auto loaded?

--> Directory Listing, File notification

2) You are currently asked to work on building a data pipeline, you have noticed that you are currently working with data source with 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?

--> Delta Live Tables

3) What is the main difference between CREATE STREAMING LIVE TABLE vs CREATE LIVE TABLE?

--> CREATE STREAMING LIVE TABLE is used when working with streaming data sources and incremental data

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

--> Data lakehouse enables both batch and streaming analytics

5) What is the underlying process that makes the Auto Loaded work?

--> Structured Streaming

6) Question about cluster pool usage

7) Grant full privileges query

8) Which of the following are stored in the control flow?

--> Web Application

9) You have noticed the data scientist team is using the notebook versioning feature. You have recommended to switch to using Data repos. What could

be the reason?

--> Databricks Repo allows you to add comments and select changes you want to commit

10) spark query to connect sqllite

--> org.apache.spark.sql.jdbc

11) Spark query to select table data-->spark.table("tablename")

12) gold and silver layer difference

13) The delta live table pipeline is configured to run in production mode using continuous pipeline mode. what is expected outcome after

clicking start to update the pipeline?

--> All dataset will be updated at set intervals until the pipeline is shutdown. The compute resources will be deployed for the update

and terminated when the pipeline is stopped

14) In the absence of table owner, who can grant access?

--> Workplace admin

15) DROP ROW constraint question

16) Wehbook alert

17) Why delta lake house preferred over traditional data warehouse?--> ACID characteristics

18) There is table with PII information. Audit team wants property to notify the type of data in the table. What are the options.

--> Create or replace table user\_pii

COMMENT "Contains PII"

19) Engineer loads previous day data using COPY INTO command. He runs the command but count not changed. --> Previous day data already loaded

20) Task dependency question

21) two different teams created report but report data looks different, why? --> They need to use same trusted resource

22) Question on delta table

23) Question on which command is outside data bricks repo?

Databricks Certified Data Engineer Associate 15/10/2022

Please note that the sentence formation of the question is different from the exam.

1. Syntax for creating an empty table with a specified schema whether it exists or not using sql.

2. Which technology helps in overcoming siloed architecture which is used for extremely specific use cases? Data lake house

3. What is the advantage data lake house gets by embracing open-source technologies? Avoiding Vendor Lock-in

4. What is used to convert long tables to wide tables? PIVOT

5. Syntax for creating a table using a CSV file located in a particular path using sql.

6. Syntax for converting a raw streaming table to a Bronze table using pyspark (python).

7. What is used to combine smaller files and increase performance? OPTIMIZE

8. When is MERGE INTO used instead of INSERT INTO?

9. What does DLT pipeline CONSTRAINT EXPECT ON VIOLATION FAIL UPDATE do? When the constraint is violated, the DLT pipeline will stop and fail.

10. What is stored in customers’ cloud account? Data

11. Where is the database stored when the command CREATE DATABASE company? dbfs:/hive/warehouse

12. What should be done if in a python notebook, a sql query is to be used in a new cell but nothing should change in existing cells? %sql in first line

13. How to write a function which returns the addition of two numbers in python?

14. How is data in databricks stored? Parquet

15. Why does autoloader store all the fields in string when the data source is a JSON file? JSON is a text-based format.

16. Where can you check the owner of a table when you are not sure about your permissions regarding that table? Data explorer owner field (not sure)

17. Where can you check the permissions of data?

18. Syntax for granting all permissions of a table to a team.

19. Which data object is to be created when you want the data to be not stored physically and do not need the data across users or sessions while minimizing costs? Temporary view

20. A SQL Endpoint is to be used for refreshing a dashboard, but it should only run when necessary. There are multiple queries linked to multiple datasets. The data used by the dashboard is automatically generated by python jobs. How should the SQL Endpoint for refreshing the dashboard be configured so that the time taken for refreshing is less?

21. A pipeline is built in python for transformations and the gold table is used by data analysts for making dashboards in SQL. The input for the pipeline is streaming data. Now the team wants to build a DLT pipeline. What changes must be made to make this possible?

22. A data engineer cloned a repo from the central repository and is using it in local repository. Some changes are made to the central repository after cloning. If the data engineer wants to sync the changes what should be done? Git pull

23. Which details are mandatory while creating a new pipeline?

24. DROP TABLE deleted underlying data and metadata. What kind of table was it? Managed

25. Syntax to use FILTER higher order function on an array column.

26. Which trigger type is to be used while writing streaming data in completed mode when you want it to be run in as many batches as needed? .trigger(availableNow = True) (not sure)

27. Which workload is always compatible with autoloader. (options stream loads, batch loads, machine learning loads …) stream load (not sure)

28. Which kind of usecase is suitable for using gold tables? For dashboards

29. What is true about silver tables and bronze tables? Silver tables are cleaner than bronze tables

30. Data engineers found that the queries going to a non-running SQL endpoint are taking longer. What should be done to reduce the run time. Serverless

31. When should we use the “fields depend on” attribute for a newly added task in a job.

32. How to programmatically apply a complex scheduling parameters used in one job to other jobs? Cron scheduling (not sure)

33. Where should we go to schedule a query which is used daily to see the time taken for ingestion of each data source?

34. What is the result when sales table is joined with store table using LEFT JOIN? (The tables are given in the exam)

35. What is the exact reason for using STREAM() in the code

CREATE STREAMING LIVE TABLE loyal\_customers

AS SELECT customer\_id

FROM STREAM(LIVE.customers)

WHERE loyalty = “high”

36. A pipeline has two live table datasets and three streaming live table datasets. Assume previous unprocessed data exists. What is the behavior of data and compute when the pipeline is in continuous and development mode?

37. What should be added to the below code so that it can handle streaming data? Change .read to .readStream

InputDF = ( spark .read

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ .format(“delta”)

.option("maxFilesPerTrigger", 1)

38. How and where to schedule a job to run it daily?

39. Which SQL statement is used to get the count of NULL values in a column of a table?

SELECT count\_if(column IS NULL) FROM table

40. What should be done to send a webhook message to all members when the value is zero in the column “transaction”? Create a alert in new wehook destination (not sure)

41. When are single node clusters used? For interactive querying on small workloads

42. What is the advantage of creating an external table using a parquet source file than a csv source file by CREATE TABLE AS SELECT syntax? CREATE TABLE AS SELECT cannot be used to create an external table (not sure)

43. How to use a python variable table\_name while running a SQL query? Spark.sql(f”SELECT \* FROM {table\_name}”)

Questions:

1. Grant ALL Privileges on Table.
2. Grant Usage permission to see the list of tables available in the Database.
3. If the Original Owner of the Database had left the organization , who can provide the required access to the newly joined team member
4. Difference between Silver and Gold
5. What kind of data is stored in Bronze layer
6. Silver to Gold Pyspark -- Having agg(max("Products"))
7. Benefits of Delta lake over Data lake - ACID properties and Transaction.
8. Vaccum - not able to see the history of the table. What would be the reason ?
9. Operations which can perform outside of Databricks Repo - Clone , Create , Push , Merge(Merge is the answer).
10. Databricks Repo over Databricks Notebooks versioning --> Offers specific things to commit.
11. Delta lake benefits
12. Pyspark - Trigger to process for every 5 sec - .trigger(processingTime='5 seconds')
13. 1 Question on Production Mode -
14. Alert using Webhook as destination
15. How to reduce the cost of SQL endpoint as the team decided to use once or twice in a day - use Autostop future.
16. Team wants to schedule the query to run only for 1 week to monitor the results.

Add an enddate to Schedule.

1. Multiple users are running queries simultaneously - option to increase the performance - increase the max bound.
2. One question on there is a difference in the result between the dataanalysts and Data engineers - use the notebooks to collaborate.
3. Query to remove duplicates - union , inner join , left join , full outer - (UNION answer)
4. Autoloader - to identify only the new files to do an incremental load.
5. Autoloader works for incremental processing - Check pointing and Writeaheadlogging.
6. Connecting to SQLLite - Org.apache.spark.sql.jdbc
7. Which one is in Control plane - Web application
8. Engineer wants to check the quality of table produced in Delta live table - Click on table and check Quality statistics
9. Delta table creation syntax
10. How to check the Database location Describe command
11. Select query to retrieve previous version - Version as of timestamp
12. Delete records greater than 5 (some column name is given) - Delete from table where columnname > 5
13. Drop command on table ran but still able to see the files in data folder - External table
14. If condition and process is true then execute : Ex: If dept == "cse" and process :
15. Comment syntax while creating the table.
16. Merge - to avoid loading duplicates.
17. Insert into to append rows to an existing table
18. Table to physically store data and make it available to other users across the different notebooks and clusters. - read this kind of question properly. View is also correct. But they are stressing to store data physically which is Table.
19. Create SQL function
20. Count(Distinct \*) - to identify the distinct count.
21. Benefit of high order functions - helps to deal with array kind of data from Json files.
22. Streaming Live vs Live tables
23. Constraint drop row - if the condition fails it drops the row and add it to event log.
24. Particular task is taking time - where to monitor. - Jobs - Active runs
25. Question on Data explorer to grant permission from UI.

Databricks associate engineering exam questions (26/04/24) - Sathvik

(Q) Which command will return the location of database sales

**(Ans) DESCRIBE DATABASE sales**

(Q) Advantage of using array functions

**(Ans) Ability to work with complex data from JSON files**

(Q) Command used to write data into a delta table without any duplicate entries

**(Ans) Merge command**

(Q) Analyst created a **table “**sales”and wants it to be tested by engineering team, the engineering team uses python and not sql, which command will the engineering team use?

**(Ans) spark.table(“sales”)**

(Q) A team wants to restore the table to a version that is 3 days old after making a mistake, but not they are not able to restore as data has been deleted, what happened here?

**(Ans) VACUUM command was run**

(Q) Git operation that can be performed outside databricks repos?

**(Ans) Merge**

(Q) When does a team want to use cluster pools?

**(Ans) For a report to be refreshed as quickly as possible**

The above are the questions I remember, I also received questions w.r.t DLT, what is autoloader and its format?  
  
How to grant all privileges to a new user?

How to change or modify owner to a table?

Question related to Union command.

How are bronze and raw schema related?

Which of the following hops is describing silver->gold?

Question related to constraint condition drop row.