

Topic 1 - Exam A

Question #1

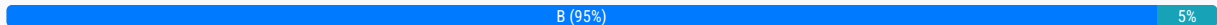
Topic 1

A data organization leader is upset about the data analysis team's reports being different from the data engineering team's reports. The leader believes the siloed nature of their organization's data engineering and data analysis architectures is to blame. Which of the following describes how a data lakehouse could alleviate this issue?

- A. Both teams would autoscale their work as data size evolves
- B. Both teams would use the same source of truth for their work**
- C. Both teams would reorganize to report to the same department
- D. Both teams would be able to collaborate on projects in real-time
- E. Both teams would respond more quickly to ad-hoc requests

Correct Answer: B

Community vote distribution



- prasioso Highly Voted 1 year, 1 month ago

Databricks Lakehouse enables using data as the single source of truth. Duplicating data often results in data silos in organizations. Correct answer B.

upvoted 7 times
- mascarenhaslucas Most Recent 1 week, 2 days ago

**Selected Answer: B**

The answer is B!

upvoted 1 times
- poo\_san 4 weeks ago

**Selected Answer: A**

B is correct

upvoted 1 times
- bettermakeme 2 months, 1 week ago

B is correct answer, I got 100%. all questions came from <https://www.udemy.com/course/practice-exams-databricks-certified-data-engineer-associate-t/?couponCode=APR2024>

upvoted 1 times
- Itmma 3 months ago

**Selected Answer: B**

B is correct

upvoted 1 times

🗨️ **Itmma** 3 months ago

B is correct  
upvoted 1 times

🗨️ **shyemko** 5 months, 2 weeks ago

**Selected Answer: B**

B is correct  
upvoted 1 times

🗨️ **SerGrey** 5 months, 4 weeks ago

**Selected Answer: B**

Correct is B  
upvoted 1 times

🗨️ **VijayKula** 8 months, 2 weeks ago

**Selected Answer: B**

Correct is B  
upvoted 1 times

🗨️ **oscar\_nadie** 8 months, 2 weeks ago

**Selected Answer: B**

Correct is B  
upvoted 1 times

🗨️ **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: B**

Correct ans B  
upvoted 1 times

🗨️ **d\_b47** 8 months, 4 weeks ago

**Selected Answer: B**

Both teams would use the same source of truth for their work  
upvoted 1 times

🗨️ **vprraja03** 9 months, 2 weeks ago

There are 2 versions in Databricks Certified Data Engineer Associate, which version we need to pick for this exam ?  
upvoted 4 times

🗨️ **vcrrhugo** 9 months, 2 weeks ago

B. Both teams would use the same source of truth for their work

A data lakehouse is designed to unify the data engineering and data analysis architectures by integrating features of both data lakes and data warehouses. One of the key benefits of a data lakehouse is that it provides a common, centralized data repository (the "lake") that serves as a single source of truth for data storage and analysis. This allows both data engineering and data analysis teams to work with the same consistent data sets, reducing discrepancies and ensuring that the reports generated by both teams are based on the same underlying data.

Option B addresses the issue of data consistency and alignment between the two teams, which is a common challenge in organizations with separate data engineering and data analysis architectures. By using the same source of truth, the data lakehouse helps alleviate this issue and promotes better collaboration and data integrity.

upvoted 2 times

🗨️ **james\_donquixote** 1 year ago

**Selected Answer: B**

Correct letter B  
upvoted 2 times

🗨️ **[Removed]** 1 year ago

**Selected Answer: B**

Correct letter B  
upvoted 3 times

🗨️ **Data\_4ever** 1 year, 2 months ago

**Selected Answer: B**

Unity Catalog in Databricks helps to eliminate Data Silos in an organization by having one single source of truth data.  
upvoted 4 times

Which of the following describes a scenario in which a data team will want to utilize cluster pools?


- A. An automated report needs to be refreshed as quickly as possible.**
- B. An automated report needs to be made reproducible.
- C. An automated report needs to be tested to identify errors.
- D. An automated report needs to be version-controlled across multiple collaborators.
- E. An automated report needs to be runnable by all stakeholders.

**Correct Answer: E**

Community vote distribution

A (97%)


3%

 **Data\_4ever** Highly Voted 1 year, 2 months ago

**Selected Answer: A**

Using cluster pools reduces the cluster startup time. So in this case, the reports can be refreshed quickly and not having to wait long for the cluster to start

upvoted 14 times

 **mascarenhaslucas** Most Recent 1 week, 2 days ago

**Selected Answer: A**

I believe it's A!

upvoted 1 times

 **poo\_san** 4 weeks ago

**Selected Answer: A**

A is the correct answer as cluster pools are used to speed up the cluster startup time

upvoted 1 times

 **M15** 1 month, 4 weeks ago

Considering the recommendation to create pools based on workloads and to pre-populate pools to ensure instances are available when cluster need them, the most suitable option would be:

E. An automated report needs to be runnable by all stakeholders.

This aligns with the concept of pre-populating pools to ensure that instances are readily available when needed, enabling the automated report to be executed promptly whenever stakeholders require it without waiting for instance acquisition.

upvoted 2 times

 **benni\_ale** 2 months, 2 weeks ago

**Selected Answer: A**

A : I think cluster pools are used mainly to accelerate cluster start up by using vms somehow.

upvoted 1 times

 **ltmma** 3 months ago

**Selected Answer: A**

A is correct

upvoted 1 times

🗨️ 👤 **Huepig** 3 months, 1 week ago

**Selected Answer: A**

<https://www.databricks.com/blog/2019/11/11/databricks-pools-speed-up-data-pipelines.html>

upvoted 1 times

🗨️ 👤 **agAshish** 4 months, 2 weeks ago

E is correct for sure. For data team , their tasks is not just to refresh a report. They equally want to share the cluster for running their queries. Please read at below:

<https://docs.databricks.com/en/compute/pool-best-practices.html#create-pools-based-on-workloads>

upvoted 1 times

🗨️ 👤 **SerGrey** 5 months, 4 weeks ago

**Selected Answer: A**

A is correct

upvoted 1 times

🗨️ 👤 **Ajinkyavsawant7** 6 months, 3 weeks ago

**Selected Answer: A**

A is correct

upvoted 1 times

🗨️ 👤 **anandpsg101** 8 months, 1 week ago

**Selected Answer: A**

A is correct

upvoted 2 times

🗨️ 👤 **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: A**

Cluster pools are allows us to reduce the start time Ans A

upvoted 1 times

🗨️ 👤 **d\_b47** 8 months, 4 weeks ago

**Selected Answer: A**

.Cluster pools allow us to reserve VM's ahead of time, which means that its start-up time will be faster.

upvoted 1 times

🗨️ 👤 **len** 8 months, 4 weeks ago

Option: A is correct.

upvoted 1 times

🗨️ 👤 **alexitogs** 9 months, 2 weeks ago

**Selected Answer: A**

Cluster pools allow us to reserve VM's ahead of time, which means that its start-up time will be faster.

upvoted 1 times

🗨️ 👤 **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: A**

A. An automated report needs to be refreshed as quickly as possible.

Cluster pools are typically used in distributed computing environments, such as cloud-based data platforms like Databricks. They allow you to pre-allocate a set of compute resources (a cluster) for specific tasks or workloads. In this case, if an automated report needs to be refreshed as quickly as possible, you can allocate a cluster pool with sufficient resources to ensure fast data processing and report generation. This helps ensure that the report is generated with minimal latency and can be delivered to stakeholders in a timely manner. Cluster pools allow you to optimize resource allocation for high-demand, time-sensitive tasks like real-time report generation.

upvoted 3 times

🗨️ 👤 **Gajen100** 10 months, 3 weeks ago

**Selected Answer: A**

An automated report needs to be refreshed as quickly as possible.

upvoted 1 times

Which of the following is hosted completely in the control plane of the classic Databricks architecture?

- A. Worker node
- B. JDBC data source
- C. Databricks web application**
- D. Databricks Filesystem
- E. Driver node



**Correct Answer: E**

Community vote distribution

C (100%)

  **h79** Highly Voted 1 year, 2 months ago



I disagree with this answer. I think its the databricks web app that is always in the control plane  
upvoted 11 times

  **XiltroX** 1 year, 2 months ago

Agreed. Its Web app for sure  
upvoted 1 times

  **XiltroX** 1 year, 2 months ago

I think I meant to say that its option C. Not sure why I said that I agree with the answer in Examtopics. Option C for sure.  
upvoted 1 times

  **vctrhugo** Highly Voted 9 months, 2 weeks ago

**Selected Answer: C**

C. Databricks web application

In the classic Databricks architecture, the control plane includes components like the Databricks web application, the Databricks REST API, and the Databricks Workspace. These components are responsible for managing and controlling the Databricks environment, including cluster provisioning, notebook management, access control, and job scheduling.

The other options, such as worker nodes, JDBC data sources, Databricks Filesystem (DBFS), and driver nodes, are typically part of the data plane or the execution environment, which is separate from the control plane. Worker nodes are responsible for executing tasks and computations, JDBC data sources are used to connect to external databases, DBFS is a distributed file system for data storage, and driver nodes are responsible for coordinating the execution of Spark jobs.

upvoted 5 times

  **mascarenhaslucas** Most Recent 1 week, 2 days ago

**Selected Answer: C**

The answer is C! According to the Databricks documentation, a cluster consists of one driver node and zero or more worker nodes, by default the driver node uses the same instance type as the worker node.

upvoted 1 times

  **pierrickaoasis** 1 month, 3 weeks ago

Who decide the correct answer on this website ? CertilQ says C ; ITExams says E.... For me it's C  
upvoted 1 times

  **benni\_ale** 2 months, 2 weeks ago

**Selected Answer: C**

Nodes are on the Data Plane. I think the Web App is the only one in the Control Pane.  
upvoted 1 times

🗨️ **ltmma** 3 months ago

**Selected Answer: C**

C is correct  
upvoted 1 times

🗨️ **kirshoff** 3 months, 2 weeks ago

Answer is C: [https://docs.databricks.com/en/\\_images/databricks-architecture-aws.png](https://docs.databricks.com/en/_images/databricks-architecture-aws.png)  
upvoted 1 times

🗨️ **agAshish** 4 months, 2 weeks ago

E.Driver Node , is the correct answer.  
In the classic Databricks architecture, the control plane includes components responsible for managing and coordinating the execution of tasks. The driver node is part of the control plane, and it handles the coordination and execution of the overall Spark application.  
upvoted 1 times

🗨️ **Isio05** 1 month, 3 weeks ago

Cluster nodes (both driver and worker) are located on customer cloud account. So E is not the correct answer here.  
upvoted 1 times

🗨️ **poundmanluffy** 5 months, 3 weeks ago

**Selected Answer: C**

Web application always resides in Control Plane  
upvoted 2 times

🗨️ **SerGrey** 5 months, 4 weeks ago

**Selected Answer: C**

C is correct  
upvoted 1 times

🗨️ **CHHIPA** 6 months, 2 weeks ago

**Selected Answer: C**

CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC  
upvoted 1 times

🗨️ **Ajinkyavsawant7** 6 months, 3 weeks ago

**Selected Answer: C**

C is correct  
upvoted 1 times

🗨️ **Huroye** 7 months, 1 week ago

It is the web UI. C  
upvoted 1 times

🗨️ **Sriramiyer92** 8 months, 3 weeks ago

Reading material: <https://learn.microsoft.com/en-us/azure/databricks/getting-started/overview>  
upvoted 2 times

🗨️ **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: C**

Correct ans C  
upvoted 1 times

🗨️ **d\_b47** 8 months, 4 weeks ago

**Selected Answer: C**

web application  
upvoted 1 times

🗨️ **Lipon23** 9 months, 3 weeks ago

**Selected Answer: C**

Databricks Web App for sure  
upvoted 1 times

Question #4

Topic 1

Which of the following benefits of using the Databricks Lakehouse Platform is provided by Delta Lake?

- A. The ability to manipulate the same data using a variety of languages
- B. The ability to collaborate in real time on a single notebook
- C. The ability to set up alerts for query failures
- D. The ability to support batch and streaming workloads**
- E. The ability to distribute complex data operations

**Correct Answer:** D

Community vote distribution

D (100%)

  **mascarenhaslucas** 1 week, 2 days ago



**Selected Answer: D**

The answer is D!  
upvoted 1 times

  **Itmma** 3 months ago



**Selected Answer: D**

D is correct  
upvoted 2 times

  **VijayKula** 8 months, 2 weeks ago

**Selected Answer: D**

Answer is D  
upvoted 1 times

  **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: D**

Correct and D  
upvoted 1 times

  **vctrhugo** 9 months, 2 weeks ago

## Question #5

Topic 1

Which of the following describes the storage organization of a Delta table?

- A. Delta tables are stored in a single file that contains data, history, metadata, and other attributes.
- B. Delta tables store their data in a single file and all metadata in a collection of files in a separate location.
- C. Delta tables are stored in a collection of files that contain data, history, metadata, and other attributes**
- D. Delta tables are stored in a collection of files that contain only the data stored within the table.
- E. Delta tables are stored in a single file that contains only the data stored within the table.

**Correct Answer: C**


Community vote distribution

C (100%)

  **mascarenhaslucas** 1 week, 2 days ago

**Selected Answer: C**

The answer is C!  
upvoted 1 times

  **benni\_ale** 2 months, 2 weeks ago

**Selected Answer: C**

GPT4:  
Delta tables in Databricks use:  
Parquet format files for data storage.  
A\_delta\_log folder for JSON log files that track transactions.  
Scheme enforcement in metadata to ensure consistency.  
Checkpoint files to speed up the rebuilding of the table state.  
upvoted 3 times

  **Itmma** 3 months ago

**Selected Answer: C**

C is correct  
upvoted 1 times



🗄️ 👤 **SerGrey** 5 months, 4 weeks ago

**Selected Answer: C**

C is correct  
upvoted 1 times

🗄️ 👤 **VijayKula** 8 months, 2 weeks ago

Answer is C  
upvoted 1 times

🗄️ 👤 **Sriramiyer92** 8 months, 3 weeks ago

Reading Material:  
5 reasons to choose Delta format (on Databricks)  
<https://medium.com/datalex/5-reasons-to-use-delta-lake-format-on-databricks-d9e76cf3e77d>  
upvoted 2 times

🗄️ 👤 **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: C**

Correct ans C  
upvoted 1 times

🗄️ 👤 **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: C**

C. Delta tables are stored in a collection of files that contain data, history, metadata, and other attributes.

Delta tables store data in a structured manner using Parquet files, and they also maintain metadata and transaction logs in separate directories. This organization allows for versioning, transactional capabilities, and metadata tracking in Delta Lake. Thank you for pointing out the error, and appreciate your understanding.

upvoted 3 times

🗄️ 👤 **andie123** 10 months, 1 week ago

**Selected Answer: C**

C is the right answer  
upvoted 2 times

🗄️ 👤 **Atnafu** 11 months, 2 weeks ago

C  
Delta tables in Databricks Delta Lake are stored in a collection of files organized in a directory structure. This directory structure includes data files, transaction log files, and metadata files. These files are stored in a specified location, typically in a distributed file system such as Hadoop Distributed File System (HDFS) or Amazon S3.

upvoted 1 times

🗄️ 👤 **prasioso** 1 year, 1 month ago

First selected D as I assumed the data to be stored in the Delta lake and the transaction log to be stored separately. However, documentation states when a user creates a Delta Lake table, that table's transaction log is automatically created in the `_delta_log` subdirectory. The `deltalog` contains multiple files hence a collection of files. Answer C.

upvoted 3 times

🗄️ 👤 **Data\_4ever** 1 year, 2 months ago

**Selected Answer: C**

C is the right option  
upvoted 3 times

🗄️ 👤 **knivesz** 1 year, 2 months ago

**Selected Answer: C**

C , respuesta correcta  
upvoted 1 times

🗄️ 👤 **XiltroX** 1 year, 2 months ago

C is correct answer  
<https://docs.delta.io/latest/delta-faq.html#:~:text=Delta%20Lake%20uses%20versioned%20Parquet,directory%20to%20provide%20ACID%20transactions.>  
upvoted 2 times

## Question #6

Topic 1

Which of the following code blocks will remove the rows where the value in column age is greater than 25 from the existing Delta table my\_table and save the updated table?


- A. `SELECT * FROM my_table WHERE age > 25;`
- B. `UPDATE my_table WHERE age > 25;`
- C. `DELETE FROM my_table WHERE age > 25;`**
- D. `UPDATE my_table WHERE age <= 25;`
- E. `DELETE FROM my_table WHERE age <= 25;`

**Correct Answer: C**

Community vote distribution

C (92%)


8%

 **mascahenhaslucas** 1 week, 2 days ago

**Selected Answer: C**

The answer is C!

upvoted 1 times

 **Svengance** 2 months ago

**Selected Answer: A**

there is not delete history option just the vacuum with its parameters of time retention.

upvoted 1 times

 **bettermakeme** 2 months, 3 weeks ago

Answer is C. Just finished exam-got 100% [Databricks Associate Exam Practice Exams] All questions came from Databricks Certified Data Engineer Associate

[https://www.udemy.com/share/10aEFa3@9M\\_uT6vrKbnl68tOK96kfy-YWitjwzLTIVCrzPs-0hGUu8fyX8V4Tn\\_x\\_y65bwLm/](https://www.udemy.com/share/10aEFa3@9M_uT6vrKbnl68tOK96kfy-YWitjwzLTIVCrzPs-0hGUu8fyX8V4Tn_x_y65bwLm/)


upvoted 3 times

 **ltmma** 3 months ago

**Selected Answer: C**


C is correct

upvoted 1 times

 **SerGrey** 5 months, 4 weeks ago

**Selected Answer: C**C. `DELETE FROM my_table WHERE age > 25;`



upvoted 1 times

 **VijayKula** 8 months, 2 weeks ago

**Selected Answer: C**

Answer is C



upvoted 1 times

  **DavidRou** 8 months, 3 weeks ago

**Selected Answer: C**

C is the correct answer as the SELECT statement allows to query a table, the UPDATE statement allows to modify values in columns. If you want to remove rows that don't match a specific condition you must use DELETE

upvoted 2 times

  **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: C**



C is correct

upvoted 1 times

  **ArindamNath** 9 months, 1 week ago

C is correct



upvoted 1 times

  **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: C**



C. DELETE FROM my\_table WHERE age > 25;

upvoted 1 times

  **nb1000** 11 months, 3 weeks ago

C is correct

upvoted 1 times

  **prasioso** 1 year, 1 month ago

C is correct. use DELETE FROM to delete existing records from the table. UPDATE is used to modify existing records. SELECT only creates a view, it does not alter the table records.

upvoted 2 times

  **Varma\_Saraswathula** 1 year, 2 months ago

C - is correct answer

upvoted 1 times

  **knivesz** 1 year, 2 months ago

**Selected Answer: C**

C es correcto



upvoted 2 times

  **surrabhi\_4** 1 year, 2 months ago

**Selected Answer: C**

option c

upvoted 1 times

  **XiltroX** 1 year, 2 months ago

C is the correct answer

upvoted 3 times

## Question #7

Topic 1

A data engineer has realized that they made a mistake when making a daily update to a table. They need to use Delta time travel to restore the table to a version that is 3 days old. However, when the data engineer attempts to time travel to the older version, they are unable to restore the data because the data files have been deleted.

Which of the following explains why the data files are no longer present?

- A. The VACUUM command was run on the table**
- B. The TIME TRAVEL command was run on the table
- C. The DELETE HISTORY command was run on the table

D. The OPTIMIZE command was run on the table

E. The HISTORY command was run on the table

**Correct Answer: C**

Community vote distribution

A (100%)

🗳️ **Feroz\_Raza** Highly Voted 7 months ago

**Selected Answer: A**

There is no DELETE HISTORY command in Databricks

VACCUUM command can remove history and we can also specify the retention period with VACCUUM Command. Default Retention period is 7 days.

To allow changing the default retention period you can run the following command

```
ALTER TABLE your_table SET TBLPROPERTIES ('delta.retentionDurationCheck.enabled' = 'true');
```

upvoted 6 times

🗳️ **potaryxkug** Most Recent 3 days, 9 hours ago

A is the good answer

upvoted 1 times

🗳️ **mascarenhaslucas** 1 week, 2 days ago

The answer is A!

upvoted 1 times

🗳️ **bettermakeme** 2 months, 3 weeks ago

Answer is A. Just finished exam-got 100% [Databricks Associate Exam Practice Exams] All questions came from Databricks Certified Data Engineer Associate

[https://www.udemy.com/share/10aEFa3@9M\\_uT6vrKbnl68tOK96kfy-YWitjwzLTIVCrzPs-0hGUu8fyX8V4Tn\\_x\\_y65bwLm/](https://www.udemy.com/share/10aEFa3@9M_uT6vrKbnl68tOK96kfy-YWitjwzLTIVCrzPs-0hGUu8fyX8V4Tn_x_y65bwLm/)

upvoted 1 times

🗳️ **ltmma** 3 months ago

**Selected Answer: A**

A is correct

upvoted 1 times

🗳️ **SerGrey** 5 months, 4 weeks ago

**Selected Answer: A**

A i correct

upvoted 1 times

🗳️ **Huroye** 7 months ago

I agree with the first post. A is the correct answer. There is no such thing as a Delete History Command

upvoted 2 times

🗳️ **awofalus** 7 months, 2 weeks ago

**Selected Answer: A**

right answer is A

upvoted 1 times

🗳️ **vivekr** 7 months, 3 weeks ago

i think B is the answer, plz let me know if not correct

upvoted 1 times

🗳️ **vivekr** 7 months, 3 weeks ago

but vacuum allows to vacuum anything that's older than 7 days right



upvoted 1 times

🗳️ **VijayKula** 8 months, 2 weeks ago

**Selected Answer: A**



Answer is A Vacuum

upvoted 1 times

  **Sriramiyer92** 8 months, 3 weeks ago

Reading Material: <https://learn.microsoft.com/en-us/azure/databricks/delta/vacuum#example-syntax-for-vacuum>


upvoted 1 times

  **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: A**

A is correct

upvoted 2 times

  **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: A**

A. The VACUUM command was run on the table

The VACUUM command in Delta Lake is used to clean up and remove unnecessary data files that are no longer needed for time travel or query purposes. When you run VACUUM with certain retention settings, it can delete older data files, which might include versions of data that are older than the specified retention period. If the data engineer is unable to restore the table to a version that is 3 days old because the data files have been deleted, it's likely because the VACUUM command was run on the table, removing the older data files as part of data cleanup.



upvoted 3 times

  **cpalmier** 10 months ago

A is Correct!

Does DELETE HISTORY command exist?

upvoted 1 times

  **Atnafu** 11 months, 2 weeks ago



A

When the data engineer attempted to time travel to an older version of the table, the data files were no longer present because the VACUUM command was run on the table. The VACUUM command in Delta Lake is used to clean up files that are no longer necessary for the current version of the table. It permanently removes older versions of data files and transaction log files that are no longer needed for queries or time travel.

By running the VACUUM command, the data engineer inadvertently deleted the data files of the version they were trying to restore, making it impossible to access that specific version of the table through Delta time travel.


VACUUM [db\_name.]table\_name [RETAIN num\_hrs] [DRY RUN]

upvoted 1 times

  **Atnafu** 11 months, 2 weeks ago



VACUUM my\_database.my\_table RETAIN 30

upvoted 1 times

  **Atnafu** 11 months, 2 weeks ago

If not specified, the default retention period of 7 days is used.

upvoted 2 times

  **Majjiji** 1 year, 1 month ago

**Selected Answer: A**

The most likely reason why the data files are no longer present when the data engineer attempts to time travel to an older version of a Delta table is that the VACUUM command was run on the table. The VACUUM command removes files that are no longer in use by the Delta table, including files that are required for time travel. Therefore, if the VACUUM command is run on a Delta table, it can make it impossible to use time travel to recover older versions of the table.

upvoted 4 times

## Question #8

Topic 1

Which of the following Git operations must be performed outside of Databricks Repos?

A. Commit

B. Pull

C. Push

D. Clone

**E. Merge**




**Correct Answer: D**

Community vote distribution

E (100%)

  **ZSun**  1 year ago


According to the most recent one, all command is feasible in Repos  
upvoted 17 times

  **NickWerbung**  9 months, 3 weeks ago

Not valid anymore...  
<https://docs.databricks.com/en/repos/ci-cd-techniques-with-repos.html>  
upvoted 10 times



  **Isio05**  1 month, 3 weeks ago

Confirmed on live environment - merging is now possible directly in Databricks Repos  
upvoted 1 times

  **ltmma** 3 months ago



**Selected Answer: E**

E is correct  
upvoted 1 times



  **Bob123456** 3 months, 3 weeks ago

Why not option B Pull



The following tasks are not supported by Databricks Repos, and must be performed in your Git provider:  
Create a pull request  
Delete branches  
Merge and rebase branches \*  
upvoted 1 times

  **Isio05** 1 month, 3 weeks ago

Pull is not the same as pull request. Pulls are updating local version of the repo to the one present on remote. And it's surely feasible in Databricks repos.  
upvoted 2 times

  **vvg130** 5 months, 2 weeks ago

The new answer is F - Delete .  
upvoted 4 times

  **SerGrey** 5 months, 4 weeks ago

**Selected Answer: E**

E is correct  
upvoted 1 times

  **nedlo** 6 months, 3 weeks ago

i think it supports merge now  
<https://docs.databricks.com/en/repos/git-operations-with-repos.html>  
"If an operation such as pull, rebase, or merge causes a merge conflict, the Repos UI shows a list of files with conflicts and options for resolving the conflicts.

You have two primary options:

Use the Repos UI to resolve the conflict."  
upvoted 1 times

- 🗄️ 👤 **Huroye** 7 months ago  
E is the correct answer given the selections. You can clone.  
upvoted 1 times
- 🗄️ 👤 **mokrani** 7 months, 3 weeks ago  
According to the recent version, all commands are supported under Repos !  
upvoted 1 times
- 🗄️ 👤 **KalavathiP** 8 months, 3 weeks ago  
**Selected Answer: E**  
E is correct  
upvoted 2 times
- 🗄️ 👤 **fred\_camargo** 1 year ago  
**Selected Answer: E**  
Merge is the correct answer  
upvoted 2 times
- 🗄️ 👤 **Majjiji** 1 year, 1 month ago  
**Selected Answer: E**  
For following tasks, work in your Git provider:  
  
Create a pull request.  
Resolve merge conflicts.  
Merge or delete branches.  
Rebase a branch.  
  
<https://docs.databricks.com/repos/index.html>  
upvoted 2 times
- 🗄️ 👤 **Varma\_Saraswathula** 1 year, 2 months ago  
Merge - A  
upvoted 1 times
- 🗄️ 👤 **naxacod574** 1 year, 2 months ago  
merge is not supported  
upvoted 1 times
- 🗄️ 👤 **SireeJ** 1 year, 2 months ago  
Option: C  
upvoted 1 times
- 🗄️ 👤 **Data\_4ever** 1 year, 2 months ago  
**Selected Answer: E**  
MERGE is the only git operation that is listed in the options that cannot be performed with Databricks repos. CLONE is absolutely possible  
upvoted 3 times

#### Question #9

Topic 1

Which of the following data lakehouse features results in improved data quality over a traditional data lake?

A. A data lakehouse provides storage solutions for structured and unstructured data.

**B. A data lakehouse supports ACID-compliant transactions.**

C. A data lakehouse allows the use of SQL queries to examine data.

D. A data lakehouse stores data in open formats.

E. A data lakehouse enables machine learning and artificial Intelligence workloads.

**Correct Answer: C**

Community vote distribution

B (100%)

 **ThomasReps** Highly Voted 1 year ago

**Selected Answer: B**

Option B - ACID Properties Source: <https://www.databricks.com/blog/2021/08/30/frequently-asked-questions-about-the-data-lakehouse.html#five> -> "Lakehouse tackles the fundamental issues that make data swamps out of data lakes. It adds ACID transactions to ensure consistency as multiple parties concurrently read or write data."

upvoted 5 times

 **mascahenhaslucas** Most Recent 1 week, 2 days ago

**Selected Answer: B**

The answer is B!

upvoted 1 times

 **bm.bala.murugan.sb** 1 month ago

**Selected Answer: B**

B is Correct

upvoted 1 times

 **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: B**

b is correct as acid transactions are related to data quality

upvoted 1 times

 **benni\_ale** 2 months, 2 weeks ago

**Selected Answer: B**

b is correct


upvoted 2 times

 **Itmma** 3 months ago

**Selected Answer: B**

B is correct

upvoted 1 times

 **SerGrey** 5 months, 2 weeks ago

**Selected Answer: B**

Correct andwer is B

upvoted 1 times

 **Garyn** 5 months, 3 weeks ago

**Selected Answer: B**

B. A data lakehouse supports ACID-compliant transactions.

ACID (Atomicity, Consistency, Isolation, Durability) compliance ensures that transactions are processed reliably and consistently, which is crucial for maintaining data integrity and quality. It helps in preventing data inconsistencies or errors that might occur during concurrent transactions data operations. By supporting ACID-compliant transactions, a data lakehouse can enhance data quality by providing mechanisms to ensure reliability and consistency of data operations, which is a notable improvement over traditional data lakes that might lack such transactional capabilities.

upvoted 2 times



🗨️ **Huroye** 7 months, 1 week ago

point of correction: B is the correct answer, not A as posted in my previous comment

upvoted 1 times

🗨️ **Huroye** 7 months, 1 week ago

A is the correct answer. Allowing SQL query to examine data does not improve the quality of the data. For example, if the person writes the wrong query they will get a wrong answer. So how does that improve the quality of the data?

upvoted 1 times

🗨️ **awofalus** 7 months, 2 weeks ago

**Selected Answer: B**

Correct : B

upvoted 1 times

🗨️ **VijayKula** 8 months, 2 weeks ago

**Selected Answer: B**

Acid Transaction is the correct answer

upvoted 1 times

🗨️ **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: B**

B is correct

upvoted 1 times

🗨️ **d\_b47** 8 months, 4 weeks ago

**Selected Answer: B**

ACID is the correct ans.

upvoted 1 times

🗨️ **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: B**

B. A data lakehouse supports ACID-compliant transactions.

One of the key features of a data lakehouse that results in improved data quality over a traditional data lake is its support for ACID (Atomicity, Consistency, Isolation, Durability) transactions. ACID transactions provide data integrity and consistency guarantees, ensuring that operations on the data are reliable and that data is not left in an inconsistent state due to failures or concurrent access.

In a traditional data lake, such transactional guarantees are often lacking, making it challenging to maintain data quality, especially in scenarios involving multiple data writes, updates, or complex transformations. A data lakehouse, by offering ACID compliance, helps maintain data quality by providing strong consistency and reliability, which is crucial for data pipelines and analytics.

upvoted 4 times

🗨️ **hany\_ds** 10 months, 1 week ago

B is correct answer. Only B deals with data quality i.e. ACID transactions make sure all the transactions are complete, isolated, consistent and durable.

upvoted 1 times

🗨️ **mehroosali** 11 months, 2 weeks ago

**Selected Answer: B**

B is correct

upvoted 2 times

## Question #10

Topic 1

A data engineer needs to determine whether to use the built-in Databricks Notebooks versioning or version their project using Databricks Repos. Which of the following is an advantage of using Databricks Repos over the Databricks Notebooks versioning?

A. Databricks Repos automatically saves development progress

**B. Databricks Repos supports the use of multiple branches**

- C. Databricks Repos allows users to revert to previous versions of a notebook
- D. Databricks Repos provides the ability to comment on specific changes
- E. Databricks Repos is wholly housed within the Databricks Lakehouse Platform

**Correct Answer: B**

*Community vote distribution*

B (100%)

🗳️ 👤 **Majiji** Highly Voted 👍 1 year, 1 month ago

**Selected Answer: B**

While both Databricks Notebooks versioning and Databricks Repos allow for version control of code, Databricks Repos provides the additional benefit of supporting the use of multiple branches. This allows for multiple versions of a notebook or project to be developed in parallel, facilitating collaboration among team members and simplifying the process of merging changes into a single main branch.

upvoted 8 times

🗳️ 👤 **benni\_ale** Most Recent ⌚ 1 month, 3 weeks ago

**Selected Answer: B**

b , multiple branches are not supported at all without a git integration and databricks repos have built in UI for governing such a thing

upvoted 1 times

🗳️ 👤 **benni\_ale** 1 month, 3 weeks ago

b , multiple branches are not supported at all without a git integration and databricks repos have built in UI for governing such a thing

upvoted 1 times

🗳️ 👤 **benni\_ale** 2 months, 2 weeks ago

**Selected Answer: B**

B is correct

upvoted 1 times

🗳️ 👤 **Itmma** 3 months ago

**Selected Answer: B**

B is correct

upvoted 1 times

🗳️ 👤 **SerGrey** 5 months, 2 weeks ago

**Selected Answer: B**

Correct answer is B

upvoted 1 times

🗳️ 👤 **awofalus** 7 months, 2 weeks ago

**Selected Answer: B**

Correct : B

upvoted 1 times

🗳️ 👤 **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: B**

B is correct

upvoted 1 times

🗨️ **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: B**

B. Databricks Repos supports the use of multiple branches.

An advantage of using Databricks Repos over the built-in Databricks Notebooks versioning is the ability to work with multiple branches. Branching is a fundamental feature of version control systems like Git, which Databricks Repos is built upon. It allows you to create separate branches for different tasks, features, or experiments within your project. This separation helps in parallel development and experimentation without affecting the main branch or the work of other team members.

Branching provides a more organized and collaborative development environment, making it easier to merge changes and manage different development efforts. While Databricks Notebooks versioning also allows you to track versions of notebooks, it may not provide the same level of flexibility and collaboration as branching in Databricks Repos.

upvoted 2 times

🗨️ **hany\_ds** 10 months, 1 week ago

B

built in databricks notebook versioning does not allow multiple branches.

upvoted 1 times

🗨️ **Atnafu** 11 months, 2 weeks ago

B

An advantage of using Databricks Repos over the Databricks Notebooks versioning is that Databricks Repos supports the use of multiple branches. With Databricks Repos, you can create and manage multiple branches of your codebase, enabling parallel development, collaboration and the ability to work on different features or bug fixes simultaneously.

upvoted 2 times

🗨️ **Varma\_Saraswathula** 1 year, 2 months ago

B. Databricks Repos supports the use of multiple branches

upvoted 1 times

🗨️ **sdas1** 1 year, 2 months ago

Option B

upvoted 2 times

🗨️ **surabhi\_4** 1 year, 2 months ago

**Selected Answer: B**

option B

upvoted 2 times

🗨️ **XiltroX** 1 year, 2 months ago

**Selected Answer: B**

Question #11

Topic 1

A data engineer has left the organization. The data team needs to transfer ownership of the data engineer's Delta tables to a new data engineer. The new data engineer is the lead engineer on the data team.

Assuming the original data engineer no longer has access, which of the following individuals must be the one to transfer ownership of the Delta tables in Data Explorer?

- A. Databricks account representative
- B. This transfer is not possible
- C. Workspace administrator**
- D. New lead data engineer
- E. Original data engineer

**Correct Answer: D**

Community vote distribution

C (100%)

🗨️ **carlosmps** 1 month ago

Option c  
upvoted 1 times

🗄️ **Svengance** 2 months ago

**Selected Answer: C**

workspace admin  
upvoted 2 times

🗄️ **benni\_ale** 2 months, 2 weeks ago

**Selected Answer: C**

workspace admin  
upvoted 1 times

🗄️ **Itmma** 3 months ago

**Selected Answer: C**

C is correct  
upvoted 1 times

🗄️ **SerGrey** 5 months, 2 weeks ago

**Selected Answer: C**

Correct answer is C  
upvoted 1 times

🗄️ **Garyn** 5 months, 3 weeks ago

**Selected Answer: C**

The ownership of Delta tables in Data Explorer can be transferred by the current owner, a metastore admin, or the owner of the container<sup>1</sup>. In case, since the original data engineer no longer has access, the Workspace Administrator (Option C) would be the most appropriate individual to transfer the ownership of the Delta tables to the new data engineer. This is because the Workspace Administrator typically has the necessary permissions to manage such resources<sup>2</sup>. Please note that the exact process may vary depending on the specific configurations and permissions set up in your workspace. It's always a good idea to consult with your organization's IT or data governance team to ensure the correct procedures are followed.

upvoted 2 times

🗄️ **Huroye** 7 months, 1 week ago

The answer is C - the workspace admin. How can it be the new DE? You do not even know if the new DE has access. That was not stated and you cannot consider it.

upvoted 2 times

🗄️ **VijayKula** 8 months, 2 weeks ago

**Selected Answer: C**

Workspace Administrator  
upvoted 1 times

🗄️ **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: C**

C is correct  
upvoted 1 times

🗄️ **d\_b47** 8 months, 4 weeks ago

**Selected Answer: C**

It should be "workspace administrator"  
upvoted 1 times

🗄️ **Niteesh** 11 months, 3 weeks ago

**Selected Answer: C**

It should be "workspace administrator"  
upvoted 1 times

🗄️ **abdullahmyahya** 1 year ago

**Selected Answer: C**

It's C  
upvoted 1 times

🗄️ 👤 **ThomasReps** 1 year ago

**Selected Answer: C**

It's C.

upvoted 1 times

🗄️ 👤 **Bob123456** 1 year, 1 month ago

why can't it be a 'Workspace administrator' instead of option 'D'

upvoted 1 times

🗄️ 👤 **Majjji** 1 year, 1 month ago

**Selected Answer: C**

The Workspace administrator must be the one to transfer ownership of the Delta tables in Data Explorer in this scenario.

upvoted 2 times

🗄️ 👤 **Varma\_Saraswathula** 1 year, 2 months ago

Option C -

<https://docs.databricks.com/sql/admin/transfer-ownership.html>

upvoted 3 times

🗄️ 👤 **naxacod574** 1 year, 2 months ago

Option C

upvoted 1 times

## Question #12

Topic 1

A data analyst has created a Delta table sales that is used by the entire data analysis team. They want help from the data engineering team to implement a series of tests to ensure the data is clean. However, the data engineering team uses Python for its tests rather than SQL.

Which of the following commands could the data engineering team use to access sales in PySpark?

A. `SELECT * FROM sales`

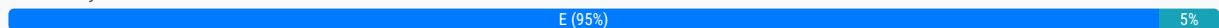
B. There is no way to share data between PySpark and SQL.

C. `spark.sql("sales")` D. `spark.delta.table("sales")`

**E. `spark.table("sales")`**

**Correct Answer: D**

Community vote distribution





🗄️ 👤 **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: E**

E is correct

upvoted 1 times

  **benni\_ale** 2 months, 2 weeks ago



**Selected Answer: E**

e is correct  
upvoted 2 times

  **ltmma** 3 months ago



**Selected Answer: E**

E is correct  
upvoted 1 times

  **SerGrey** 5 months, 2 weeks ago

**Selected Answer: E**

Correct answer is E  
upvoted 1 times



  **Garyn** 5 months, 3 weeks ago

**Selected Answer: E**

E. spark.table("sales")

The spark.table() function in PySpark allows access to a registered table within the SparkSession. In this case, "sales" is the name of the Delta table created by the data analyst, and the spark.table() function enables access to this table for performing data engineering tests using PySpark.

upvoted 4 times



  **csd** 5 months, 3 weeks ago

C is correct Answer  
upvoted 1 times

  **awofalus** 7 months, 2 weeks ago

**Selected Answer: E**

Correct is E  
upvoted 1 times

  **KalavathiP** 8 months, 3 weeks ago



**Selected Answer: E**

E is correct  
upvoted 1 times

  **d\_b47** 8 months, 4 weeks ago

**Selected Answer: E**

delta is default.  
upvoted 1 times

  **Atnafu** 11 months, 2 weeks ago

E

The spark.table() function in PySpark allows you to access tables registered in the catalog, including Delta tables. By specifying the table name ("sales"), the data engineering team can read the Delta table and perform various operations on it using PySpark.

Option A, SELECT \* FROM sales, is a SQL syntax and cannot be directly used in PySpark.

Option B, "There is no way to share data between PySpark and SQL," is incorrect. PySpark provides the capability to interact with data using both SQL and DataFrame/DataSet APIs.

Option C, spark.sql("sales"), is a valid command to execute SQL queries on registered tables in PySpark. However, in this case, the "sales" argument alone is not a valid SQL query.

Option D, spark.delta.table("sales"), is a specific method provided by Delta Lake to access Delta tables directly. While it can be used to access the "sales" table, it is not the most common approach in PySpark.

upvoted 4 times

🗨️ **ThomasReps** 1 year ago

**Selected Answer: E**

It's E. As stated by others, the default format is delta

If you try to run D, you get an error, that there are no "delta"-command for spark: "AttributeError: 'SparkSession' object has no attribute 'delta' you want to explicit tell it should be delta, then you need an ".option(format='delta')" insted.

upvoted 2 times

🗨️ **Dwarakkrishna** 1 year ago

You access data in Delta tables by the table name or the table path, as shown in the following examples:

```
people_df = spark.read.table(table_name)
```

```
display(people_df)
```

upvoted 1 times

🗨️ **prasioso** 1 year, 1 month ago

I believe the answer is E as in databricks the default tables are delta tables hence spark.table should be enough. Have not seen a spark.delta.table function before.

upvoted 1 times

🗨️ **Tickxit** 1 year, 1 month ago

**Selected Answer: E**

E: spark.table or spark.read.table

upvoted 2 times

🗨️ **softthinkers** 1 year, 1 month ago

Correct Answer is D spark.delta.table("sales") And the reason that its asking for delta table not normal table if its for normal table then it should be spark.table("sales")

upvoted 1 times

🗨️ **qium** 7 months, 1 week ago

default type type is "delta".

upvoted 1 times

🗨️ **Majjji** 1 year, 1 month ago

The correct answer is D.

The data engineering team can access the Delta table sales in PySpark by using the spark.delta.table command. This command is used to create a DataFrame based on a Delta table. Therefore, the correct command is spark.delta.table("sales").

upvoted 1 times

🗨️ **Varma\_Saraswathula** 1 year, 2 months ago

Option E -

<https://spark.apache.org/docs/3.2.1/api/python/reference/api/pyspark.sql.Session.table.html>

upvoted 1 times

## Question #13

Topic 1

Which of the following commands will return the location of database customer360?

- A. DESCRIBE LOCATION customer360;
- B. DROP DATABASE customer360;
- C. DESCRIBE DATABASE customer360;**
- D. ALTER DATABASE customer360 SET DBPROPERTIES ('location' = '/user');
- E. USE DATABASE customer360;

**Correct Answer: C**

Community vote distribution

C (100%)

🗄️ **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: C**

C is correct

upvoted 1 times

🗄️ **Itmma** 3 months ago

**Selected Answer: C**

C is correct

upvoted 1 times

🗄️ **SerGrey** 5 months, 2 weeks ago

**Selected Answer: C**

Correct answer is C

upvoted 1 times

🗄️ **awofalus** 7 months, 2 weeks ago

**Selected Answer: C**

Correct :C

upvoted 1 times

🗄️ **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: C**

C is correct

upvoted 1 times

🗄️ **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: C**

C. DESCRIBE DATABASE customer360;

To retrieve the location of a database named "customer360" in a database management system like Hive or Databricks, you can use the DESCRIBE DATABASE command followed by the database name. This command will provide information about the database, including its location.

upvoted 4 times

🗄️ **Akshay67364** 10 months, 1 week ago

Option C

upvoted 1 times

🗄️ **Gowthamr02** 10 months, 2 weeks ago

Option C

upvoted 1 times

🗄️ **Varma\_Saraswathula** 1 year, 2 months ago

Option C -

<https://spark.apache.org/docs/3.0.0-preview/sql-ref-syntax-aux-describe-database.html>

upvoted 1 times

🗄️ **surrabhi\_4** 1 year, 2 months ago

**Selected Answer: C**

option c

upvoted 2 times

🗄️ **knivesz** 1 year, 2 months ago

**Selected Answer: C**

Muy facil

upvoted 2 times

🗄️ **XiltroX** 1 year, 2 months ago

**Selected Answer: C**

Correct answer

upvoted 3 times



A data engineer wants to create a new table containing the names of customers that live in France.

They have written the following command:

```
CREATE TABLE customersInFrance
  AS
SELECT id,
       firstName,
       lastName,
FROM customerLocations
WHERE country = 'FRANCE';
```

A senior data engineer mentions that it is organization policy to include a table property indicating that the new table includes personally identifiable information (PII).


Which of the following lines of code fills in the above blank to successfully complete the task?

- A. There is no way to indicate whether a table contains PII.
- B. "COMMENT PII"
- C. TBLPROPERTIES PII
- D. COMMENT "Contains PII"**
- E. PII

**Correct Answer: C**


Community vote distribution

D (100%)

 **Huroye** Highly Voted 7 months, 1 week ago

The correct answer is D. COMMENT "Contains PII". Context matters. Yes, you can use Table Property to add additional metadata. But you cannot view that property when you describe the table. With the Comment "this is ..." anyone who describe the table <DESC <table name> w see the comment.

upvoted 7 times

 **benni\_ale** Most Recent 1 month, 3 weeks ago

**Selected Answer: D**

D is correct

upvoted 1 times

 **Itmma** 3 months ago

**Selected Answer: D**

D is correct

upvoted 1 times

 **a\_51** 3 months ago



**Selected Answer: D**

<https://docs.databricks.com/en/sql/language-manual/sql-ref-syntax-ddl-create-table-using.html>



COMMENT column\_comment

A string literal to describe the column.

upvoted 1 times



  **agAshish** 4 months, 2 weeks ago

answer C :  
CREATE TABLE new\_table  
AS  
SELECT customer\_name  
FROM original\_table  
WHERE country = 'France'  
TBLPROPERTIES ('PII'='true');  
upvoted 2 times

  **SerGrey** 5 months, 2 weeks ago



**Selected Answer: D**

Correct answer is D  
upvoted 1 times

  **awofalus** 7 months, 2 weeks ago



**Selected Answer: D**

correct :D  
upvoted 1 times

  **chris\_mach** 8 months, 3 weeks ago


**Selected Answer: D**

D is correct  
upvoted 1 times

  **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: D**

D is correct  
upvoted 1 times

  **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: D**

D. COMMENT "Contains PII"  
upvoted 1 times

  **Gems1** 10 months, 3 weeks ago

D  
Ref:<https://www.databricks.com/discover/pages/data-quality-management>  
CREATE TABLE my\_table (id INT COMMENT 'Unique Identification Number', name STRING COMMENT 'PII', age INT COMMENT 'PII')  
TBLPROPERTIES ('contains\_pii'=True)  
COMMENT 'Contains PII';  
upvoted 4 times

  **Atnafu** 11 months, 2 weeks ago

D  
The COMMENT keyword is used to add a comment to a table. The comment can be used to provide additional information about the table, such as its purpose or the data that it contains.

In this case, the data engineer wants to add a comment to the customersInFrance table indicating that the table contains PII. The following line of code will do this:

Code snippet  
COMMENT "Contains PII"  
Use code with caution. Learn more  
This will add the comment "Contains PII" to the customersInFrance table.

The other options are not valid ways to indicate that a table contains PII. The TBLPROPERTIES keyword is used to set the table properties, but there is no property for indicating whether a table contains PII. The PII keyword is not a valid keyword in SQL.

Therefore, the only valid way to indicate that a table contains PII is to use the COMMENT keyword.

upvoted 2 times

🗋️ 👤 **Virendev** 1 year, 1 month ago

**Selected Answer: D**

syntax of C is wrong.  
upvoted 2 times

🗋️ 👤 **[Removed]** 1 year ago

Exactly. The correct syntax for table properties is: TBLPROPERTIES ('foo'='bar');  
upvoted 1 times

🗋️ 👤 **softthinkers** 1 year, 1 month ago

Correct answer should be C as command creates a new table called "customersInFrance" with the properties of Personally Identifiable Information (PII) and selects the columns ID, FIRSTNAME, LASTNAME, ADDRESS, and PHONE\_NUMBER from the existing "customers" table where the country is France.  
upvoted 2 times

🗋️ 👤 **Varma\_Saraswathula** 1 year, 2 months ago

D  
<https://learn.microsoft.com/en-us/azure/databricks/sql/language-manual/sql-ref-syntax-ddl-create-table-using>  
upvoted 1 times

🗋️ 👤 **naxacod574** 1 year, 2 months ago

Option D  
upvoted 1 times

🗋️ 👤 **XiltroX** 1 year, 2 months ago

**Selected Answer: D**

Option D is the correct answer  
upvoted 1 times

## Question #15

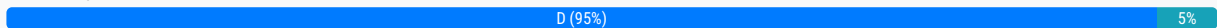
Topic 1

Which of the following benefits is provided by the array functions from Spark SQL?

- A. An ability to work with data in a variety of types at once
- B. An ability to work with data within certain partitions and windows
- C. An ability to work with time-related data in specified intervals
- D. An ability to work with complex, nested data ingested from JSON files**
- E. An ability to work with an array of tables for procedural automation

**Correct Answer: B**

Community vote distribution



🗋️ 👤 **BharaniRaj** 1 month ago

**Selected Answer: D**

D is the right answer  
upvoted 1 times

🗋️ 👤 **benni\_ale** 2 months, 2 weeks ago



**Selected Answer: D**

i thought sql arrays are usually seen in json files read  
upvoted 1 times



  **Itmma** 3 months ago

**Selected Answer: E**

E is correct  
upvoted 1 times

  **SerGrey** 5 months, 2 weeks ago

Correct answer is D  
upvoted 1 times



  **Garyn** 5 months, 3 weeks ago

**Selected Answer: D**

D. An ability to work with complex, nested data ingested from JSON files

Array functions in Spark SQL enable users to work efficiently with arrays and complex, nested data structures that are often ingested from JSON files or other nested data formats. These functions allow manipulation, querying, and extraction of elements from arrays and nested structures within the dataset, facilitating operations on complex data types within Spark SQL.

upvoted 4 times



  **Huroye** 7 months, 1 week ago

Correct answer is D. Array provides complex nesting of data and it is easy to query. That's why we use arrays for defining data domains.  
upvoted 1 times

  **awofalus** 7 months, 2 weeks ago



**Selected Answer: D**

D is correct  
upvoted 1 times

  **VijayKula** 8 months, 2 weeks ago



**Selected Answer: D**

D is the correct answer  
upvoted 1 times

  **chris\_mach** 8 months, 3 weeks ago



**Selected Answer: D**

array functions allow you to work with JSON data  
upvoted 1 times

  **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: D**

D is right ans  
upvoted 1 times

  **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: D**

D. An ability to work with complex, nested data ingested from JSON files

Array functions in Spark SQL are primarily used for working with arrays and complex, nested data structures, such as those often encountered when ingesting JSON files. These functions allow you to manipulate and query nested arrays and structures within your data, making it easier to extract and work with specific elements or values within complex data formats.

While some of the other options (such as option A for working with different data types) are features of Spark SQL or SQL in general, array functions specifically excel at handling complex, nested data structures like those found in JSON files.

upvoted 2 times

🗨️ 👤 **Atnafu** 11 months, 2 weeks ago

Array functions in Spark SQL allow you to work with complex, nested data ingested from JSON files. These functions can be used to extract data from nested structures, manipulate data within nested structures, and aggregate data within nested structures.

The other options are not benefits provided by the array functions from Spark SQL.

Option A: Array functions do not allow you to work with data in a variety of types at once.

Option B: Array functions do not allow you to work with data within certain partitions and windows.

Option C: Array functions do not allow you to work with time-related data in specified intervals.

Option E: Array functions do not allow you to work with an array of tables for procedural automation.

Therefore, the only benefit provided by the array functions from Spark SQL is the ability to work with complex, nested data ingested from JSON files.

upvoted 4 times

🗨️ 👤 **prasioso** 1 year, 1 month ago

**Selected Answer: D**

Correct answer is D. Spark SQL Array functions allow us to work with nested datasets in JSON files

upvoted 2 times

🗨️ 👤 **Varma\_Saraswathula** 1 year, 2 months ago

Option D

upvoted 1 times

🗨️ 👤 **naxacod574** 1 year, 2 months ago

Option D

upvoted 1 times

🗨️ 👤 **sdas1** 1 year, 2 months ago

option D

upvoted 1 times

🗨️ 👤 **surrabhi\_4** 1 year, 2 months ago

**Selected Answer: D**

## Question #16

Topic 1

Which of the following commands can be used to write data into a Delta table while avoiding the writing of duplicate records?

A. DROP

B. IGNORE

**C. MERGE**

D. APPEND

E. INSERT

**Correct Answer: C**

Community vote distribution



🗨️ 👤 **BharaniRaj** 1 month ago

**Selected Answer: C**

C is the right answer

upvoted 1 times

🗨️ 👤 **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: C**

C merge

upvoted 1 times

🗄️ 👤 **SerGrey** 5 months, 2 weeks ago

**Selected Answer: C**

Correct answer is C  
upvoted 1 times

🗄️ 👤 **awofalus** 7 months, 2 weeks ago

**Selected Answer: C**

C is correct  
upvoted 1 times

🗄️ 👤 **J\_1\_2** 8 months ago

**Selected Answer: C**

Merge is correct  
upvoted 1 times

🗄️ 👤 **DavidRou** 8 months, 2 weeks ago

MERGE INTO is the one to choose if you want to avoid duplicates.  
upvoted 2 times

🗄️ 👤 **chris\_mach** 8 months, 3 weeks ago

**Selected Answer: C**

Merge is correct  
upvoted 1 times

🗄️ 👤 **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: C**

Merge will avoid duplicates by comparing the results based on primary key columns  
upvoted 1 times

🗄️ 👤 **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: C**

C. MERGE

The MERGE command is used to write data into a Delta table while avoiding the writing of duplicate records. It allows you to perform an "ups" operation, which means that it will insert new records and update existing records in the Delta table based on a specified condition. This helps maintain data integrity and avoid duplicates when adding new data to the table.

upvoted 3 times

🗄️ 👤 **Atnafu** 11 months, 2 weeks ago

C. MERGE

To write data into a Delta table while avoiding the writing of duplicate records, you can use the MERGE command. The MERGE command in Delta Lake allows you to combine the ability to insert new records and update existing records in a single atomic operation.

The MERGE command compares the data being written with the existing data in the Delta table based on specified matching criteria, typically using a primary key or unique identifier. It then performs conditional actions, such as inserting new records or updating existing records, depending on the comparison results.

By using the MERGE command, you can handle the prevention of duplicate records in a more controlled and efficient manner. It allows you to synchronize and reconcile data from different sources while avoiding duplication and ensuring data integrity.

Therefore, option C, MERGE, is the correct command to use when writing data into a Delta table while avoiding the writing of duplicate records.


upvoted 2 times

🗄️ 👤 **softthinkers** 1 year, 1 month ago

Answer is C. AS DROP is used to remove a table or database  
IGNORE is used to skip errors while executing a query.  
INSERT will add new records but will not avoid duplication so Merge is right answer  
upvoted 2 times

🗄️ 👤 **Varma\_Saraswathula** 1 year, 2 months ago

Ans - C  
<https://docs.databricks.com/sql/language-manual/delta-merge-into.html>  
upvoted 2 times

 **naxacod574** 1 year, 2 months ago

Option C


upvoted 1 times

 **XiltroX** 1 year, 2 months ago

**Selected Answer: D**

Wrong answer. The correct answer is D.

upvoted 1 times

 **Oleskie** 1 year, 2 months ago

'C' is a correct answer. <https://docs.databricks.com/sql/language-manual/delta-merge-into.html>

upvoted 4 times

 **XiltroX** 1 year, 2 months ago

Thanks for the clarification


upvoted 1 times

 **knivesz** 1 year, 2 months ago

**Selected Answer: C**

la unica opcion posible

upvoted 3 times

 **knivesz** 1 year, 2 months ago

Respuesta correcta C

A) DROP: Elimina registros, B) IGNORE : NO existe C) MERGE: EN base a la data, registra, actualiza o elimina registros, D) NO existe E) Sc inserta

## Question #17

Topic 1

A data engineer needs to apply custom logic to string column city in table stores for a specific use case. In order to apply this custom logic at scale, the data engineer wants to create a SQL user-defined function (UDF).

Which of the following code blocks creates this SQL UDF?

- A. 

```
CREATE FUNCTION combine_nyc(city STRING)
  RETURNS STRING
  RETURN CASE
    WHEN city = "brooklyn" THEN "new york"
    ELSE city
  END;
```
- B. 

```
CREATE UDF combine_nyc(city STRING)
  RETURNS STRING
  CASE
    WHEN city = "brooklyn" THEN "new york"
    ELSE city
  END;
```
- C. 

```
CREATE UDF combine_nyc(city STRING)
  RETURN CASE
    WHEN city = "brooklyn" THEN "new york"
    ELSE city
  END;
```
- D. 

```
CREATE FUNCTION combine_nyc(city STRING)
  RETURN CASE
    WHEN city = "brooklyn" THEN "new york"
    ELSE city
  END;
```
- E. 

```
CREATE UDF combine_nyc(city STRING)
  RETURNS STRING
  RETURN CASE
    WHEN city = "brooklyn" THEN "new york"
    ELSE city
  END;
```

**Correct Answer: E**

Community vote distribution

A (100%)

- Flexron **Highly Voted** 1 year, 2 months ago  
E is wrong, the right answer is A.  
<https://www.databricks.com/blog/2021/10/20/introducing-sql-user-defined-functions.html>  
upvoted 10 times
- XiltroX **Highly Voted** 1 year, 2 months ago  
**Selected Answer: A**  
The answer E is incorrect. A user defined function is never written as CREATE UDF. The correct way is CREATE FUNCTION. So that leaves us with the choices A and D. Out of that, in D, there is no such thing as RETURN CASE so the correct answer is A.  
upvoted 7 times
- Redwings538 1 year, 1 month ago  
Both A and D use RETURN CASE  
upvoted 4 times
- benni\_ale **Most Recent** 1 month, 3 weeks ago  
**Selected Answer: A**  
A is correct  
upvoted 1 times
- a\_51 3 months ago  
**Selected Answer: A**  
<https://docs.databricks.com/en/udf/index.html#language-sql>  
Answer E is not correct when having CREATE UDF rather than CREATE FUNCTION  
upvoted 1 times
- isamrat28 4 months, 1 week ago  
Correct Answer is A  
upvoted 1 times
- agAshish 4 months, 2 weeks ago  
D, should be the answer. The function is not returning STRING, it is applied on a string column  
upvoted 2 times
- SerGrey 5 months, 2 weeks ago  
**Selected Answer: A**  
Correct answer is A  
upvoted 1 times
- Huroye 7 months, 1 week ago  
Correct answer is A. It is not E. First, you do not need to specify "UDF" in the syntax. You need to specify the return type and then what you want to return, usually your processed output.  
CREATE FUNCTION myUDF(udf STRING)  
RETURNS STRING  
<...udf....>  
upvoted 3 times
- VijayKula 8 months, 1 week ago  
**Selected Answer: A**  
Create Function Return String  
upvoted 1 times
- VijayKula 8 months, 2 weeks ago  
**Selected Answer: A**  
Correct is A  
upvoted 1 times
- DavidRou 8 months, 2 weeks ago  
A is the right answer.  
The template to use is the following:  
CREATE FUNCTION <name\_function> (<function\_parameter>, ..) RETURNS <return\_type>  
<body>  
upvoted 2 times



- 🗨️ **AtanuChat** 8 months, 3 weeks ago  
First need create a function and then need to register it as UDF  
upvoted 1 times
- 🗨️ **KalavathiP** 8 months, 3 weeks ago  
**Selected Answer: A**  
A is correct  
upvoted 1 times
- 🗨️ **d\_b47** 8 months, 4 weeks ago  
**Selected Answer: A**  
no UDF only FUNCTION needed.  
upvoted 1 times
- 🗨️ **akk\_1289** 11 months ago  
CORRECT ANSWER IS : A  
<https://www.databricks.com/blog/2021/10/20/introducing-sql-user-defined-functions.html>  
upvoted 1 times
- 🗨️ **Atnafu** 11 months, 2 weeks ago  
A  
-- Create the UDF  
CREATE FUNCTION custom\_logic(city STRING) RETURNS STRING  
BEGIN  
-- Custom logic goes here  
-- Example: Return the uppercase version of the city  
RETURN UPPER(city);  
END;  
upvoted 1 times
- 🗨️ **Atnafu** 11 months, 2 weeks ago  
from datetime import datetime  
current\_day = datetime.now().strftime('%A')  
if current\_day == 'Sunday':  
# Run the final query  
spark.sql("SELECT ... FROM ... WHERE ...")  
else:  
# Handle other scenarios or skip the final query  
pass  
upvoted 1 times
- 🗨️ **malani** 1 year, 1 month ago  
D also works  
upvoted 1 times

## Question #18

Topic 1

A data analyst has a series of queries in a SQL program. The data analyst wants this program to run every day. They only want the final query in the program to run on Sundays. They ask for help from the data engineering team to complete this task.

Which of the following approaches could be used by the data engineering team to complete this task?

A. They could submit a feature request with Databricks to add this functionality.

**B. They could wrap the queries using PySpark and use Python's control flow system to determine when to run the final query**

C. They could only run the entire program on Sundays.

D. They could automatically restrict access to the source table in the final query so that it is only accessible on Sundays.

E. They could redesign the data model to separate the data used in the final query into a new table.

**Correct Answer: B**

Community vote distribution

B (100%)

  **Atnafu**  11 months, 2 weeks ago

B

The answer is B.

Option A: Submitting a feature request with Databricks to add this functionality is not a feasible solution because it would require Databricks to implement new functionality.

Option C: Only running the entire program on Sundays would be inconvenient for the data analyst because they would have to remember to run the program on Sundays.


Option D: Automatically restricting access to the source table in the final query so that it is only accessible on Sundays would be difficult to implement and would not be a reliable solution.

Option E: Redesigning the data model to separate the data used in the final query into a new table would be a major undertaking and would not be a feasible solution for this specific problem.

Therefore, the only feasible solution to the problem is to wrap the queries using PySpark and use Python's control flow system to determine when to run the final query.

python code

upvoted 9 times

  **vibha2119** 2 weeks, 5 days ago

Also to add for option C is:

the requirement says: data analysts want to run the sql program every day, but only the final query to run on Sundays. So the option C violates the requirements


upvoted 1 times

  **SerGrey**  5 months, 2 weeks ago

**Selected Answer: B**

Correct answer is B


upvoted 1 times

  **VijayKula** 8 months, 2 weeks ago

**Selected Answer: B**

B is correct

upvoted 1 times

  **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: B**

B is correct ans

upvoted 1 times

  **d\_b47** 8 months, 4 weeks ago

**Selected Answer: B**

B is correct.

upvoted 1 times

🗨️ **mehroosali** 11 months, 2 weeks ago

**Selected Answer: B**

B is correct  
upvoted 1 times

🗨️ **[Removed]** 1 year ago

**Selected Answer: B**

B is correct  
upvoted 1 times

🗨️ **XiltroX** 1 year, 2 months ago

**Selected Answer: B**

B is the correct answer  
upvoted 1 times

🗨️ **mimzzz** 1 year, 2 months ago

i think the answer is correct  
upvoted 1 times

🗨️ **knivesz** 1 year, 2 months ago

**Selected Answer: B**

Respuesta Correcta es B  
upvoted 1 times

## Question #19

Topic 1

A data engineer runs a statement every day to copy the previous day's sales into the table transactions. Each day's sales are in their own file in the location "/transactions/raw".

Today, the data engineer runs the following command to complete this task:

```
COPY INTO transactions
FROM "/transactions/raw"
FILEFORMAT = PARQUET;
```

After running the command today, the data engineer notices that the number of records in table transactions has not changed.

Which of the following describes why the statement might not have copied any new records into the table?

- A. The format of the files to be copied were not included with the FORMAT\_OPTIONS keyword.
- B. The names of the files to be copied were not included with the FILES keyword.
- C. The previous day's file has already been copied into the table**
- D. The PARQUET file format does not support COPY INTO.
- E. The COPY INTO statement requires the table to be refreshed to view the copied rows.

**Correct Answer: C**

Community vote distribution



🗨️ **Nika12** 4 months, 3 weeks ago

**Selected Answer: C**

Just got 100% on the test. C was correct.  
upvoted 2 times

🗄️ 👤 **SerGrey** 5 months, 2 weeks ago

**Selected Answer: C**

Correct answer is C  
upvoted 1 times

🗄️ 👤 **Garyn** 5 months, 3 weeks ago

**Selected Answer: C**

C. The previous day's file has already been copied into the table.

The COPY INTO statement is generally used to copy data from files or a location into a table. If the data engineer runs this statement daily to copy the previous day's sales into the "transactions" table and the number of records hasn't changed after today's execution, it's possible that the data from today's file might not have differed from the data already present in the table.

If the files in the "/transactions/raw" location are expected to contain distinct data for each day and the number of records in the table remain the same, it implies that the data engineer might have already copied today's data previously, or today's data was identical to the data already present in the table.

Options A, B, D, and E don't accurately explain why the statement might not have copied new records into the table based on the provided scenario.

upvoted 1 times

🗄️ 👤 **awofalus** 7 months, 2 weeks ago

**Selected Answer: C**

C is correct  
upvoted 1 times

🗄️ 👤 **kishanu** 8 months, 1 week ago

If the table "transaction" is an external table, then option E, if its internal C should suffice.

upvoted 1 times

🗄️ 👤 **DavidRou** 8 months, 2 weeks ago

**Selected Answer: C**

COPY INTO statement does skip already copied rows.  
upvoted 1 times

🗄️ 👤 **KalavathiP** 8 months, 3 weeks ago

**Selected Answer: C**

C is correct ans  
upvoted 1 times

🗄️ 👤 **ezeik** 8 months, 4 weeks ago

**Selected Answer: E**

E is the correct answer, because immediately after using copy into you might query the cached version of the table.  
upvoted 3 times

🗄️ 👤 **AndreFR** 10 months ago

**Selected Answer: C**

<https://docs.databricks.com/en/ingestion/copy-into/index.html>

The COPY INTO SQL command lets you load data from a file location into a Delta table. This is a re-triable and idempotent operation; files in the source location that have already been loaded are skipped.

if there are no new records, the only consistent choice is C no new files were loaded because already loaded files were skipped.

upvoted 1 times

🗄️ 👤 **Atnafu** 11 months, 2 weeks ago

C

The COPY INTO statement copies the data from the specified files into the target table. If the previous day's file has already been copied into the table, then the COPY INTO statement will not copy any new records into the table.

upvoted 1 times

junction 1 year ago

**Selected Answer: C**

COPY INTO

Loads data from a file location into a Delta table. This is a retrievable and idempotent operation—files in the source location that have already been loaded are skipped.

upvoted 1 times

testdb 1 year, 1 month ago

**Selected Answer: B**

Answer: B

FILES = ('f1.json', 'f2.json', 'f3.json', 'f4.json', 'f5.json')

<https://docs.databricks.com/ingestion/copy-into/examples.html>

upvoted 1 times

[Removed] 1 year ago

The correct answer is letter C.

The use of specific file names with keyword "FILES" is optional as the syntax of COPY INTO declares:

[ FILES = ( file\_name [, ...] ) | PATTERN = glob\_pattern ]

When keyword FILES is not used in the statement all files of the directory is used once (because this operation is idempotent).

upvoted 2 times

Varma\_Saraswathula 1 year, 2 months ago

C-

<https://docs.databricks.com/ingestion/copy-into/tutorial-notebook.html>

Because this action is idempotent, you can run it multiple times but data will only be loaded once.

upvoted 1 times

XiltroX 1 year, 2 months ago

**Selected Answer: C**

Option C is the correct answer.

upvoted 3 times

mimzzz 1 year, 2 months ago

i am not sure whether C is the correct answer, but A is definitely not right

upvoted 1 times

sdas1 1 year, 2 months ago

option C

upvoted 1 times

knivesz 1 year, 2 months ago

**Selected Answer: C**

Respuesta C, por descarte, A) No es necesario B) No se coloca FILES D) PARQUET si es soportado E) No es necesario refrescar la vista, ya que se esta copiando un archivo

upvoted 2 times

## Question #20

Topic 1

A data engineer needs to create a table in Databricks using data from their organization's existing SQLite database.

They run the following command:

```
CREATE TABLE jdbc_customer360
USING _____
OPTIONS (
  url "jdbc:sqlite:/customers.db",
  dbtable "customer360"
)
```

Which of the following lines of code fills in the above blank to successfully complete the task?

**A. org.apache.spark.sql.jdbc**

B. autoloader

C. DELTA

D. sqlite

E. org.apache.spark.sql.sqlite

**Correct Answer: E**

Community vote distribution

A (100%)

🗳️ **rafahb** Highly Voted 1 year, 2 months ago  
A is correct  
upvoted 8 times

🗳️ **benni\_ale** Most Recent 1 month, 3 weeks ago  
Selected Answer: A  
A is correct  
upvoted 1 times

🗳️ **SerGrey** 5 months, 2 weeks ago  
Correct answer is A  
upvoted 1 times

🗳️ **Huroye** 7 months, 1 week ago  
I think the correct answer is A. All that is missing the the jdbc drive. org.apache.spark.sql.jdbc  
upvoted 2 times

🗳️ **chris\_mach** 8 months, 3 weeks ago  
Selected Answer: A  
A is correct  
upvoted 1 times

🗳️ **KalavathiP** 8 months, 3 weeks ago  
Selected Answer: A  
A is correct  
upvoted 1 times

🗳️ **juliom6** 1 year ago  
Selected Answer: A  
must be "USING JDBC", there is no such thing as "USING org.apache.spark.sql.jdbc". <https://docs.databricks.com/external-data/jdbc.html#language-sql>  
upvoted 1 times

🗳️ **juliom6** 1 year ago  
I correct myself <https://docs.yugabyte.com/preview/integrations/apache-spark/spark-sql/>  
upvoted 4 times

🗳️ **Majiji** 1 year, 1 month ago  
Selected Answer: A  
To specify the JDBC driver and other options, the using clause should be followed by the fully qualified name of the JDBC data source, which org.apache.spark.sql.jdbc.  
upvoted 2 times

🗳️ **Varma\_Saraswathula** 1 year, 2 months ago  
Answer A -  
CREATE TABLE new\_employees\_table  
USING JDBC  
OPTIONS (  
url "<jdbc\_url>",  
dbtable "<table\_name>",  
user '<username>',  
password '<password>'  
) AS  
SELECT \* FROM employees\_table\_vw  
upvoted 1 times

- 🗨️ **naxacod574** 1 year, 2 months ago  
JDBC - Option A  
upvoted 1 times
- 🗨️ **XiltroX** 1 year, 2 months ago  
**Selected Answer: A**  
Option A is correct answer  
upvoted 2 times
- 🗨️ **sdas1** 1 year, 2 months ago  
option A  
upvoted 2 times
- 🗨️ **surrabhi\_4** 1 year, 2 months ago  
**Selected Answer: A**  
option A  
upvoted 1 times
- 🗨️ **knivesz** 1 year, 2 months ago  
**Selected Answer: A**  
Es JDBC osea la A, pregunta con truco para confundir  
upvoted 1 times
- 🗨️ **knivesz** 1 year, 2 months ago  
es JDBC  
upvoted 3 times

#### Question #21

Topic 1

A data engineering team has two tables. The first table `march_transactions` is a collection of all retail transactions in the month of March. The second table `april_transactions` is a collection of all retail transactions in the month of April. There are no duplicate records between the tables. Which of the following commands should be run to create a new table `all_transactions` that contains all records from `march_transactions` and `april_transactions` without duplicate records?


- A. `CREATE TABLE all_transactions AS  
SELECT * FROM march_transactions  
INNER JOIN SELECT * FROM april_transactions;`
- B. `CREATE TABLE all_transactions AS  
SELECT * FROM march_transactions  
UNION SELECT * FROM april_transactions;`**
- C. `CREATE TABLE all_transactions AS  
SELECT * FROM march_transactions  
OUTER JOIN SELECT * FROM april_transactions;`
- D. `CREATE TABLE all_transactions AS  
SELECT * FROM march_transactions  
INTERSECT SELECT * from april_transactions;`
- E. `CREATE TABLE all_transactions AS`

SELECT \* FROM march\_transactions  
MERGE SELECT \* FROM april\_transactions;

**Correct Answer: B**


Community vote distribution

B (100%)

 **SerGrey** 5 months, 2 weeks ago

**Selected Answer: B**

B is correct  
upvoted 2 times

 **awofalus** 7 months, 2 weeks ago

**Selected Answer: B**

Correct: B  
upvoted 1 times

 **ezeik** 9 months ago

UNION [ALL | DISTINCT]

Returns the result of subquery1 plus the rows of subquery2`.

If ALL is specified duplicate rows are preserved.

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

<https://docs.databricks.com/en/sql/language-manual/sql-ref-syntax-qry-select-setops.html#examples>

upvoted 3 times

 **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: B**

B. CREATE TABLE all\_transactions AS  
SELECT \* FROM march\_transactions  
UNION SELECT \* FROM april\_transactions;

To create a new table all\_transactions that contains all records from march\_transactions and april\_transactions without duplicate records, you should use the UNION operator, as shown in option B. This operator combines the result sets of the two tables while automatically removing duplicate records.

upvoted 1 times

 **Atnafu** 11 months, 2 weeks ago

B  
CREATE TABLE all\_transactions AS  
SELECT \* FROM march\_transactions  
UNION  
SELECT \* FROM april\_transactions;

upvoted 1 times

 **prasioso** 1 year, 1 month ago

**Selected Answer: B**

Answer is B.  
upvoted 1 times

 **surrabhi\_4** 1 year, 2 months ago

**Selected Answer: B**

option B  
upvoted 1 times

 **XiltroX** 1 year, 2 months ago

**Selected Answer: B**

Answer is correct  
upvoted 2 times



A data engineer only wants to execute the final block of a Python program if the Python variable `day_of_week` is equal to 1 and the Python variable `review_period` is `True`.

Which of the following control flow statements should the data engineer use to begin this conditionally executed code block?


- A. `if day_of_week = 1 and review_period:`
- B. `if day_of_week = 1 and review_period = "True":`
- C. `if day_of_week == 1 and review_period == "True":`
- D. `if day_of_week == 1 and review_period:`**
- E. `if day_of_week = 1 & review_period: = "True":`

**Correct Answer: C**

Community vote distribution

D (93%)


7%

 **4be8126** Highly Voted 1 year, 2 months ago

The correct control flow statement to begin the conditionally executed code block would be D. `if day_of_week == 1 and review_period:`.

This statement will check if the variable `day_of_week` is equal to 1 and if the variable `review_period` evaluates to a truthy value. The use of the double equal sign (`==`) in the comparison of `day_of_week` is important, as a single equal sign (`=`) would be used to assign a value to the variable instead of checking its value. The use of a single ampersand (`&`) instead of the keyword `and` is not valid syntax in Python. The use of quotes around `True` in options B and C will result in a string comparison, which will not evaluate to `True` even if the value of `review_period` is `True`.


upvoted 15 times

 **Mircuz** Most Recent 3 months, 2 weeks ago

**Selected Answer: D**

C fits if you're looking for a string `== 'True'`, in this case you are using a boolean so D

upvoted 2 times

 **SerGrey** 5 months, 2 weeks ago

**Selected Answer: D**

D is correct

upvoted 1 times

 **Garyn** 5 months, 3 weeks ago

**Selected Answer: D**

D. `if day_of_week == 1 and review_period:`

- In Python, the equality comparison operator is `==`, not `=`. `==` is used to check if two values are equal.
- The logical operator "and" is used to combine two conditions, ensuring that both conditions (`day_of_week == 1` and `review_period`) are true the subsequent code block to execute.
- `day_of_week == 1` checks if the variable `day_of_week` is equal to the integer value 1.
- `review_period` is already assumed to be a Boolean variable since it is stated to be `True` (without quotes) in the question. Therefore, it should be compared to a string `"True"`.

Therefore, option D correctly represents the condition for executing the final block of the Python program based on the given conditions.

upvoted 1 times

 **awofalus** 7 months, 2 weeks ago

**Selected Answer: D**

D is correct

upvoted 1 times

 **VijayKula** 8 months, 1 week ago

**Selected Answer: D**

`review_period == "true"` is different from `review_period == True`

upvoted 1 times

🗨️ **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: D**

D. if day\_of\_week == 1 and review\_period:

The correct control flow statement to begin the conditionally executed code block is option D. In Python, the == operator is used for equality comparison, and and is used for logical "and" operations. So, this statement checks if day\_of\_week is equal to 1 and review\_period is True (a boolean value), which is the correct way to express the conditions you mentioned.

upvoted 1 times

🗨️ **[Removed]** 9 months, 3 weeks ago

**Selected Answer: D**

Answer is D

upvoted 1 times

🗨️ **Atnafu** 11 months, 2 weeks ago

D

if day\_of\_week == 1 and review\_period:

upvoted 1 times

🗨️ **prasioso** 1 year, 1 month ago

**Selected Answer: D**

in python value comparison is done by double equal signs (==). in case of boolean values that are TRUE these may be omitted. Quotes around True would result in string comparison and here we are comparing to a bool value.

upvoted 2 times

🗨️ **Bob123456** 1 year, 1 month ago

Answer is 'D'

day\_of\_week=1

review\_period = True

1)

if day\_of\_week == 1 and review\_period:

print("yes")

output:

Above code block's output is yes

2)

if day\_of\_week == 1 and review\_period == "True":

print("yes")

output:

There is no output for above code block

upvoted 1 times

🗨️ **Majjji** 1 year, 1 month ago

**Selected Answer: D**

The data engineer should use option D: if day\_of\_week == 1 and review\_period:. This statement checks if the variable day\_of\_week is equal to 1 and if the variable review\_period is True. It uses the double equal sign (==) to compare the values of the variables, and does not use quotes around the keyword True, which is a boolean value.

upvoted 1 times

🗨️ **surrabhi\_4** 1 year, 2 months ago

**Selected Answer: D**

option D

upvoted 2 times

🗨️ **XiltroX** 1 year, 2 months ago

**Selected Answer: C**

I believe the right answer is C

upvoted 1 times

🗨️ **Igkofficialwork** 1 year ago

It's not C. Conditional check of "True" is treated as a string and not Boolean. Hence D is the right answer

upvoted 3 times

## Question #23

Topic 1

A data engineer is attempting to drop a Spark SQL table `my_table`. The data engineer wants to delete all table metadata and data. They run the following command:

```
DROP TABLE IF EXISTS my_table -
```

While the object no longer appears when they run `SHOW TABLES`, the data files still exist.


Which of the following describes why the data files still exist and the metadata files were deleted?

- A. The table's data was larger than 10 GB
- B. The table's data was smaller than 10 GB
- C. The table was external**
- D. The table did not have a location
- E. The table was managed

**Correct Answer:** C

*Community vote distribution*

C (100%)

 **SerGrey** 5 months, 2 weeks ago

**Selected Answer: C**


C is correct

upvoted 1 times

 **hemanthgvs** 8 months ago

THE QUESTION SHOULD BE "Which of the following describes why the metadata files still exist and the data files were deleted?"

upvoted 1 times

 **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: C**

C. The table was external

The reason why the data files still exist while the metadata files were deleted is because the table was external. When a table is external in Spark SQL (or in other database systems), it means that the table metadata (such as schema information and table structure) is managed externally, and Spark SQL assumes that the data is managed and maintained outside of the system. Therefore, when you execute a DROP TABLE statement for an external table, it removes only the table metadata from the catalog, leaving the data files intact.

On the other hand, for managed tables (option E), Spark SQL manages both the metadata and the data files. When you drop a managed table, it deletes both the metadata and the associated data files, resulting in a complete removal of the table.

upvoted 2 times

 **surrabhi\_4** 1 year, 2 months ago

**Selected Answer: C**

Option C

upvoted 2 times

## Question #24

Topic 1

A data engineer wants to create a data entity from a couple of tables. The data entity must be used by other data engineers in other sessions. It also must be saved to a physical location.

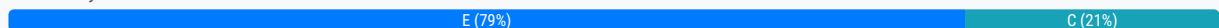
Which of the following data entities should the data engineer create?


- A. Database
- B. Function
- C. View
- D. Temporary view

**E. Table**

**Correct Answer: C**

Community vote distribution



 **Bob123456** **Highly Voted** 1 year, 1 month ago

Questions says :

1. The data entity must be used by other data engineers in other sessions.
2. It also must be saved to a physical location.

Here View doesn't store data in physical location , from the options only table stores data in physical location

So answer should be 'E' which is Table.

upvoted 28 times

🗨️ **Nika12** Highly Voted 4 months, 3 weeks ago

**Selected Answer: E**

Just got 100% in the exam. Table was a correct answer.  
upvoted 7 times

🗨️ **ADVIT** 4 months, 3 weeks ago

Wow! Congratz!  
upvoted 1 times

🗨️ **benni\_ale** Most Recent 1 month, 3 weeks ago

**Selected Answer: E**

physical location means table  
upvoted 2 times

🗨️ **agAshish** 4 months, 2 weeks ago

E. as view doesnt has any location  
upvoted 1 times

🗨️ **SerGrey** 5 months, 2 weeks ago

**Selected Answer: E**

E is correct  
upvoted 1 times

🗨️ **Garyn** 5 months, 3 weeks ago

**Selected Answer: E**

E. Table

Usage by Other Sessions: Tables in a database are persistent data structures that can be accessed by multiple users and sessions concurrently.

Saved to a Physical Location: Tables store data physically in a structured manner on disk or in a storage system, making them suitable for long-term storage.

Usage by Other Data Engineers: Other data engineers can query, access, and work with the data within the table, making it a feasible choice for shared access among multiple users or sessions.

While other entities like views or temporary views can provide different ways to represent or filter data, a table fits the criteria best when the data engineer requires a persistent physical storage entity accessible by other sessions and users for data manipulation, retrieval, and storage.

upvoted 1 times

🗨️ **RafaelCFC** 6 months, 1 week ago

**Selected Answer: C**

I think the key to the answer is that it refers to the Data Entity, and not to the data itself, when it mentions "the Data Entity must be used by other Data Engineers", and "It must be saved to a physical location". From this PoV, both C and E would be correct, however, creating a new table would incur in processing to a static state the relationship from "a couple of tables". While this makes sense to many use cases, this would require either a Workflow or a DLT to make it work, which goes over the requested scope. C is the best answer for the requested scenario.

upvoted 3 times

🗨️ **Vikram1710** 6 months, 3 weeks ago

Key point to remember during answering this question:  
"It also must be saved to a physical location"

So answer should be 'E' which is Table.

upvoted 2 times

🗨️ **rbeeraka** 6 months, 3 weeks ago

C is the right answer. View is a data entity and its definition is physically saved so other users can consume view  
upvoted 1 times

🗨️ **Huroye** 7 months, 1 week ago

The correct answer is E because it has to be physically saved. View is in memory.  
upvoted 1 times

awofalus 7 months, 2 weeks ago

Selected Answer: E

Correct : E

upvoted 1 times

vctrhugo 9 months, 2 weeks ago

Selected Answer: E

E. Table

To create a data entity that can be used by other data engineers in other sessions and must be saved to a physical location, you should create a table. Tables in a database are physical storage structures that hold data, and they can be accessed and shared by multiple users and sessions. By creating a table, you provide a permanent and structured storage location for the data entity that can be used across different sessions and by other users as needed.

Options like databases (A) can be used to organize tables, views (C) can provide virtual representations of data, and temporary views (D) are temporary in nature and don't save data to a physical location. Functions (B) are typically used for processing data or performing calculations not for storing data.

upvoted 2 times

[Removed] 9 months, 3 weeks ago

Selected Answer: E

View does not have a physical location so answer has to be E

upvoted 1 times

Kartz130789 10 months, 2 weeks ago

Selected Answer: E

View Doesn't physical location

upvoted 2 times

ehsanmor18 11 months, 1 week ago

The answer is E: "Table"

In the context described, creating a "Table" is the most suitable choice. Tables in SQL are data entities that exist independently of any session and are saved in a physical location. They can be accessed and manipulated by other data engineers in different sessions, which aligns with the requirements stated.

A "Database" is a collection of tables, views, and other database objects. A "Function" is a stored procedure that performs an operation. A "View" is a virtual table based on the result-set of an SQL statement, but it is not stored physically. A "Temporary view" is a feature that allows you to store the result of a query as a view that disappears once your session with the database is closed.

upvoted 2 times

keksssd 11 months, 1 week ago

Selected Answer: E

answer e

upvoted 1 times

Atnafu 11 months, 2 weeks ago

C

A view is a virtual table that is created from a query on one or more tables. Views are stored in the database and can be used by other data engineers in other sessions.

The other options are not correct.

Option A: A database is a collection of tables.

Option B: A function is a named block of code that can be executed.

Option D: A temporary view is a view that is only stored in memory and is not saved to a physical location.

Option E: A table is a physical collection of data.

upvoted 1 times

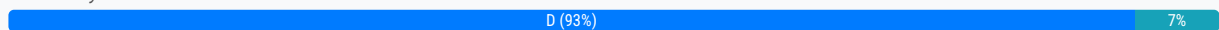
A data engineer is maintaining a data pipeline. Upon data ingestion, the data engineer notices that the source data is starting to have a lower level of quality. The data engineer would like to automate the process of monitoring the quality level.

Which of the following tools can the data engineer use to solve this problem?

- A. Unity Catalog
- B. Data Explorer
- C. Delta Lake
- D. Delta Live Tables**
- E. Auto Loader

**Correct Answer: C**

Community vote distribution



**XiltroX** **Highly Voted** 1 year, 2 months ago

**Selected Answer: D**

The answer is incorrect. The correct answer is Delta Live Tables or (C)  
<https://docs.databricks.com/delta-live-tables/expectations.html>  
upvoted 15 times

**mimzzz** 1 year ago  
upon reading this i think you are right  
upvoted 1 times

**benni\_ale** **Most Recent** 1 month, 3 weeks ago

**Selected Answer: D**

delta live table  
upvoted 1 times

**SerGrey** 5 months, 2 weeks ago

**Selected Answer: D**

Correct is D  
upvoted 1 times

**awofalus** 7 months, 2 weeks ago

**Selected Answer: D**

Correct: D  
upvoted 1 times

**awofalus** 7 months, 2 weeks ago

**Selected Answer: D**


D is correct  
upvoted 1 times

**DQCR** 9 months, 2 weeks ago

**Selected Answer: D**

Delta Live Tables is a declarative framework for building reliable, maintainable, and testable data processing pipelines. You define the transformations to perform on your data and Delta Live Tables manages task orchestration, cluster management, monitoring, data quality, and error handling.

Quality is explicitly mentioned in the definition.  
upvoted 3 times

 **vctrhugo** 9 months, 2 weeks ago


**Selected Answer: D**

D. Delta Live Tables

Delta Live Tables is a tool provided by Databricks that can help data engineers automate the monitoring of data quality. It is designed for managing data pipelines, monitoring data quality, and automating workflows. With Delta Live Tables, you can set up data quality checks and alerts to detect issues and anomalies in your data as it is ingested and processed in real-time. It provides a way to ensure that the data quality meets your desired standards and can trigger actions or notifications when issues are detected.

While the other tools mentioned may have their own purposes in a data engineering environment, Delta Live Tables is specifically designed for data quality monitoring and automation within the Databricks ecosystem.

upvoted 3 times

 **Atnafu** 11 months, 2 weeks ago

D

Delta Live Tables.


Delta Live Tables is a tool that can be used to automate the process of monitoring the quality level of data in a data pipeline. Delta Live Tables provides a number of features that can be used to monitor data quality, including:

**Data lineage:** Delta Live Tables tracks the lineage of data as it flows through the data pipeline. This allows the data engineer to see where the data came from and how it has been transformed.

**Data quality checks:** Delta Live Tables allows the data engineer to define data quality checks that can be run on the data as it is ingested. The checks can be used to identify data that is not meeting the expected quality standards.

**Alerts:** Delta Live Tables can be configured to send alerts when data quality checks fail. This allows the data engineer to be notified of potential problems with the data pipeline.

upvoted 1 times

 **Majiji** 1 year, 1 month ago

**Selected Answer: B**

The data engineer can use the Data Explorer tool to monitor the quality level of the ingested data. Data Explorer is a feature of Databricks that provides data profiling and data quality metrics to monitor the health of data pipelines.

upvoted 1 times

 **Majiji** 1 year, 1 month ago

After reading docs and more investigation I think in the terms of managing the data quality D would be better answer

upvoted 3 times

 **4be8126** 1 year, 2 months ago

**Selected Answer: B**

B. Data Explorer can be used to monitor the quality level of data. It provides an interactive interface to analyze the data and define quality rules to identify issues. Data Explorer also offers automated validation rules that can be used to monitor data quality over time.

upvoted 1 times

## Question #26

Topic 1

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 Production mode using the Continuous 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 at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing.

B. All datasets will be updated once and the pipeline will persist without any processing. The compute resources will persist but go unused.

**C. 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.**

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




E. All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.

**Correct Answer: E**

Community vote distribution



C (100%)

  **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: C**

daje gianluca

upvoted 1 times

  **SerGrey** 5 months, 2 weeks ago

**Selected Answer: C**

Correct is C

upvoted 2 times

  **Garyn** 5 months, 3 weeks ago

**Selected Answer: C**

C. 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.

Explanation:

Continuous Pipeline Mode in Production mode implies that the pipeline continuously processes incoming data updates at set intervals, ensuring the datasets are kept up-to-date as new data arrives.

Since the pipeline is set to Continuous Pipeline Mode, it will keep running and updating the datasets until it is manually shut down.

The compute resources are allocated dynamically to process and update the datasets as needed, and they will be terminated when the pipeline is stopped or shut down.

This mode allows for real-time or near-real-time updates to the datasets from the streaming/live tables, ensuring that the data remains current and reflects the changes occurring in the data sources.

upvoted 4 times

  **awofalus** 7 months, 2 weeks ago

**Selected Answer: C**

Correct : C

upvoted 1 times

  **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: C**

C. 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.

In the scenario described:

The Delta Live Table pipeline is configured in Production mode, which means it will continuously process data using the Continuous Pipeline Mode.

There are both STREAMING LIVE TABLE datasets and LIVE TABLE datasets defined.

When you click Start to update the pipeline in Continuous Pipeline Mode:

All datasets, including both STREAMING LIVE TABLE and LIVE TABLE datasets, will be updated at set intervals.

Compute resources will be deployed for the update, ensuring that the pipeline processes data.

The compute resources will be terminated when the pipeline is stopped or shut down.

This setup allows for continuous data processing while efficiently managing compute resources, and the pipeline can be stopped when no longer needed.

upvoted 4 times

  **Sandy\_17** 10 months ago

**Selected Answer: C**

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.

upvoted 1 times

- 🗳️ 👤 **say88** 10 months, 3 weeks ago  
No answer is correct. Prod Continuous mode processes data at set intervals until pipe is shutdown. However, compute must be always-on and will not terminate. <https://docs.databricks.com/delta-live-tables/updates.html#continuous-triggered>  
upvoted 3 times
- 🗳️ 👤 **Inhaler\_boy** 10 months ago  
You could be correct:  
"Triggered pipelines can reduce resource consumption and expense since the cluster runs only long enough to execute the pipeline. However, new data won't be processed until the pipeline is triggered. Continuous pipelines require an always-running cluster, which is more expensive but reduces processing latency."  
upvoted 1 times
- 🗳️ 👤 **Inhaler\_boy** 10 months ago  
Actually it could make A the correct answer?  
upvoted 2 times
- 🗳️ 👤 **Atnafu** 11 months, 2 weeks ago  
C.  
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.  
  
In a Delta Live Table pipeline running in Continuous Pipeline Mode, when you click Start to update the pipeline, the following outcome is expected:  
  
All datasets defined using STREAMING LIVE TABLE and LIVE TABLE against Delta Lake table sources will be updated at set intervals. The compute resources will be deployed for the update process and will be active during the execution of the pipeline. The compute resources will be terminated when the pipeline is stopped or shut down.  
This mode allows for continuous and periodic updates to the datasets as new data arrives or changes in the underlying Delta Lake tables occur. The compute resources are provisioned and utilized during the update intervals to process the data and perform the necessary operations.  
upvoted 2 times
- 🗳️ 👤 **chays** 1 year ago  
**Selected Answer: C**  
Answer: C  
upvoted 3 times
- 🗳️ 👤 **Er5** 1 year, 1 month ago  
Answer: C  
Pipeline mode - This specifies how the pipeline will be run. Choose the mode based on latency and cost requirements.  
\* Triggered pipelines run once and then shut down until the next manual or scheduled update.  
\* Continuous pipelines run continuously, ingesting new data as it arrives.  
upvoted 2 times
- 🗳️ 👤 **hrabiabw** 1 year, 2 months ago  
Answer: D  
Official Databricks practice exam with answers - question 36  
upvoted 3 times
- 🗳️ 👤 **SHINGX** 1 year, 2 months ago  
Correct Answer is C. That question in the practice test is using a Triggered Pipeline. This question is using a Continuous.  
upvoted 4 times
- 🗳️ 👤 **hrabiabw** 1 year, 2 months ago  
Yes, You're right. Thanks.  
upvoted 3 times
- 🗳️ 👤 **XiltroX** 1 year, 2 months ago  
E is not the right answer. The correct answer is C  
<https://www.databricks.com/product/delta-live-tables>  
upvoted 3 times

In order for Structured Streaming to reliably track the exact progress of the processing so that it can handle any kind of failure by restarting and/or reprocessing, which of the following two approaches is used by Spark to record the offset range of the data being processed in each trigger?

**A. Checkpointing and Write-ahead Logs**

- B. Structured Streaming cannot record the offset range of the data being processed in each trigger.
- C. Replayable Sources and Idempotent Sinks
- D. Write-ahead Logs and Idempotent Sinks
- E. Checkpointing and Idempotent Sinks

**Correct Answer: E**

Community vote distribution

A (88%)

13%

 **squidy24** 1 month ago

**Selected Answer: A**

The answer is A

"Structured Streaming is a scalable and fault-tolerant stream processing engine built on the Spark SQL engine. ... Finally, the system ensures end-to-end exactly-once fault-tolerance guarantees through checkpointing and Write-Ahead Logs." - Apache Spark Structured Streaming Programming Guide

upvoted 1 times

 **keensolution** 1 month, 2 weeks ago

Nice information and i hope best [url=https://keensolution.in/data-visualization-services/]Data visualization agencies in India[/url]

upvoted 1 times

 **bita7** 1 month, 2 weeks ago

The answer is Checkpointing and idempotent sinks (E)

How does structured streaming achieves end to end fault tolerance:

- First, Structured Streaming uses checkpointing and write-ahead logs to record the offset range of data being processed during each trigger interval.
- Next, the streaming sinks are designed to be `_idempotent_`—that is, multiple writes of the same data (as identified by the offset) do not result in duplicates being written to the sink.

Taken together, replayable data sources and idempotent sinks allow Structured Streaming to ensure end-to-end, exactly-once semantics under any failure condition

upvoted 1 times


 **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: A**

1 checkpointing and write ahead logs to record the offset range of data being processed

2 checkpointing and idempotent sinks achieve end to end fault tolerance


upvoted 1 times

 **SerGrey** 5 months, 2 weeks ago

**Selected Answer: A**

Correct is A

upvoted 1 times

  **juadaves** 9 months, 2 weeks ago

The answer is Checkpointing and idempotent sinks

How does structured streaming achieves end to end fault tolerance:

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

Next, the streaming sinks are designed to be \_idempotent\_ —that is, multiple writes of the same data (as identified by the offset) do not result duplicates being written to the sink.

Taken together, replayable data sources and idempotent sinks allow Structured Streaming to ensure end-to-end, exactly-once semantics under any failure condition.

upvoted 3 times

  **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: A**

A. Checkpointing and Write-ahead Logs

To reliably track the exact progress of processing and handle failures in Spark Structured Streaming, Spark uses both checkpointing and write ahead logs. Checkpointing allows Spark to periodically save the state of the streaming application to a reliable distributed file system, which can be used for recovery in case of failures. Write-ahead logs are used to record the offset range of data being processed, ensuring that the system can recover and reprocess data from the last known offset in the event of a failure.

upvoted 2 times

  **akk\_1289** 11 months ago

A:



The engine uses checkpointing and write-ahead logs to record the offset range of the data being processed in each trigger.

-- in the link search for "The engine uses " you'll find the answer.

[https://spark.apache.org/docs/latest/structured-streaming-programming-](https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html#:~:text=The%20engine%20uses%20checkpointing%20and,being%20processed%20in%20each%20trigger.)

[guide.html#:~:text=The%20engine%20uses%20checkpointing%20and,being%20processed%20in%20each%20trigger.](https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html#:~:text=The%20engine%20uses%20checkpointing%20and,being%20processed%20in%20each%20trigger.)

upvoted 2 times

  **Atnafu** 11 months, 2 weeks ago

A. Checkpointing and Write-ahead Logs.

Checkpointing is a process of periodically saving the state of the streaming computation to a durable storage system. This ensures that if the streaming computation fails, it can be restarted from the last checkpoint and resume processing from where it left off.

Write-ahead logs are a type of log that records all changes made to a dataset. This allows Structured Streaming to recover from failures by replaying the write-ahead logs from the last checkpoint.

upvoted 3 times

  **mimzzz** 1 year ago

why i think both A & E are correct? <https://learn.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-streaming-exactly-once#:~:text=Use%20idempotent%20sinks>

upvoted 2 times

  **ZSun** 1 year ago

spark handle streaming failure through:

1. track the progress/offset(This is option A)

2. fix failure(This is option E)

But the question is "two approaches ... record the offset range"

Therefore, A

upvoted 5 times

  **chays** 1 year ago

**Selected Answer: A**

Answer is A:

The engine uses checkpointing and write-ahead logs to record the offset range of the data being processed in each trigger. The streaming sinks are designed to be idempotent for handling reprocessing. Together, using replayable sources and idempotent sinks, Structured Streaming can ensure end-to-end exactly-once semantics under any failure.

[https://spark.apache.org/docs/latest/structured-streaming-programming-](https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html#:~:text=The%20engine%20uses%20checkpointing%20and,being%20processed%20in%20each%20trigger.)

[guide.html#:~:text=The%20engine%20uses%20checkpointing%20and,being%20processed%20in%20each%20trigger.](https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html#:~:text=The%20engine%20uses%20checkpointing%20and,being%20processed%20in%20each%20trigger.)

upvoted 3 times



  **prasioso** 1 year, 1 month ago

**Selected Answer: A**

Answer is A.

From Spark documentation: Every streaming source is assumed to have offsets to track the read position in the stream. The engine uses checkpointing and write-ahead logs to record the offset range of the data being processed in each trigger.

upvoted 2 times

  **Majiji** 1 year, 1 month ago

**Selected Answer: E**

E. Checkpointing and Idempotent Sinks are the two approaches used by Spark to record the offset range of the data being processed in each trigger, enabling Structured Streaming to reliably track the exact progress of the processing so that it can handle any kind of failure by restarting and/or reprocessing. Checkpointing periodically checkpoints the state of the streaming query to a fault-tolerant storage system, while idempotent sinks ensure that data can be written multiple times to the sink without affecting the final result.

upvoted 1 times

  **Majiji** 1 year, 1 month ago

Answer is A:

The engine uses checkpointing and write-ahead logs to record the offset range of the data being processed in each trigger. The streaming sinks are designed to be idempotent for handling reprocessing. Together, using replayable sources and idempotent sinks, Structured Streaming can ensure end-to-end exactly-once semantics under any failure.

<https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html#:~:text=The%20engine%20uses%20checkpointing%20and,being%20processed%20in%20each%20trigger.>

upvoted 1 times

  **4be8126** 1 year, 2 months ago

**Selected Answer: E**



The answer is E. Checkpointing and Idempotent Sinks are used by Spark to record the offset range of the data being processed in each trigger. Checkpointing helps to recover the query from the point of failure and Idempotent Sinks ensure that the output of a streaming query is consistent even in the face of failures and retries.

upvoted 1 times

  **XiltroX** 1 year, 2 months ago

Wrong answer. Please check official databricks documentation to confirm that the right answer is A.

upvoted 5 times

  **XiltroX** 1 year, 2 months ago

**Selected Answer: A**

E is a partial answer. The two correct answers are A and E. Structured streaming is important because it uses these two methods to make sure there is fault tolerance and Exactly-once guarantee of data

upvoted 4 times

Question #28

Topic 1

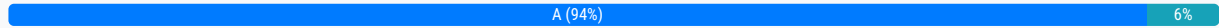
Which of the following describes the relationship between Gold tables and Silver tables?

**A. Gold tables are more likely to contain aggregations than Silver tables.**

- B. Gold tables are more likely to contain valuable data than Silver tables.
- C. Gold tables are more likely to contain a less refined view of data than Silver tables.
- D. Gold tables are more likely to contain more data than Silver tables.
- E. Gold tables are more likely to contain truthful data than Silver tables.

**Correct Answer: C**

Community vote distribution



☐ **jaromarg** 1 week, 5 days ago

In some data processing pipelines, particularly those following a "Bronze-Silver-Gold" data lakehouse architecture, Silver tables are indeed considered a more refined version of raw or Bronze data. Gold tables, which represent the final stage of data processing, typically contain highly refined, aggregated, and ready-to-consume data.

Therefore, it's common for Gold tables to contain aggregations, as they often represent the final, summarized, and aggregated view of the data. On the other hand, Silver tables may contain partially aggregated or cleansed data but are not typically the final destination for aggregated data. "Gold tables are more likely to contain aggregations than Silver tables" is accurate, making option A a valid choice.

upvoted 1 times

☐ **Dusica** 2 weeks, 5 days ago

A; row data = bronze data > silver data > golden data

C is so opposite and wrong

upvoted 1 times

☐ **carlosmps** 1 month ago

Raw Data > Bronze Data > Silver Data > Golden Data

upvoted 1 times

☐ **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: A**

correct is A

upvoted 1 times

☐ **SerGrey** 5 months, 2 weeks ago

**Selected Answer: A**

Correct is A

upvoted 1 times

☐ **awofalus** 7 months, 2 weeks ago

**Selected Answer: A**

Correct: A

upvoted 1 times

☐ **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: A**

A. Gold tables are more likely to contain aggregations than Silver tables.

In some data processing pipelines, especially those following a typical "Bronze-Silver-Gold" data lakehouse architecture, Silver tables are often considered a more refined version of the raw or Bronze data. Silver tables may include data cleansing, schema enforcement, and some initial transformations.



Gold tables, on the other hand, typically represent a stage where data is further enriched, aggregated, and processed to provide valuable insights for analytical purposes. This could indeed involve more aggregations compared to Silver tables.

upvoted 4 times

☐ **Inhaler\_boy** 10 months, 1 week ago

To me it seems A and E is equally correct. Truthful is not very defined in the question. But Gold layer typically have more rules and transformations in order to be consumed by business and reports. So it could be interpreted as more "truthful". Or am I wrong here?

upvoted 2 times

  **Atnafu** 11 months, 2 weeks ago

B

2 Type of Tables in Delta Lake data lake architecture

Gold tables are the most refined and valuable tables in the data lake, while Silver tables are less refined and less valuable.

Gold tables are typically used for downstream analysis and reporting, while Silver tables are typically used for data exploration and experimentation.

Gold tables typically contain the most refined, high-quality, and valuable data in an organization's data architecture. They often represent the final output or result of data processing pipelines, where data has undergone extensive cleansing, transformation, and aggregation. Gold tables are typically used for critical business analysis, reporting, and decision-making processes.

Option A: Gold tables are not necessarily more likely to contain aggregations than Silver tables.

Option C: Gold tables are more likely to contain a more refined view of data than Silver tables.

Option D: Gold tables are not necessarily more likely to contain more data than Silver tables.

upvoted 1 times

  **Inhaleboy** 10 months, 1 week ago

The data itself should be the same. However the Transformations are not. Gold Layer, as I understand it, is more probable to have more transformations as its ready for reports and business consumptions. So A?

"The Gold layer is for reporting and uses more de-normalized and read-optimized data models with fewer joins. The final layer of data transformations and data quality rules are applied here."

<https://www.databricks.com/glossary/medallion-architecture>

upvoted 2 times

  **4be8126** 1 year, 2 months ago

**Selected Answer: B**



The correct answer is B. Gold tables are typically considered to be the most valuable and trusted data assets in an organization. They represent the final, refined view of the data after all cleaning, transformations, and enrichments have been performed. Silver tables are the intermediate tables that feed into the Gold tables, and are typically used to perform data cleansing, filtering, and enrichment before the data is promoted to Gold.

upvoted 1 times

  **XiltroX** 1 year, 2 months ago

Dude you are providing all the wrong answers and giving baseless explanations without any link to a documentation or something. Please stop misleading people.

upvoted 6 times

  **rafahb** 1 year, 2 months ago

**Selected Answer: A**

A is correct



upvoted 2 times

  **surrabhi\_4** 1 year, 2 months ago

**Selected Answer: A**

Option A

upvoted 4 times

  **XiltroX** 1 year, 2 months ago

**Selected Answer: A**

THE ANSWER C IS INCORRECT! Silver tables usually contain data that is commonly a little more refined than Bronze tables. Meaning they contain data that is likely cleaned and contains no duplicates. Gold tables usually contain aggregate or "corrected" data.

upvoted 4 times

Question #29

Topic 1

Which of the following describes the relationship between Bronze tables and raw data?

A. Bronze tables contain less data than raw data files.

B. Bronze tables contain more truthful data than raw data.

C. Bronze tables contain aggregates while raw data is unaggregated.

D. Bronze tables contain a less refined view of data than raw data.

**E. Bronze tables contain raw data with a schema applied.**

**Correct Answer: C**


Community vote distribution

E (100%)

 **XiltroX** Highly Voted 1 year, 2 months ago


**Selected Answer: E**

Bronze tables are basically raw ingested data, often with schema borrowed from the original data source or table. Correct answer is E.  
upvoted 11 times

 **benni\_ale** Most Recent 1 month, 3 weeks ago

**Selected Answer: E**

still i am not sure about the schema as i thought that correct types are usually defined in silver while in bronze are all strings  
upvoted 1 times

 **SerGrey** 5 months, 2 weeks ago

**Selected Answer: E**

Correct is E  
upvoted 2 times

 **awofalus** 7 months, 2 weeks ago

**Selected Answer: E**

E is correct  
upvoted 1 times

 **DavidRou** 7 months, 3 weeks ago

**Selected Answer: E**

E is the right answer. Bronze data are simply a more structured (in terms of schema) version of raw data to be found in the "landing area".  
upvoted 2 times

 **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: E**

E. Bronze tables contain raw data with a schema applied.

In a typical data processing pipeline following a "Bronze-Silver-Gold" data lakehouse architecture, Bronze tables are the initial stage where raw data is ingested and transformed into a structured format with a schema applied. The schema provides structure and meaning to the raw data making it more usable and accessible for downstream processing.

Therefore, Bronze tables contain the raw data but in a structured and schema-enforced format, which makes them distinct from the unprocessed, unstructured raw data files.  
upvoted 2 times

 **akk\_1289** 11 months ago

Ans : E

The Bronze layer is where we land all the data from external source systems. The table structures in this layer correspond to the source system table structures "as-is," along with any additional metadata columns that capture the load date/time, process ID, etc. The focus in this layer is quick Change Data Capture and the ability to provide an historical archive of source (cold storage), data lineage, auditability, reprocessing if needed without rereading the data from the source system.

<https://www.databricks.com/glossary/medallion-architecture#:~:text=Bronze%20layer%20%28raw%20data%29>  
upvoted 3 times

 **akk\_1289** 11 months ago

Ans: E

<https://www.databricks.com/glossary/medallion-architecture#:~:text=Bronze%20layer%20%28raw%20data%29>  
upvoted 1 times



🗨️ 👤 **Atnafu** 11 months, 2 weeks ago

E

Bronze tables are the foundation of the Delta Lake data lake architecture. They are created from raw data files and contain a schema that describes the data. This makes it easy to query and analyze the data in Bronze tables.

Raw data files, on the other hand, do not have a schema applied. This means that it can be difficult to query and analyze the data in raw data files.

Option A: Bronze tables typically contain more data than raw data files, because they include the schema.

Option B: There is no indication that Bronze tables contain more truthful data than raw data.

Option C: Bronze tables can contain aggregates, but they do not have to.

Option D: Bronze tables typically contain a more refined view of data than raw data, because they include the schema.

upvoted 1 times

🗨️ 👤 **Atnafu** 11 months, 2 weeks ago

Sorry this is meant to be on question #30

upvoted 1 times

🗨️ 👤 **Atnafu** 11 months, 2 weeks ago

never mind :)

upvoted 1 times

🗨️ 👤 **rafahb** 1 year, 2 months ago

**Selected Answer: E**

E option

upvoted 2 times

🗨️ 👤 **surrabhi\_4** 1 year, 2 months ago

**Selected Answer: E**

Option E

upvoted 3 times

## Question #30

Topic 1

Which of the following tools is used by Auto Loader process data incrementally?

A. Checkpointing

**B. Spark Structured Streaming**

C. Data Explorer

D. Unity Catalog

E. Databricks SQL

**Correct Answer: B**

Community vote distribution

B (100%)

🗨️ 👤 **XiltroX** **Highly Voted** 👍 1 year, 2 months ago

**Selected Answer: B**

B is the correct answer. Checkpointing is a method that is part of structured streaming.

upvoted 7 times

  **benni\_ale** Most Recent 1 month, 3 weeks ago

**Selected Answer: B**

run moley run



upvoted 1 times

  **RBKasemodel** 5 months ago

The answer should be A.

Auto Loader is used by Structured Streaming to process data incrementally, not the other way around.



upvoted 2 times

  **SerGrey** 5 months, 2 weeks ago

**Selected Answer: B**

Correct is B

upvoted 1 times

  **awofalus** 7 months, 2 weeks ago

**Selected Answer: B**

B is correct


upvoted 1 times

  **anandpsg101** 8 months, 1 week ago

**Selected Answer: B**

B is orrect

upvoted 1 times

  **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: B**

B. Spark Structured Streaming

The Auto Loader process in Databricks is typically used in conjunction with Spark Structured Streaming to process data incrementally. Spark Structured Streaming is a real-time data processing framework that allows you to process data streams incrementally as new data arrives. The Auto Loader is a feature in Databricks that works with Structured Streaming to automatically detect and process new data files as they are added to a specified data source location. It allows for incremental data processing without the need for manual intervention.

upvoted 2 times

  **akk\_1289** 11 months ago

ans:A



How does Auto Loader track ingestion progress?

As files are discovered, their metadata is persisted in a scalable key-value store (RocksDB) in the checkpoint location of your Auto Loader pipeline. This key-value store ensures that data is processed exactly once.

In case of failures, Auto Loader can resume from where it left off by information stored in the checkpoint location and continue to provide exactly once guarantees when writing data into Delta Lake. You don't need to maintain or manage any state yourself to achieve fault tolerance or exactly once semantics.

<https://docs.databricks.com/ingestion/auto-loader/index.html>

upvoted 1 times

  **akk\_1289** 11 months ago

ans:B


How does Auto Loader track ingestion progress?

As files are discovered, their metadata is persisted in a scalable key-value store (RocksDB) in the checkpoint location of your Auto Loader pipeline. This key-value store ensures that data is processed exactly once.

In case of failures, Auto Loader can resume from where it left off by information stored in the checkpoint location and continue to provide exactly once guarantees when writing data into Delta Lake. You don't need to maintain or manage any state yourself to achieve fault tolerance or exactly once semantics.

<https://docs.databricks.com/ingestion/auto-loader/index.html>

upvoted 1 times

 **Atnafu** 11 months, 2 weeks ago

B

Auto Loader uses Spark Structured Streaming to process data incrementally. Spark Structured Streaming is a streaming engine that can be used to process data as it arrives. This makes it ideal for processing data that is being generated in real time.

Option A: Checkpointing is a technique used to ensure that data is not lost in case of a failure. It is not used to process data incrementally.

Option C: Data Explorer is a data exploration tool that can be used to explore data. It is not used to process data incrementally.

Option D: Unity Catalog is a metadata management tool that can be used to store and manage metadata about data assets. It is not used to process data incrementally.

Option E: Databricks SQL is a SQL engine that can be used to query data. It is not used to process data incrementally.

upvoted 2 times

 **surrabhi\_4** 1 year, 2 months ago

**Selected Answer: B**

Option B

upvoted 2 times

Question #31

Topic 1

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")
  .trigger(
    .table("new_sales")
  )
)
```

If the data engineer only wants the query to execute a micro-batch to process data every 5 seconds, which of the following lines of code should the data engineer use to fill in the blank?

A. trigger("5 seconds")

B. trigger()

C. trigger(once="5 seconds")


**D. trigger(processingTime="5 seconds")**

E. trigger(continuous="5 seconds")

**Correct Answer: D**

Community vote distribution

D (100%)

 **XiltroX** **Highly Voted** 1 year, 2 months ago

D is the correct answer

upvoted 5 times

4be8126 Highly Voted 1 year, 2 months ago

**Selected Answer: D**

The correct line of code to fill in the blank to execute a micro-batch to process data every 5 seconds is:

D. trigger(processingTime="5 seconds")

Option A ("trigger("5 seconds)") would not work because it does not specify that the trigger should be a processing time trigger, which is necessary to trigger a micro-batch processing at regular intervals.

Option B ("trigger()") would not work because it would use the default trigger, which is not a processing time trigger.

Option C ("trigger(once="5 seconds)") would not work because it would only trigger the query once, not at regular intervals.

Option E ("trigger(continuous="5 seconds)") would not work because it would trigger the query to run continuously, without any pauses in between, which is not what the data engineer wants.

upvoted 5 times

benni\_ale Most Recent 1 month, 3 weeks ago

**Selected Answer: D**

correct syntax is D

upvoted 1 times

awofalus 7 months, 2 weeks ago

**Selected Answer: D**

Correct: D

upvoted 1 times

vctrhugo 9 months, 2 weeks ago

**Selected Answer: D**

```
# ProcessingTime trigger with two-seconds micro-batch interval
df.writeStream \
  .format("console") \
  .trigger(processingTime='2 seconds') \
  .start()
```

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

upvoted 2 times

AndreFR 10 months ago

**Selected Answer: D**

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

upvoted 1 times

Atnafu 11 months, 2 weeks ago

```
D
val query = sourceTable
.writeStream
.format("delta")
.outputMode("append")
.trigger(Trigger.ProcessingTime("5 seconds"))
.start(destinationTable)
```

upvoted 1 times

vctrhugo 9 months, 2 weeks ago

This is Scala example. Exam should be 100% on Python.

upvoted 2 times

rafahb 1 year, 2 months ago

**Selected Answer: D**

D os correct

upvoted 2 times

surrabhi\_4 1 year, 2 months ago

**Selected Answer: D**

Option D

upvoted 3 times

A dataset has been defined using Delta Live Tables and includes an expectations clause:

CONSTRAINT valid\_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION DROP ROW


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

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

**Correct Answer: D**

Community vote distribution

C (100%)

 **XiltroX** Highly Voted 1 year, 2 months ago

**Selected Answer: C**

I am simply appalled by the number of wrong answers in this series of questions. The statement in the question already says "ON VIOLATE DROP ROW" which means if condition is violated, there will be nothing saved to quarantine table and a log of all invalid entries will be recorded. All invalid data that doesn't meet condition will be dropped.

So C is the correct answer.


upvoted 14 times

 **rafahb** Highly Voted 1 year, 2 months ago

**Selected Answer: C**

C is correct

upvoted 5 times

 **benni\_ale** Most Recent 1 month, 3 weeks ago

**Selected Answer: C**

C is correct

upvoted 1 times

 **SerGrey** 5 months, 1 week ago

**Selected Answer: C**

C is correct

upvoted 1 times

🗨️ **Garyn** 5 months, 3 weeks ago

**Selected Answer: C**

C. Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.

Explanation:

The defined expectation specifies that if the timestamp is not greater than '2020-01-01', the row will be considered in violation of the constraint. The ON VIOLATION DROP ROW clause states that rows that violate the constraint will be dropped from the target dataset.

Additionally, the expectation clause will log these violations in the event log, indicating which records did not meet the specified constraint criteria.

This behavior ensures that the rows failing the defined constraint are not included in the target dataset and are logged as invalid in the event log for reference or further investigation, maintaining data integrity within the dataset based on the specified constraints.

upvoted 2 times

🗨️ **Huroye** 7 months, 1 week ago

who choses these answers? The correct answer is C. The record is dropped. This is not about the default behavior. It is explicit.

upvoted 1 times

🗨️ **DavidRou** 7 months, 3 weeks ago

**Selected Answer: C**

Right answer: C

Invalid rows will be dropped as requested by the constraint and flagged as such in log files. If you need a quarantine table, you'll have to write more code.

upvoted 1 times

🗨️ **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: C**

C. Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.

With the defined constraint and expectation clause, when a batch of data is processed, any records that violate the expectation (in this case, where the timestamp is not greater than '2020-01-01') will be dropped from the target dataset. These dropped records will also be recorded as invalid in the event log, allowing for auditing and tracking of the data quality issues without causing the entire job to fail.

upvoted 2 times

🗨️ **AndreFR** 10 months ago

**Selected Answer: C**

<https://docs.databricks.com/en/delta-live-tables/expectations.html>

upvoted 2 times

🗨️ **Atnafu** 11 months, 2 weeks ago

C

When a batch of data is processed in Delta Live Tables and contains data that violates the defined expectations or constraints, the expected behavior is that the records violating the expectation are dropped from the target dataset. Additionally, these violated records are recorded as invalid in the event log.

upvoted 1 times

🗨️ **mehroosali** 11 months, 2 weeks ago

**Selected Answer: C**

C is correct

upvoted 1 times

🗨️ **SHINGX** 1 year, 2 months ago

B is correct. This question is number 35 on the practice test on databricks partner academy. <https://partner-academy.databricks.com/> correct answer is "Records that violate the expectation are added to the target dataset and recorded as invalid in the event log"

upvoted 2 times

🗨️ **SHINGX** 1 year, 2 months ago

Sorry, D

upvoted 1 times

🗨️ **SHINGX** 1 year, 2 months ago

I was wrong, the ON VIOLATION DROP ROW makes C the correct answer

upvoted 5 times

Which of the following describes when to use the CREATE STREAMING LIVE TABLE (formerly CREATE INCREMENTAL LIVE TABLE) syntax over the CREATE LIVE TABLE syntax when creating Delta Live Tables (DLT) tables using SQL?

- A. CREATE STREAMING LIVE TABLE should be used when the subsequent step in the DLT pipeline is static.
- B. CREATE STREAMING LIVE TABLE should be used when data needs to be processed incrementally**
- C. CREATE STREAMING LIVE TABLE is redundant for DLT and it does not need to be used.
- D. CREATE STREAMING LIVE TABLE should be used when data needs to be processed through complicated aggregations.
- E. CREATE STREAMING LIVE TABLE should be used when the previous step in the DLT pipeline is static.

**Correct Answer: B**

*Community vote distribution*

B (100%)

 **XiltroX** Highly Voted 1 year, 2 months ago

**Selected Answer: B**

B is the correct answer.  
upvoted 6 times

 **4he8126** Highly Voted 1 year, 2 months ago

Question #34

Topic 1

A data engineer is designing a data pipeline. The source system generates files in a shared directory that is also used by other processes. As a result, the files should be kept as is and will accumulate in the directory. The data engineer needs to identify which files are new since the previous run in the pipeline, and set up the pipeline to only ingest those new files with each run.  
Which of the following tools can the data engineer use to solve this problem?

- A. Unity Catalog
- B. Delta Lake
- C. Databricks SQL
- D. Data Explorer
- E. Auto Loader**

**Correct Answer: E**

Community vote distribution

E (100%)




 **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: E**

E is correct

upvoted 1 times

 **SerGrey** 5 months, 1 week ago

**Selected Answer: E**

E is correct

upvoted 1 times

 **Huroye** 7 months ago

the data engineer needs to identify which files are new since the previous run. This seems to be an analysis effort. If that is the case, and I might be wrong, then DB SQL is the correct answer.

upvoted 1 times

## Question #35

Topic 1

Which of the following Structured Streaming queries is performing a hop from a Silver table to a Gold table?

- A. 

```
(spark.readStream.load(rawSalesLocation)
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("append")
  .table("newSales")
)
```
- B. 

```
(spark.read.load(rawSalesLocation)
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("append")
  .table("newSales")
)
```
- C. 

```
(spark.table("sales")
  .withColumn("avgPrice", col("sales") / col("units"))
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("append")
  .table("newSales")
)
```
- D. 

```
(spark.table("sales")
  .filter(col("units") > 0)
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("append")
  .table("newSales")
)
```
- E. 

```
(spark.table("sales")
  .groupBy("store")
  .agg(sum("sales"))
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("complete")
  .table("newSales")
)
```



**Correct Answer: E**

Community vote distribution

  **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: E**



E - Aggregations are performed from silver to gold  
upvoted 1 times

  **agAshish** 4 months, 2 weeks ago

Answer should be A , for writestream data should be stream only and not static  
upvoted 1 times

  **benni\_ale** 1 month, 3 weeks ago

that's a good point but i would say that A is performing a raw data ingestion into bronze  
upvoted 1 times

  **SerGrey** 5 months, 1 week ago

**Selected Answer: E**

E is correct  
upvoted 1 times

  **surya\_lolla** 7 months, 1 week ago

The best practice is to use "Complete" as output mode instead of "append" when working with aggregated tables. Since gold layer is work fir aggregated tables, the only option with output mode as complete is option E.  
upvoted 2 times

  **DavidRou** 7 months, 3 weeks ago

**Selected Answer: E**

E is the right answer. The "gold layer" is used to store aggregated clean data, E is the only answer in wich aggregation is performed.  
upvoted 1 times



  **tocs** 8 months, 2 weeks ago

**Selected Answer: E**

E as we're doing an aggregation and we're rewriting the whole table and not just appending.  
upvoted 2 times


  **GhaneshK** 10 months, 2 weeks ago

E is correct as it includes group by as well by store.  
upvoted 2 times

  **rafahb** 1 year, 2 months ago


**Selected Answer: E**

E option  
upvoted 3 times

  **surrabhi\_4** 1 year, 2 months ago

**Selected Answer: E**

Option E  
upvoted 3 times

  **XiltroX** 1 year, 2 months ago

E is the correct answer.  
upvoted 4 times

## Question #36

Topic 1

A data engineer has three tables in a Delta Live Tables (DLT) pipeline. They have configured the pipeline to drop invalid records at each table. They notice that some data is being dropped due to quality concerns at some point in the DLT pipeline. They would like to determine at which table in their pipeline the data is being dropped.

Which of the following approaches can the data engineer take to identify the table that is dropping the records?

- A. They can set up separate expectations for each table when developing their DLT pipeline.
- B. They cannot determine which table is dropping the records.
- C. They can set up DLT to notify them via email when records are dropped.
- D. They can navigate to the DLT pipeline page, click on each table, and view the data quality statistics**
- E. They can navigate to the DLT pipeline page, click on the "Error" button, and review the present errors.

**Correct Answer: E**

Community vote distribution

D (100%)



  **vctrhugo** **Highly Voted** 9 months, 2 weeks ago

**Selected Answer: D**

D. They can navigate to the DLT pipeline page, click on each table, and view the data quality statistics.

To identify the table in a Delta Live Tables (DLT) pipeline where data is being dropped due to quality concerns, the data engineer can navigate the DLT pipeline page, click on each table in the pipeline, and view the data quality statistics. These statistics often include information about records dropped, violations of expectations, and other data quality metrics. By examining the data quality statistics for each table in the pipeline, the data engineer can determine at which table the data is being dropped.


upvoted 10 times

  **benni\_ale** **Most Recent** 1 month, 3 weeks ago

**Selected Answer: D**

I would say D but I have never really tested it, still other solutions smell wrong

upvoted 1 times

  **agAshish** 4 months, 2 weeks ago

D is correct

By clicking on each table in the DLT pipeline page, the data engineer may be able to access data quality statistics, error logs, or other information related to dropped records. This can help them pinpoint at which table in the pipeline the data is being dropped.


upvoted 1 times

  **Diewrine** 6 months, 2 weeks ago

**Selected Answer: D**

E is for when an error occur. But pipeline is defined to drop some records that will not result on error



upvoted 2 times

  **awofalus** 7 months, 2 weeks ago

**Selected Answer: D**

D is correct

upvoted 1 times

  **Atnafu** 11 months, 2 weeks ago

E

When records are dropped due to quality concerns in a DLT pipeline, the errors are logged in the event log. The data engineer can navigate to DLT pipeline page and click on the "Error" button to view the present errors. The errors will show the table where the records were dropped. Option A: Setting up separate expectations for each table will not help the data engineer determine which table is dropping the records.

Option B: The data engineer cannot determine which table is dropping the records without looking at the event log.

Option C: Setting up DLT to notify the data engineer via email when records are dropped will not help the data engineer determine which table dropping the records.

Option D: Viewing the data quality statistics for each table will not help the data engineer determine which table is dropping the records.

upvoted 2 times

  **DavidRou** 7 months, 3 weeks ago

Don't you have to select a table generated in a single step of the pipeline to access the errors through the button though? Probably D is the right one here

upvoted 1 times



  **prasioso** 1 year, 1 month ago

**Selected Answer: D**

Think answer is D.

The pipeline is configured to drop invalid records, i.e. a SQL equivalent query with a ON VIOLATION DROP ROW clause. This will not result in failed pipeline execution because there are no errors. Instead, you'd have to go to each table and review the quality characteristics.

upvoted 4 times

  **Atnafu** 11 months, 2 weeks ago

Option D is incorrect because viewing the data quality statistics for each table will not help the data engineer identify which table is dropping the records. The data quality statistics will show the overall quality of the data in each table, but they will not show which table is dropping records.

For example, if the data quality statistics for a table show that 10% of the records are invalid, this does not mean that 10% of the records are being dropped. The invalid records could be being updated, inserted, or deleted.

upvoted 2 times

  **[Removed]** 1 year, 1 month ago

Is this for v2 or v3

upvoted 3 times

## Question #37

Topic 1

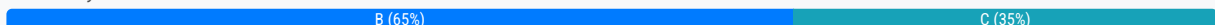
A data engineer has a single-task Job that runs each morning before they begin working. After identifying an upstream data issue, they need to set up another task to run a new notebook prior to the original task.




Which of the following approaches can the data engineer use to set up the new task?

- A. They can clone the existing task in the existing Job and update it to run the new notebook.
- B. They can create a new task in the existing Job and then add it as a dependency of the original task.**
- C. They can create a new task in the existing Job and then add the original task as a dependency of the new task.
- D. They can create a new job from scratch and add both tasks to run concurrently.
- E. They can clone the existing task to a new Job and then edit it to run the new notebook.

**Correct Answer: E**

Community vote distribution



  **Data\_4ever** **Highly Voted**  1 year, 2 months ago

**Selected Answer: B**

B is the right answer.

upvoted 14 times

🗨️ **Redwings538** Highly Voted 1 year, 2 months ago

**Selected Answer: B**

It seems there is some confusion on what dependency means in this case. Option B is correct because adding the new task as a dependency the original task means that the new task will run BEFORE the original task, which is the goal defined in the question.

upvoted 11 times

🗨️ **kokosz** Most Recent 3 weeks, 6 days ago

**Selected Answer: B**

B is the right answer.

upvoted 2 times

🗨️ **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: B**

original depends on new

upvoted 1 times

🗨️ **Mircuz** 3 months, 2 weeks ago

**Selected Answer: C**

C because the new task has to run prior the original one

upvoted 2 times

🗨️ **Nika12** 4 months, 3 weeks ago

**Selected Answer: B**

Just got 100% on the test. B was correct.

upvoted 6 times

🗨️ **Shaxxie** 5 months ago

This has become more of a English grammatical test as the word dependency is confusing people. When the Original task has a dependency the new task this means the original task needs to depend on the new task. So it's Option C.

upvoted 1 times

🗨️ **Garyn** 5 months, 3 weeks ago

**Selected Answer: C**

The data engineer can create a new task in the existing Job and then add the original task as a dependency of the new task (Option C). This way the new task will run first, and once it's completed, the original task will run. Here are the steps to do this:

Click Workflows in the sidebar and click New and select Job.

The Tasks tab appears with the create task dialog.

Replace Add a name for your job... with your job name.

Enter a name for the task in the Task name field.

In the Type drop-down menu, select the type of task to run.

Configure the cluster where the task runs.

To add dependent libraries, click + Add next to Dependent libraries.

You can pass parameters for your task.

Please note that the exact process may vary depending on the specific configurations and permissions set up in your workspace. It's always a good idea to consult with your organization's IT or data governance team to ensure the correct procedures are followed.

upvoted 3 times

🗨️ **Tinendra** 5 months, 3 weeks ago

Answer is C



upvoted 5 times

🗨️ **nedlo** 6 months, 1 week ago

**Selected Answer: B**

I am pretty sure its B - "they need to set up another task to run a new notebook prior to the original task." - so NEW task need to run BEFORE ORIGINAL task. So NEW TASK should be DEPENDENCY of ORIGINAL TASK (or in other words: original task is dependent on new task)

upvoted 1 times

  **ObeOne** 7 months, 2 weeks ago

"A data engineer has a single-task Job that runs each morning before they begin working. After identifying an upstream data issue, they need set up another task to run a new notebook prior to the original task."



In the tasks UI of the Job:

1. Create a \*new task\*
2. Select \*original task\*
3. In \*original task\* for "depends on" enter \*new task" - as \*new task\* needs to run prior to \*original task\*, ie, original task has a dependency c new task

from 1. create new task ..... from 3. original task has a dependency on new task

Answer is C ... They can \*create a new task\* in the existing Job and then add the \*original task as a dependency of the new task\*.

upvoted 5 times

  **awofalus** 7 months, 2 weeks ago

**Selected Answer: C**



Correct is C because original task will run after the newer, and then, depend on it

upvoted 2 times

  **AndreFR** 6 months ago

I disagree. "the original task as a dependency of the new task" means that the original task needs to run first.

upvoted 1 times

  **ObeOne** 7 months, 3 weeks ago

C is correct



upvoted 2 times

  **DavidRou** 7 months, 3 weeks ago

Right answer: B

We need to add the new task as a dependency of the original one because the question says that it needs to be run before the original task.


upvoted 1 times

  **kbaba101** 8 months ago

This is a Grammar issue not a Databricks issue:

Add A as a dependency of B means A must run before B.

upvoted 1 times

  **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: C**

C. They can create a new task in the existing Job and then add the original task as a dependency of the new task.


To set up a new task that runs a new notebook prior to the original task in an existing Job, you can create a new task within the same Job and then set the original task as a dependency for the new task. This way, the new task will execute before the original task when the Job is triggered.

upvoted 3 times

  **AndreFR** 6 months ago

I disagree. "the original task as a dependency of the new task" means that the original task needs to run first.

upvoted 1 times

  **[Removed]** 9 months, 3 weeks ago

**Selected Answer: B**

B is the right answer

## Question #38

## Topic 1

An engineering manager wants to monitor the performance of a recent project using a Databricks SQL query. For the first week following the project's release, the manager wants the query results to be updated every minute. However, the manager is concerned that the compute resources used for the query will be left running and cost the organization a lot of money beyond the first week of the project's release. Which of the following approaches can the engineering team use to ensure the query does not cost the organization any money beyond the first week of the project's release?

- A. They can set a limit to the number of DBUs that are consumed by the SQL Endpoint.
- B. They can set the query's refresh schedule to end after a certain number of refreshes.

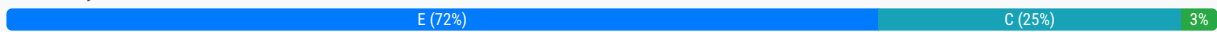
C. They cannot ensure the query does not cost the organization money beyond the first week of the project's release.

D. They can set a limit to the number of individuals that are able to manage the query's refresh schedule.

**E. They can set the query's refresh schedule to end on a certain date in the query scheduler.**

**Correct Answer: E**

Community vote distribution



**Nika12** Highly Voted 4 months, 3 weeks ago

**Selected Answer: E**

Just got 100% on the test. E was correct. C was not in the available options.

upvoted 13 times

**BigDaddyAus** Highly Voted 1 year, 1 month ago

The query scheduler only gives the option on what the interval is to run the query. It does not provide a way to stop after x iterations or at a point in time.

The question is confusing. From what I found the only option is to limit users access to the query (and therefore query scheduler).

<https://docs.databricks.com/security/auth-ctrl/access-control/query-acl.html>

Not convinced how this would be helping the organization save money if no-one is manually stopping the schedule.

Answer C seems most correct

Answer D can be achieved using acl however how is this helpful in the use case described?

upvoted 10 times

**aspix82** Most Recent 3 weeks, 6 days ago

Answer is E

upvoted 1 times

**data\_arch** 3 months, 3 weeks ago

**Selected Answer: E**

Answer is E

It's true natively the query can't be scheduled to stop, but the scheduler allow us to use cron syntax.

So we can define the year, month and days of the first week and the trigger won't run after that

upvoted 3 times

**Def21** 5 months, 1 week ago

**Selected Answer: C**

The query scheduler does not give option to have end date (or iterations). Dashboards might give one, but the question specifically mentions queries.

<https://learn.microsoft.com/en-gb/azure/databricks/sql/user/queries/schedule-query>

upvoted 2 times

**Garyn** 5 months, 3 weeks ago

**Selected Answer: E**

E. They can set the query's refresh schedule to end on a certain date in the query scheduler.

Explanation:

Query Scheduler: Databricks offers a Query Scheduler that allows users to schedule the execution of SQL queries at specific intervals or for specific durations.

Setting a Specific End Date: The team can configure the query's refresh schedule to conclude or end on a certain date. By specifying an end date within the first week of the project's release, the query will automatically stop refreshing after that date. This action ensures that compute resources aren't continuously utilized beyond the specified timeframe, preventing unnecessary costs.

This approach allows the team to control and limit the execution of the query to the required duration without incurring additional costs beyond the first week of the project's release.

upvoted 3 times

**mokrani** 6 months, 2 weeks ago

C is the correct answer

Source : <https://docs.databricks.com/en/sql/user/queries/schedule-query.html>

upvoted 1 times

🗄️ 👤 **god\_father** 7 months, 3 weeks ago

**Selected Answer: E**

E is the correct answer.

From the docs:

If a dashboard is configured for automatic updates, it has a Scheduled button at the top, rather than a Schedule button. To stop automatically updating the dashboard and remove its subscriptions:

Click Scheduled.

In the Refresh every drop-down, select Never.

Click Save. The Scheduled button label changes to Schedule.

Source: <https://learn.microsoft.com/en-us/azure/databricks/sql/user/dashboards/>

upvoted 2 times

🗄️ 👤 **Def21** 5 months, 1 week ago

This is Dashboard, not SQL query.

upvoted 2 times

🗄️ 👤 **kishore1980** 7 months, 3 weeks ago

**Selected Answer: E**

Option E is correct answer

upvoted 1 times

🗄️ 👤 **kishore1980** 7 months, 3 weeks ago

**Selected Answer: B**

The picker scrolls and allows you to choose:

An interval: 1-30 minutes, 1-12 hours, 1 or 30 days, 1 or 2 weeks

Since the schedule picker allows to choose interval to refresh query every 1 or 2 weeks. If we choose 1 week the schedule ends after a week. the answer is B.

upvoted 1 times

🗄️ 👤 **kishore1980** 7 months, 3 weeks ago

Based on my explanation E can be the correct answer.

upvoted 1 times

🗄️ 👤 **damaldon** 9 months, 2 weeks ago

Correct Answer E.

upvoted 1 times

🗄️ 👤 **[Removed]** 9 months, 3 weeks ago

**Selected Answer: C**

agree with BigDaddyAus

upvoted 1 times

🗄️ 👤 **Inhaler\_boy** 10 months, 1 week ago

**Selected Answer: C**

Answer is C. According to documentation it cant be scheduled up until a certain date. It has to be in intervals and then canceled manually. They don't mention end date. Only start date and intervals.

<https://docs.databricks.com/en/workflows/jobs/schedule-jobs.html>

upvoted 2 times

🗄️ 👤 **Inhaler\_boy** 10 months, 1 week ago

Also this link seems to verify C as the correct answer:

<https://docs.databricks.com/en/sql/user/queries/schedule-query.html>

upvoted 1 times



🗨️ 👤 **Atnafu** 11 months, 2 weeks ago

E

Option A: The query will still run, but it will be throttled if it exceeds the DBU limit.

Option B: The query will still run, but it will only run a certain number of times before it stops.

Option C: The engineering team can ensure

Option D: The query will still run, but only the individuals who are authorized to manage the refresh schedule will be able to stop it.

E-Answer

Therefore, the correct answer is that the engineering team can set the query's refresh schedule to end on a certain date in the query schedule ensure the query does not cost the organization any money beyond the first week of the project's release.

upvoted 1 times

🗨️ 👤 **ashubhar09** 8 months ago

Refresh schedule doesn't have any option to expire. So E is not correct option. <https://docs.databricks.com/en/sql/user/queries/schedule-query.html>

upvoted 1 times

🗨️ 👤 **LANDIS** 11 months, 3 weeks ago

Answer is E

<https://docs.databricks.com/sql/user/queries/schedule-query.html#schedule-a-query>

upvoted 2 times

🗨️ 👤 **chays** 1 year ago

**Selected Answer: C**

agree with BigDaddyAus

upvoted 2 times

🗨️ 👤 **Tickxit** 1 year, 1 month ago

**Selected Answer: C**

I agree with BigDaddyAus, I don't see any option to end the query scheduler.

upvoted 2 times

## Question #39

Topic 1

A data analysis team has noticed that their Databricks SQL queries are running too slowly when connected to their always-on SQL endpoint. They claim that this issue is present when many members of the team are running small queries simultaneously. They ask the data engineering team for help. The data engineering team notices that each of the team's queries uses the same SQL endpoint.

Which of the following approaches can the data engineering team use to improve the latency of the team's queries?

A. They can increase the cluster size of the SQL endpoint.

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

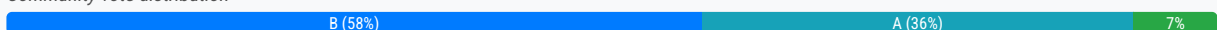
C. They can turn on the Auto Stop feature for the SQL endpoint.

D. They can turn on the Serverless 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."

**Correct Answer: B**

Community vote distribution



🗨️ 👤 **damaldon** **Highly Voted** 🍌 9 months, 2 weeks ago

Answer is B.

According to databricks documentation:

-Sequentially -> Increase cluster size  
-Concurrent --> Scale out cluster  
upvoted 26 times

🗨️ **mokrani** Highly Voted 7 months, 2 weeks ago

Answer B is correct

For those who's selected the same answer as the question 40 in the Databricks exam training, be careful because it's quite different:

- Here the question is about simultaneously runs -> Scale Out clusters (involves adding more clusters)
- In the Databricks exam training, the question is about "sequentially run queries" -> Scale Up (increasing the size of the nodes)

Please refer to the this accepted answer

<https://community.databricks.com/t5/data-engineering/sequential-vs-concurrency-optimization-questions-from-query/td-p/36696>

upvoted 14 times

🗨️ **benni\_ale** Most Recent 1 month, 3 weeks ago

Selected Answer: B

simultaneously probably means concurrently so scaling out the cluster is better

upvoted 1 times

🗨️ **sakis213** 2 months, 2 weeks ago

Selected Answer: B

B is correct

upvoted 1 times

🗨️ **niharam2021** 4 months, 1 week ago

A data analysis team has noticed that their Databricks SQL queries are running too slowly when connected to their always-on SQL endpoint. They claim that this issue is present when many members of the team are running small queries simultaneously.

upvoted 2 times

🗨️ **agAshish** 4 months, 2 weeks ago

Answer is A, Q40 -- <https://files.training.databricks.com/assessments/practice-exams/PracticeExam-DataEngineerAssociate.pdf>

upvoted 3 times

🗨️ **K\_yamini** 4 months, 1 week ago

the question on Practice set is slightly different if you look closely :- In the first scenario, the data analyst notes slow query performance for sequentially run queries on a SQL endpoint that isn't shared with other users. This suggests that the problem may be related to the configuration or performance of the SQL endpoint itself rather than contention with other users.

In the second scenario, the data analysis team experiences slow query performance when multiple team members are running queries simultaneously on the same SQL endpoint. This indicates potential resource contention or limitations on the SQL endpoint when handling concurrent queries from multiple users.

Given these differences, the approaches to address the issues may also differ:

upvoted 1 times

🗨️ **Nika12** 4 months, 3 weeks ago

Selected Answer: B

Just got 100% on the exam. B was correct. Also, here is the link to good explanation:

<https://docs.databricks.com/en/compute/cluster-config-best-practices.html>

upvoted 5 times

🗨️ **Ody\_\_** 5 months ago

Selected Answer: A

A is correct

upvoted 1 times

🗨️ **Ody\_\_** 5 months, 1 week ago

Selected Answer: A

correct answer is A

Question 40: <https://files.training.databricks.com/assessments/practice-exams/PracticeExam-DataEngineerAssociate.pdf>



upvoted 2 times

🗨️ **SerGrey** 5 months, 1 week ago

Selected Answer: B

B is correct

upvoted 2 times

  **nedlo** 6 months, 1 week ago


**Selected Answer: B**

its B because its "simultaneously by many users" so you have to scale it horizontally by increasing number of nodes :  
<https://community.databricks.com/t5/data-engineering/sequential-vs-concurrency-optimization-questions-from-query/td-p/36696>  
upvoted 3 times

  **pc1337xd** 7 months, 1 week ago



**Selected Answer: B**

Issues occur when too many users are running queries at the same time -> Increase scaling so more clusters handle the queries  
upvoted 5 times

  **god\_father** 7 months, 3 weeks ago

**Selected Answer: B**

Increasing cluster size is for vertical scalability of query execution, while scaling out cluster is for horizontal scalability of query execution  
upvoted 2 times

  **saikot** 9 months, 1 week ago

The correct answer is B  
(we can check this under databricks sql WH tool tip option. It is clearly mentioend that scaling is used to improve query "LATANCY")  
upvoted 2 times

  **vctrhugo** 9 months, 2 weeks ago



**Selected Answer: A**

A. They can increase the cluster size of the SQL endpoint.

To improve the latency of the team's queries when many members are running small queries simultaneously, you can increase the cluster size the SQL endpoint. Increasing the cluster size allocates more compute resources to handle query execution, which can help reduce query execution times and improve overall performance, especially during periods of high query concurrency.

Option B refers to adjusting scaling settings, which can also be beneficial, but increasing the cluster size (Option A) directly allocates more resources, making it a more direct approach to improving query performance.

Options C, D, and E relate to different features and configurations (Auto Stop, Serverless, and Spot Instance Policy), but they may not directly address the issue of improving query latency during high concurrency, which is the primary concern in this scenario.  
upvoted 1 times

  **[Removed]** 9 months, 3 weeks ago

**Selected Answer: A**

agree with @AndreFR  
upvoted 1 times

  **AndreFR** 10 months ago

**Selected Answer: A**

question 40 in the official databricks training exam : <https://files.training.databricks.com/assessments/practice-exams/PracticeExam-DataEngineerAssociate.pdf>  
upvoted 5 times

  **ezeik** 9 months ago

but the question is different:  
"is affecting all of their sequentially run queries."  
upvoted 4 times

  **AndreFR** 6 months ago

I agree, Answer A is incorrect. Correct answer is B, because : The key is simultaneously. The autoscaling is triggered by jobs sitting in the queue, so databricks will increase number of workers because there is a queue. If queries were running sequentially, there wouldn't be queue so increasing the cluster size would be the best choice.  
upvoted 2 times

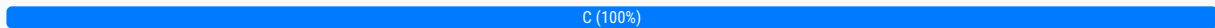
A data engineer wants to schedule their Databricks SQL dashboard to refresh once per day, but they only want the associated SQL endpoint to be running when it is necessary.

Which of the following approaches can the data engineer use to minimize the total running time of the SQL endpoint used in the refresh schedule of their dashboard?

- A. They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.
- B. They can set up the dashboard's SQL endpoint to be serverless.
- C. They can turn on the Auto Stop feature for the SQL endpoint.**
- D. They can reduce the cluster size of the SQL endpoint.
- E. They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.

**Correct Answer: C**

*Community vote distribution*



4be8126 Highly Voted 1 year, 2 months ago

Selected Answer: C

The data engineer can use the Auto Stop feature to minimize the total running time of the SQL endpoint used in the refresh schedule of their dashboard. The Auto Stop feature allows the SQL endpoint to automatically shut down when there are no active connections, which will minimize the total running time of the SQL endpoint. By scheduling the dashboard to refresh once per day, the SQL endpoint will only be running for a short period of time each day, which will minimize the total running time and reduce costs.

upvoted 8 times

SerGrey Most Recent 5 months, 1 week ago

Selected Answer: C

C is correct

upvoted 1 times

mokrani 6 months, 3 weeks ago

Why it can't be B ? . They can set up the dashboard's SQL endpoint to be serverless. ? they can use a serverless endpoint and it will only be active when required.

upvoted 1 times

awofalus 7 months, 2 weeks ago

Selected Answer: C

#### Question #41

Topic 1

A data engineer has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to an ELT job. The ELT job has its Databricks SQL query that returns the number of input records containing unexpected NULL values. The data engineer wants their entire team to be notified via a messaging webhook whenever this value reaches 100.

Which of the following approaches can the data engineer use to notify their entire team via a messaging webhook whenever the number of NULL values reaches 100?

- A. They can set up an Alert with a custom template.
- B. They can set up an Alert with a new email alert destination.
- C. They can set up an Alert with a new webhook alert destination.**
- D. They can set up an Alert with one-time notifications.
- E. They can set up an Alert without notifications.

Correct Answer: C

Community vote distribution

C (100%)

**XiltroX** Highly Voted 1 year, 2 months ago

**Selected Answer: C**

Correct answer.  
upvoted 5 times

**AndreFR** Most Recent 6 months ago

**Selected Answer: C**

<https://docs.databricks.com/en/lakehouse-monitoring/monitor-alerts.html>

Monitor alerts are created and used the same way as other Databricks SQL alerts. You create a Databricks SQL query on the monitor profile metrics table or drift metrics table. You then create a Databricks SQL alert for this query. You can configure the alert to evaluate the query at a desired frequency, and send a notification if the alert is triggered. By default, email notification is sent. You can also set up a webhook or send notifications to other applications such as Slack or Pagerduty.

upvoted 2 times

**awofalus** 7 months, 2 weeks ago

**Selected Answer: C**

C is correct  
upvoted 1 times

**vctrhugo** 9 months, 2 weeks ago

**Selected Answer: C**

C. They can set up an Alert with a new webhook alert destination.

To notify their entire team via a messaging webhook whenever the number of NULL values reaches 100, the data engineer can set up an Alert Databricks with a new webhook alert destination. This allows them to configure the alert to trigger when the specified condition (reaching 100 NULL values) is met, and the notification can be sent to the team's messaging webhook.

Option C provides the specific approach to achieve the desired outcome of notifying the team via a messaging webhook when the condition is met.

upvoted 2 times

**Atnafu** 11 months, 2 weeks ago

C

Alerts allow you to be notified when something goes wrong in your Databricks environment. You can set up alerts to be notified by email, webhook, or Slack.

Webhooks are a way to send data from one application to another. You can use a webhook to send data from Databricks to a messaging service such as Slack or PagerDuty.

One-time notifications allow you to be notified only once when an alert is triggered. This is useful if you only want to be notified about a specific event.

Custom templates allow you to customize the email or webhook notification that is sent when an alert is triggered. This is useful if you want to include additional information in the notification, such as the name of the alert or the value of the metric that triggered the alert.

upvoted 2 times

**4be8126** 1 year, 2 months ago

## Question #42

Topic 1

A single Job runs two notebooks as two separate tasks. A data engineer has noticed that one of the notebooks is running slowly in the Job's current run. The data engineer asks a tech lead for help in identifying why this might be the case.

Which of the following approaches can the tech lead use to identify why the notebook is running slowly as part of the Job?

- A. They can navigate to the Runs tab in the Jobs UI to immediately review the processing notebook.
- B. They can navigate to the Tasks tab in the Jobs UI and click on the active run to review the processing notebook.
- C. They can navigate to the Runs tab in the Jobs UI and click on the active run to review the processing notebook.**
- D. There is no way to determine why a Job task is running slowly.
- E. They can navigate to the Tasks tab in the Jobs UI to immediately review the processing notebook.

**Correct Answer: C**

Community vote distribution

C (85%)

B (15%)

🗨️ **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: C**

c is correct

upvoted 1 times

🗨️ **Garyn** 5 months, 3 weeks ago

**Selected Answer: C**

The tech lead can navigate to the Runs tab in the Jobs UI and click on the active run to review the processing notebook (Option C). This will allow them to inspect the details of the job run, including the duration of each task, which can help identify potential performance issues.

There could be several reasons why a notebook is running slowly as part of a job. For instance, there might be a delay when the job cluster has to be spun up, or the table gets delta cached in memory and copies of files will be stored on local node's storage. Even certain operations like pandas UDFs can be slow.

Please note that the exact process may vary depending on the specific configurations and permissions set up in your workspace. It's always a good idea to consult with your organization's IT or data governance team to ensure the correct procedures are followed.

upvoted 1 times

🗨️ **csd** 5 months, 3 weeks ago

C is correct answer as we monitor job and performance of task in same way in my current project .

Task tab to add another task or edit existing one

upvoted 1 times

🗨️ **awofalus** 7 months, 2 weeks ago

**Selected Answer: C**

C is correct.

upvoted 1 times

🗨️ **AndreFR** 10 months ago

**Selected Answer: C**

The job run details page contains job output and links to logs, including information about the success or failure of each task in the job run. You can access job run details from the Runs tab for the job. To view job run details from the Runs tab, click the link for the run in the Start time column in the runs list view. To return to the Runs tab for the job, click the Job ID value.

If the job contains multiple tasks, click a task to view task run details, including:

the cluster that ran the task

the Spark UI for the task

logs for the task

metrics for the task

<https://docs.databricks.com/en/workflows/jobs/monitor-job-runs.html#job-run-details>

upvoted 4 times

🗨️ **Atnafu** 11 months, 2 weeks ago

C

In the Databricks Jobs UI, the Runs tab provides detailed information about the execution of each run in a Job. By clicking on the active run associated with the notebook running slowly, you can access the specific run details, including the notebook execution logs, execution duration, resource utilization, and any error messages or warnings.

upvoted 3 times

🗨️ **Tickxit** 1 year, 1 month ago

**Selected Answer: C**

"Job runs" tab

upvoted 2 times

🗨️ **XiltroX** 1 year, 2 months ago

**Selected Answer: C**

C is the correct answer. See link

<https://docs.databricks.com/workflows/jobs/jobs.html>

upvoted 2 times

🗨️ 👤 4be8126 1 year, 2 months ago

**Selected Answer: B**

B. They can navigate to the Tasks tab in the Jobs UI and click on the active run to review the processing notebook.

The Tasks tab in the Jobs UI provides detailed information about each task in the job, including the task's execution time, the task's logs, and task's output. By clicking on the active run for the notebook that is running slowly, the tech lead can review the task's logs and output to identify any issues that might be causing the slowdown. The Runs tab provides an overview of all runs of the job, but it does not provide detailed information about each task in the job.

upvoted 2 times

🗨️ 👤 XiltroX 1 year, 2 months ago

Wrong answer. Please see documentation and you will realize the correct answer is C  
<https://docs.databricks.com/workflows/jobs/jobs.html>

upvoted 2 times

#### Question #43

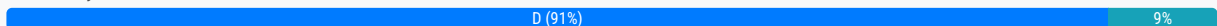
Topic 1

A data engineer has a Job with multiple tasks that runs nightly. Each of the tasks runs slowly because the clusters take a long time to start. Which of the following actions can the data engineer perform to improve the start up time for the clusters used for the Job?

- A. They can use endpoints available in Databricks SQL
- B. They can use jobs clusters instead of all-purpose clusters
- C. They can configure the clusters to be single-node
- D. They can use clusters that are from a cluster pool**
- E. They can configure the clusters to autoscale for larger data sizes

**Correct Answer: B**

Community vote distribution



🗨️ 👤 Atnafu **Highly Voted** 🍌 11 months, 2 weeks ago

D

Cluster pools are a way to pre-provision clusters that are ready to use. This can reduce the start up time for clusters, as they do not have to be created from scratch.

All-purpose clusters are not pre-provisioned, so they will take longer to start up.

Jobs clusters are a type of cluster pool, but they are not the best option for this use case. Jobs clusters are designed for long-running jobs, as they can be more expensive than other types of cluster pools.

Single-node clusters are the smallest type of cluster, and they will start up the fastest. However, they may not be powerful enough to run the Job's tasks.

Autoscaling clusters can scale up or down based on demand. This can help to improve the start up time for clusters, as they will only be created when they are needed. However, autoscaling clusters can also be more expensive than other types of cluster pool

upvoted 7 times



🗨️ **benni\_ale** Most Recent 1 month, 3 weeks ago

**Selected Answer: D**

to be fair B might seem correct but D is more appropriate for reducing start up times  
upvoted 1 times

🗨️ **Garyn** 5 months, 3 weeks ago

**Selected Answer: D**

D. They can use clusters that are from a cluster pool.

Explanation:

Cluster Pools: Cluster pools in Databricks allow for the pre-creation and management of clusters in a pool that are readily available for use. With cluster pools, clusters are pre-initialized and kept in a ready state, minimizing the startup time when tasks need to run. This reduces the overhead of cluster initialization as the clusters are already provisioned and waiting for the tasks to be assigned.

Using clusters from a pool ensures that there is no wait time for cluster initialization when the tasks start running in the nightly Job. This approach significantly reduces the time taken for clusters to start, thereby improving the overall performance and efficiency of the tasks by minimizing the overhead of cluster startup delays.

upvoted 3 times

🗨️ **DavidRou** 7 months, 3 weeks ago

**Selected Answer: D**

They must use clusters from a pool if they want to reduce the startup time.  
upvoted 3 times

🗨️ **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: D**

D. They can use clusters that are from a cluster pool.

To improve startup time for the clusters used for the Job, the data engineer can configure the clusters to be sourced from a cluster pool. Cluster pools are pre-allocated clusters that are kept in a running state, ready for use. This eliminates the need to start new clusters from scratch each time a Job runs, significantly reducing startup times.

Cluster pools are designed to optimize cluster reuse, making them an efficient choice for recurring jobs like the one described in the scenario.

Option D provides a practical solution to address the slow cluster startup time issue.

upvoted 3 times

🗨️ **AndreFR** 10 months ago

**Selected Answer: D**

You can minimize instance acquisition time by creating a pool for each instance type and Databricks runtime your organization commonly use

SOURCE : <https://docs.databricks.com/en/clusters/pool-best-practices.html>

upvoted 3 times

🗨️ **TC007** 1 year, 2 months ago

**Selected Answer: D**

D: use clusters that are from a cluster pool.

Using clusters from a cluster pool can improve the start-up time for the clusters used in the Job because the pool contains preconfigured and pre-started clusters that can be used immediately. This can save time and resources compared to starting new clusters for each task.

upvoted 4 times

🗨️ **4be8126** 1 year, 2 months ago

**Selected Answer: D**

D. They can use clusters that are from a cluster pool. Cluster pools allow you to pre-create a pool of ready-to-use clusters that can be used for running jobs, thereby eliminating the need to start new clusters each time a job runs. This can greatly reduce the startup time for each task.

upvoted 4 times

🗨️ 👤 **XiltroX** 1 year, 2 months ago

**Selected Answer: B**

B is the correct answer. Job clusters are best suited for automated tasks running on a schedule.  
upvoted 2 times

🗨️ 👤 **t30730** 1 year, 2 months ago

"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 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. U Databricks cluser pools feature to reduce the startup time"  
upvoted 1 times

🗨️ 👤 **knivesz** 1 year, 2 months ago

D es la respuesta correcta  
upvoted 2 times

#### Question #44

Topic 1

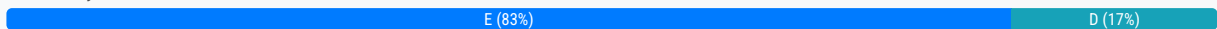
A new data engineering team team. has been assigned to an ELT project. The new data engineering team will need full privileges on the database customers to fully manage the project.

Which of the following commands can be used to grant full permissions on the database to the new data engineering team?

- A. GRANT USAGE ON DATABASE customers TO team;
- B. GRANT ALL PRIVILEGES ON DATABASE team TO customers;
- C. GRANT SELECT PRIVILEGES ON DATABASE customers TO teams;
- D. GRANT SELECT CREATE MODIFY USAGE PRIVILEGES ON DATABASE customers TO team;
- E. GRANT ALL PRIVILEGES ON DATABASE customers TO team ;**

**Correct Answer: E**

Community vote distribution



🗨️ 👤 **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: E**

e is correct  
upvoted 1 times

🗨️ 👤 **Viju\_1** 3 months ago

Examtopics not showing all the questions and asking for contributor access. I can only see qestions till 44. Anyone is able to see all 99 questions??????  
upvoted 1 times



🗨️ 👤 **akshirao** 3 months ago

i think the 99 questions have been relabelled as Q1-44 then Q1-45  
upvoted 1 times

  **Skidmee** 5 months ago

E is correct

upvoted 1 times

  **csd** 5 months, 3 weeks ago



E is correct

Template to give access-->

GRANT Privilege ON Object <object-name> TO <user or group>

ALL PRIVILEGES = gives all privilege



upvoted 4 times

  **awofalus** 7 months, 2 weeks ago

**Selected Answer: E**

E is correct

upvoted 2 times



  **DavidRou** 7 months, 3 weeks ago

**Selected Answer: D**

Right answer is E.

The template to respect is the following: GRANT <privilege> ON <resource> TO <user/group>

upvoted 2 times

  **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: E**

E. GRANT ALL PRIVILEGES ON DATABASE customers TO team;

To grant full privileges on the database "customers" to the new data engineering team, you can use the GRANT ALL PRIVILEGES command as shown in option E. This command provides the team with all possible privileges on the specified database, allowing them to fully manage it.


Option A is not correct because it grants only the USAGE privilege, which is not sufficient for full management.

Option B has the syntax reversed, and it is attempting to grant privileges on the "team" database to the "customers" database, which is not the desired action.

Option C contains incorrect syntax and should use "team" instead of "teams."

Option D has incorrect syntax and is not a valid SQL command for granting privileges in most database management systems.



upvoted 2 times

  **Atnafu** 11 months, 2 weeks ago

E

GRANT ALL PRIVILEGES ON DATABASE customers TO team;



upvoted 1 times

  **rafahb** 1 year, 2 months ago

**Selected Answer: E**

Option E

upvoted 2 times

  **XiltroX** 1 year, 2 months ago

**Selected Answer: E**

Correct answer is E. Please take note of how the questions are worded to avoid confusion and not make the wrong choice.

upvoted 3 times

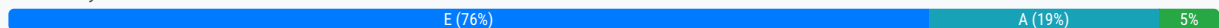
A new data engineering team has been assigned to work on a project. The team will need access to database customers in order to see what tables already exist. The team has its own group team.

Which of the following commands can be used to grant the necessary permission on the entire database to the new team?

- A. GRANT VIEW ON CATALOG customers TO team;
- B. GRANT CREATE ON DATABASE customers TO team;
- C. GRANT USAGE ON CATALOG team TO customers;
- D. GRANT CREATE ON DATABASE team TO customers;
- E. GRANT USAGE ON DATABASE customers TO team;**

**Correct Answer: E -**

Community vote distribution



☐ **Data\_4ever** Highly Voted 1 year, 2 months ago

**Selected Answer: E**

E is the correct answer.  
upvoted 13 times

☐ **nedlo** Highly Voted 6 months, 1 week ago

**Selected Answer: E**

GRANT USAGE, answer E is correct. I dont see such privilege as GRANT VIEW <https://docs.databricks.com/en/sql/language-manual/sql-ref-privileges.html#privilege-types>  
upvoted 6 times

☐ **rcpaudel** Most Recent 3 months ago

**Selected Answer: E**

The question is asking "The team will need access to database customers in order to see what tables already exist." The presumption is, how the team already has usage privilege on the catalog containing the database.  
upvoted 1 times

☐ **kishanu** 8 months ago

**Selected Answer: E**

Customers is a Database and not a catalog.  
The three-space naming convention has Catalog at the top level(catalog.database(schema).table).  
So granting a usage on catalog will expose other objects.  
upvoted 4 times

☐ **J\_1\_2** 8 months, 1 week ago

**Selected Answer: E**

Option A, "GRANT VIEW ON CATALOG customers TO team," grants the privilege to view the catalog but not access the database's tables, so might not fulfill the requirement of seeing the existing tables in the database.  
So answer is E  
upvoted 1 times

☐ **tocs** 8 months, 2 weeks ago

E.  
It can NOT be A.  
You can GRANT SELECT, but you cannot GRANT VIEW.  
VIEW is a securable OBJECT and not a PRIVILEGE TYPE, so you cannot grant it.  
See also <https://learn.microsoft.com/en-us/azure/databricks/data-governance/unity-catalog/manage-privileges/privileges>  
upvoted 2 times

🗨️ **Nina0609** 9 months, 1 week ago

It's A....I got 100% on the Data Governance section...it's A  
upvoted 2 times

🗨️ **tocs** 8 months, 2 weeks ago

There is no such thing as GRANT VIEW, so A is not a valid option  
upvoted 1 times

🗨️ **vctrhugo** 9 months, 2 weeks ago

**Selected Answer: A**

A. GRANT VIEW ON CATALOG customers TO team;

To grant the new data engineering team the necessary permission to view the tables that exist in the "customers" database, you can use the GRANT VIEW ON CATALOG command as shown in option A. This command allows the team to see the metadata and information about the tables in the specified catalog or database, which is what you want in this case.

Option B grants the CREATE privilege on the "customers" database to the team, which is not necessary for simply viewing existing tables.

Option C and Option D have the syntax reversed, attempting to grant permissions from the team to the "customers" database, which is not the desired action.

Option E grants USAGE privilege on the "customers" database to the team, which allows them to use the database but may not provide the necessary view permissions to see existing tables.

upvoted 1 times

🗨️ **AndreFR** 10 months ago

**Selected Answer: B**

SOURCE : <https://docs.databricks.com/en/sql/language-manual/security-grant.html>

The perfect answer would be : GRANT SHOW METADATA ON DATABASE customers TO team

But because, it is not suggested, we need to find an imperfect answer allowing listing tables on database customers for project team's group team

- A. GRANT VIEW ON CATALOG customers TO team -- Incorrect : "GRANT VIEW" does not exist
  - B. GRANT CREATE ON DATABASE customers TO team -- Correct : only possible answer by elimination.
  - C. GRANT USAGE ON CATALOG team TO customers -- Incorrect : "GRANT USAGE" does not allow listing tables
  - D. GRANT CREATE ON DATABASE team TO customers -- Incorrect : "team" and "customers" are inverted
  - E. GRANT USAGE ON DATABASE customers TO team -- Incorrect : "GRANT USAGE" does not allow listing tables
- upvoted 2 times

🗨️ **AndreFR** 10 months ago

In addition to what was typed previously, I'm adding an extra source : <https://docs.databricks.com/en/data-governance/table-acls/object-privileges.html#usage-privilege>

So correct syntax for what I previously wrote is :  
GRANT SHOW READ\_METADATA ON DATABASE customers TO team  
and not :  
GRANT SHOW METADATA ON DATABASE customers TO team  
upvoted 1 times

🗨️ **Office2022** 11 months ago

The correct answer is E. GRANT ALL PRIVILEGES ON DATABASE customers TO team;

The GRANT statement is used to grant privileges on a database, table, or view to a user or role. The ALL PRIVILEGES option grants all possible privileges on the specified object, such as CREATE, SELECT, MODIFY, and USAGE. The syntax of the GRANT statement is:



GRANT privilege\_type ON object TO user\_or\_role;

Therefore, to grant full permissions on the database customers to the new data engineering team, the command should be:

GRANT ALL PRIVILEGES ON DATABASE customers TO team;

Option A is incorrect because it only grants the USAGE privilege, which allows the team to access the database but not to create or modify tables or views in it.

upvoted 2 times

  **Atnafu** 11 months, 2 weeks ago

E

The GRANT USAGE ON DATABASE command grants the USAGE privilege on the specified database to the specified group. The USAGE privilege allows the group to see the tables that exist in the database, but it does not allow them to do anything else with the database.

The other commands are incorrect. The GRANT VIEW ON CATALOG command grants the VIEW privilege on the specified catalog to the specified group. The CREATE privilege allows the group to create new tables in the database. The USAGE privilege on the CATALOG does not exist.

upvoted 1 times

  **clownfishman** 11 months, 3 weeks ago

The correct is GRANT both "USAGE" and "SELECT" on DATABASE CUSTOMERS TO TEAMS

upvoted 2 times

  **chays** 1 year ago

**Selected Answer: E**

e is the right answer

upvoted 1 times

  **prasioso** 1 year, 1 month ago

**Selected Answer: E**

Think the Answer is E.

The privilege types are CREATE, MODIFY, READ\_METADATA, SELECT and USAGE. There is no such thing as GRANT VIEW. (it can however be GRANT <p-type> ON VIEW TO <team>). For our use case we want to GRANT USAGE for reading the Database. C has wrong syntax.



upvoted 2 times

  **Bob123456** 1 year, 1 month ago

option A is correct .

Question says that "The team will need access to database customers in order to see what tables already exist" this looks like they need view access

upvoted 1 times



  **chandra\_157\_447** 1 year, 1 month ago

<https://learn.microsoft.com/en-us/azure/databricks/sql/language-manual/sql-ref-privileges-hms>

In this link you can see that E is the correct answer, as there is no 'VIEW' privilege, basically view privilege is 'SELECT' privilege.

Then why will we have a VIEW privilege?, just think on process of elimination as well when answering such questions.

upvoted 2 times

  **Majiji** 1 year, 1 month ago



A:

In Databricks, the "GRANT VIEW ON CATALOG" command is used to grant permission to a user or group to view metadata in the catalog. The catalog in Databricks is a metadata management service that provides information about the data and other resources that are available within Databricks workspace.

When a user is granted the "VIEW ON CATALOG" permission, they are able to view information about databases, tables, and other resources that are available within the Databricks workspace. This information can be useful for understanding the structure and relationships of data within the workspace.

After the permission has been granted, the user will be able to view metadata in the catalog by running catalog commands, such as "SHOW DATABASES" or "DESCRIBE TABLE". Note that granting the "VIEW ON CATALOG" permission does not allow the user to modify or delete metadata in the catalog, only view it.



upvoted 1 times

  **Majiji** 1 year, 1 month ago

E:

In Databricks, the "GRANT USAGE ON DATABASE" command is used to grant a user or group the permission to use a specific database in the Databricks workspace. This command allows the user or group to perform read operations on the database, such as listing tables and running queries.

upvoted 1 times

  **Majiji** 1 year, 1 month ago

So according to the question "E" is correct

upvoted 1 times

## Question #46

Topic 1

A data engineer is running code in a Databricks Repo that is cloned from a central Git repository. A colleague of the data engineer informs them that changes have been made and synced to the central Git repository. The data engineer now needs to sync their Databricks Repo to get the changes from the central Git repository.

Which of the following Git operations does the data engineer need to run to accomplish this task?

- A. Merge
- B. Push
- C. Pull**
- D. Commit
- E. Clone

**Correct Answer: C**


Community vote distribution

C (100%)

 **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: C**

C is correct  
upvoted 2 times

 **OfficeSaracus** 3 months, 1 week ago

**Selected Answer: C**

C is correct  
upvoted 1 times

 **god\_father** 7 months, 3 weeks ago

**Selected Answer: C**

This is more of a Git question.

From the docs:

In Databricks Repos, you can use Git functionality to:

Clone, push to, and pull from a remote Git repository.

Create and manage branches for development work, including merging, rebasing, and resolving conflicts.

Create notebooks—including IPYNB notebooks—and edit them and other files.

Visually compare differences upon commit and resolve merge conflicts.

Source: <https://docs.databricks.com/en/repos/index.html>

upvoted 1 times

 **kishanu** 8 months ago

**Selected Answer: C**

pull is required from the Databricks Repo to sync the changes b/w local and central repo.  
upvoted 2 times

## Question #47

Topic 1


Which of the following is a benefit of the Databricks Lakehouse Platform embracing open source technologies?

- A. Cloud-specific integrations
- B. Simplified governance
- C. Ability to scale storage
- D. Ability to scale workloads
- E. Avoiding vendor lock-in**

**Correct Answer: E**

Community vote distribution

E (100%)

 **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: E**

E is correct

upvoted 1 times

 **UGOTCOOKIES** 4 months, 3 weeks ago

**Selected Answer: E**

E is correct as open-source is opposite of proprietary technology, so not being a proprietary means it is free of vendor lock in, if that makes sense.

upvoted 4 times

 **meow\_akk** 8 months ago

its avoiding vendor lock in : - <https://double.cloud/blog/posts/2023/01/break-free-from-vendor-lock-in-with-open-source-tech/>

upvoted 3 times

 **kishanu** 8 months ago

**Selected Answer: E**

E looks to be the correct one, as Databricks Lakeshouse platform supports Delta table which is an open-source format for storage.

upvoted 2 times

 **Rs1997** 8 months ago

D is the correct answer

upvoted 1 times



## Question #48

Topic 1

A data engineer needs to use a Delta table as part of a data pipeline, but they do not know if they have the appropriate permissions.


In which of the following locations can the data engineer review their permissions on the table?

- A. Databricks Filesystem
- B. Jobs
- C. Dashboards
- D. Repos
- E. Data Explorer**

**Correct Answer: E**

Community vote distribution


E (100%)

 **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: E**

E is correct

upvoted 1 times

 **kz\_data** 6 months, 2 weeks ago

**Selected Answer: E**

E is correct answer

upvoted 4 times

 **meow\_akk** 8 months ago

E is correct Data explorer

upvoted 2 times

## Question #49

Topic 1

Which of the following describes a scenario in which a data engineer will want to use a single-node cluster?

- A. When they are working interactively with a small amount of data**
- B. When they are running automated reports to be refreshed as quickly as possible
- C. When they are working with SQL within Databricks SQL
- D. When they are concerned about the ability to automatically scale with larger data
- E. When they are manually running reports with a large amount of data

**Correct Answer: A**

Community vote distribution

A (100%)

  **kishanu** Highly Voted 8 months ago

Selected Answer: A

Single node clusters can be used for interactive queries with small dataset  
upvoted 5 times

  **benni\_ale** Most Recent 1 month, 3 weeks ago


Selected Answer: A

A is correct  
upvoted 1 times

  **azure\_bimonster** 5 months ago

Selected Answer: A

A seems correct for this  
upvoted 2 times

  **meow\_akk** 8 months ago

ans A : A Single Node cluster is a cluster consisting of an Apache Spark driver and no Spark workers. A Single Node cluster supports Spark jobs and all Spark data sources, including Delta Lake. A Standard cluster requires a minimum of one Spark worker to run Spark jobs.  
[https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwidg8mSsYqCAxUmg2oFHbkTDJsQFnoECA4QAw&url=https%3A%2Fdocs.databricks.com%2Fen%2Fclusters%2Fsingle-node.html%23%3A~%3Atext%3DA%2520Single%2520Node%2520cluster%2520is%2Cworker%2520to%2520run%2520Spark%2520jobs%20sg=AOvVaw3PFq3\\_Qyt2gAAa4id0j6CS&opi=89978449](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwidg8mSsYqCAxUmg2oFHbkTDJsQFnoECA4QAw&url=https%3A%2Fdocs.databricks.com%2Fen%2Fclusters%2Fsingle-node.html%23%3A~%3Atext%3DA%2520Single%2520Node%2520cluster%2520is%2Cworker%2520to%2520run%2520Spark%2520jobs%20sg=AOvVaw3PFq3_Qyt2gAAa4id0j6CS&opi=89978449)  
upvoted 4 times

A data engineer has been given a new record of data:

```
id STRING = 'a1'
rank INTEGER = 6
rating FLOAT = 9.4
```

Which of the following SQL commands can be used to append the new record to an existing Delta table my\_table?

**A. INSERT INTO my\_table VALUES ('a1', 6, 9.4)**

B. my\_table UNION VALUES ('a1', 6, 9.4)

C. INSERT VALUES ( 'a1' , 6, 9.4) INTO my\_table


D. UPDATE my\_table VALUES ('a1', 6, 9.4)

E. UPDATE VALUES ('a1', 6, 9.4) my\_table

**Correct Answer: A**

Community vote distribution

A (100%)

 **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: A**

A is correct  
upvoted 1 times

 **azure\_bimonster** 5 months ago

**Selected Answer: A**

A is correct because syntax is correct  
upvoted 2 times

 **Annelijn** 5 months, 1 week ago

**Selected Answer: A**

A is correct  
upvoted 2 times

 **meow\_akk** 8 months ago

Ans A : check the correct syntax for insert into  
upvoted 3 times

A data engineer has realized that the data files associated with a Delta table are incredibly small. They want to compact the small files to form larger files to improve performance.

Which of the following keywords can be used to compact the small files?

A. REDUCE

**B. OPTIMIZE**

C. COMPACTION

D. REPARTITION

E. VACUUM

**Correct Answer: B**

*Community vote distribution*

B (100%)

  **kishanu** Highly Voted 8 months ago

**Selected Answer: B**

OPTIMIZE can be used to club small files into 1 and improve performance.

upvoted 5 times

  **benni\_ale** Most Recent 1 month, 3 weeks ago

**Selected Answer: B**

B is correct

upvoted 1 times

  **UGOTCOOKIES** 4 months, 3 weeks ago

**Selected Answer: B**

OPTIMIZE is the correct answer. Compacting small files using the OPTIMIZE command improves table performance such as by combining multiple small files into larger ones.



upvoted 2 times

  **azure\_bimonster** 5 months ago

**Selected Answer: B**

OPTIMIZE would help in this scenario

upvoted 2 times

  **nedlo** 6 months, 1 week ago

**Selected Answer: B**

Its B <https://docs.databricks.com/en/delta/optimize.html>

upvoted 2 times

  **meow\_akk** 8 months ago

Ans B : optimize is used to compact small files which in turn improves perf

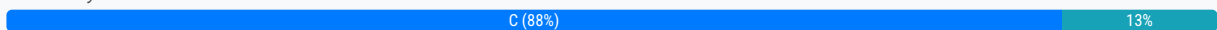
upvoted 3 times

In which of the following file formats is data from Delta Lake tables primarily stored?

- A. Delta
- B. CSV
- C. Parquet**
- D. JSON
- E. A proprietary, optimized format specific to Databricks

**Correct Answer: C**

*Community vote distribution*



  **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: C**

Parquet for data and JSON for metadata  
upvoted 1 times

  **azure\_bimonster** 5 months ago

**Selected Answer: C**

Parquet

## Question #53

Topic 1



Which of the following is stored in the Databricks customer's cloud account?

- A. Databricks web application
- B. Cluster management metadata
- C. Repos
- D. Data**
- E. Notebooks

**Correct Answer: D**



Community vote distribution

D (100%)

  **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: D**

Data is in Data plane  
upvoted 1 times

  **Bob123456** 3 months, 4 weeks ago



Answer should be B  
Because

When the customer sets up a Spark cluster, the cluster virtual machines are deployed in the data plane in the customer's cloud account.  
upvoted 1 times

  **azure\_bimonster** 5 months ago

**Selected Answer: D**

D is correct  
upvoted 2 times

  **bartfto** 5 months, 1 week ago

**Selected Answer: D**

D. Data  
upvoted 1 times

  **meow\_akk** 8 months ago

D. Data  
upvoted 4 times

Which of the following can be used to simplify and unify siloed data architectures that are specialized for specific use cases?

- A. None of these
- B. Data lake
- C. Data warehouse
- D. All of these
- E. Data lakehouse**

**Correct Answer: E**

Community vote distribution

E (100%)

☐ **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: E**

E is correct

upvoted 1 times

☐ **azure\_bimonster** 5 months ago

**Selected Answer: E**

Lakehouse, so E is correct

upvoted 2 times

☐ **kishanu** 8 months ago

**Selected Answer: E**

Data Lakehouse can be used as a single source of truth for multiple specific use cases

upvoted 3 times

A data architect has determined that a table of the following format is necessary:

employeeId	startDate	avgRating
a1	2009-01-06	5.5
a2	2018-11-21	7.1
...	...	...

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?

```
CREATE TABLE IF NOT EXISTS table_name (  
  employeeId STRING,  
A.  startDate DATE,  
  avgRating FLOAT  
)
```

- ```
CREATE OR REPLACE TABLE table_name AS
SELECT
    employeeId STRING,
    startDate DATE,
    avgRating FLOAT
USING DELTA

CREATE OR REPLACE TABLE table_name WITH COLUMNS (
    employeeId STRING,
    startDate DATE,
    avgRating FLOAT
) USING DELTA

CREATE TABLE table_name AS
SELECT
    employeeId STRING,
    startDate DATE,
    avgRating FLOAT

CREATE OR REPLACE TABLE table_name (
    employeeId STRING,
    startDate DATE,
    avgRating FLOAT
)
```
- E.**

**Correct Answer: E**



Community vote distribution

E (100%)

 **meow\_akk**  8 months ago

E is correct you dont need to specify Delta as its the default storage format for tables.

upvoted 5 times

 **benni\_ale**  1 month, 3 weeks ago

**Selected Answer: E**

E is correct

upvoted 1 times

 **Stemix** 4 months, 3 weeks ago

A and E have both correct syntax, but the question mentioned "regardless of whether a table already exists with this name". Hence the correct answer is E


upvoted 2 times

 **azure\_bimonster** 5 months ago

**Selected Answer: E**

E is correct option

upvoted 2 times

 **bartfto** 5 months, 1 week ago

**Selected Answer: E**

E. correct

upvoted 2 times





A data engineer has a Python notebook in Databricks, but they need to use SQL to accomplish a specific task within a cell. They still want all of the other cells to use Python without making any changes to those cells.

Which of the following describes how the data engineer can use SQL within a cell of their Python notebook?

- A. It is not possible to use SQL in a Python notebook
- B. They can attach the cell to a SQL endpoint rather than a Databricks cluster
- C. They can simply write SQL syntax in the cell
- D. They can add %sql to the first line of the cell**
- E. They can change the default language of the notebook to SQL

**Correct Answer: D**

*Community vote distribution*


D (100%)

 **azure\_bimonster** 5 months ago

**Selected Answer: D**

Magic command % can be used to switch the language, so D is correct


upvoted 1 times

 **bartfto** 5 months, 1 week ago

**Selected Answer: D**

D. Correct. Use %sql magic in first line.


upvoted 1 times

 **Lavpak** 6 months, 4 weeks ago

**Selected Answer: D**

Use magic command %sql

upvoted 3 times

 **MFEST** 7 months, 3 weeks ago

correct answer D

upvoted 3 times

Which of the following SQL keywords can be used to convert a table from a long format to a wide format?

- A. TRANSFORM
- B. PIVOT**
- C. SUM
- D. CONVERT
- E. WHERE

**Correct Answer: B**

*Community vote distribution*

  **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: B**

B is correct

upvoted 1 times

  **azure\_bimonster** 5 months ago

**Selected Answer: B**

Answer is B

upvoted 1 times

  **AndreFR** 6 months ago

**Selected Answer: B**

<https://docs.databricks.com/en/sql/language-manual/sql-ref-syntax-qry-select-pivot.html>

“Pivot” transforms the rows of the table\_reference by rotating unique values of a specified column list into separate columns.



SYNTAX :

```
table_reference PIVOT ( { aggregate_expression [ [ AS ] agg_column_alias ] } [, ...]
FOR column_list IN ( expression_list ) )
```

```
column_list
{ column_name |
( column_name [, ...] ) }
```

```
expression_list
{ expression [ AS ] [ column_alias ] |
{ ( expression [, ...] ) [ AS ] [ column_alias] } [, ...] }
```

upvoted 1 times

  **athu07** 7 months, 4 weeks ago

**Selected Answer: B**

PIVOT is correct.

upvoted 2 times

  **meow\_akk** 8 months ago

PIVOT is correct.

upvoted 3 times

Which of the following describes a benefit of creating an external table from Parquet rather than CSV when using a CREATE TABLE AS SELECT statement?


- A. Parquet files can be partitioned
- B. CREATE TABLE AS SELECT statements cannot be used on files
- C. Parquet files have a well-defined schema**
- D. Parquet files have the ability to be optimized
- E. Parquet files will become Delta tables

**Correct Answer: D**

Community vote distribution

C (88%)

6%

  **MDWPartners** 3 weeks, 2 days ago

**Selected Answer: C**

The keywords are "CREATE TABLE AS SELECT "  
upvoted 1 times

  **benni\_ale** 1 month, 3 weeks ago



**Selected Answer: C**

C is correct  
upvoted 1 times

  **UGOTCOOKIES** 4 months, 3 weeks ago



**Selected Answer: C**

CREATE TABLE AS SELECT adopts the schema details from the source. Parquet files have a defined schema.  
upvoted 2 times

  **bartfto** 5 months, 1 week ago

**Selected Answer: C**

C. Parquet has well defined schema online csv  
upvoted 1 times

  **Garyn** 5 months, 3 weeks ago

**Selected Answer: C**

C. Parquet files have a well-defined schema.

Explanation:

Parquet files inherently store metadata about the schema within the files themselves, allowing for a well-defined schema. This schema information includes data types, column names, and other structural information. When creating an external table from Parquet, this schema is retained, providing a structured and well-defined format for the data. It ensures consistency and enables more efficient processing, query optimization, and compatibility across various systems or tools that work with the Parquet format.

This structured schema within Parquet files offers advantages in terms of data integrity, ease of data processing, and compatibility, making it a beneficial choice over CSV, which lacks inherent schema information and might need additional handling or inference of schema during data ingestion.

upvoted 1 times

  **AndreFR** 6 months ago

**Selected Answer: B**

The key word here is : CREATE TABLE AS SELECT

not A : partitioning is not relevant in a create table as statement because the data will be created in a delta table



not C : Parquet schema is not well defined and there can be parquet files with multiple schema in a folder

not D : Parquet are already optimized and are not relevant in a create table as statement because the data will be created in a delta table

not E : both CSV & Parquet will become delta tables in a create table as statement

B : correct answer by elimination

upvoted 1 times

  **nedlo** 6 months, 1 week ago

**Selected Answer: D**

I disagree i think its D. Schema can be inferred from CSV as well, but CSV cannot provide same optimizations as Parquet



upvoted 1 times

  **FastEddie** 7 months, 3 weeks ago

**Selected Answer: C**

CTAS - CTAS automatically infer schema information from query results and do not support manual schema declaration. This means that CTAS statements are useful for external data ingestion from sources with well-defined schema, such as Parquet files and tables. CTAS statements also do not support specifying additional file options.

upvoted 4 times

  **kishore1980** 7 months, 3 weeks ago

**Selected Answer: C**

C is the correct option

upvoted 2 times

  **anandpsg101** 7 months, 4 weeks ago

**Selected Answer: C**

c is correct

upvoted 2 times


  **meow\_akk** 8 months ago

Ans : C

<https://www.databricks.com/glossary/what-is-parquet#:~:text=Columnar%20storage%20like%20Apache%20Parquet,compared%20to%20row%2Doriented%20databases.>

Columnar storage like Apache Parquet is designed to bring efficiency compared to row-based files like CSV. When querying, columnar storage you can skip over the non-relevant data very quickly. As a result, aggregation queries are less time-consuming compared to row-oriented databases.

upvoted 3 times

  **kbaba101** 8 months ago

Question #59

Topic 1

A data engineer wants to create a relational object by pulling data from two tables. The relational object does not need to be used by other data engineers in other sessions. In order to save on storage costs, the data engineer wants to avoid copying and storing physical data.

Which of the following relational objects should the data engineer create?

- A. Spark SQL Table
- B. View
- C. Database
- D. Temporary view**
- E. Delta Table

**Correct Answer: D**

Community vote distribution

D (100%)

  **meow\_akk** Highly Voted 8 months ago

D is correct.

Temp view : session based

Create temp view view\_name as query

All these are termed as session ended:



Opening a new notebook

Detaching and reattaching a cluster

Installing a python package

Restarting a cluster

upvoted 7 times

  **Garyn** Most Recent 5 months, 3 weeks ago

Selected Answer: D

D. Temporary view

Explanation:

Temporary View: A temporary view in database systems like Apache Spark provides a temporary and ephemeral representation of data based on an SQL query's result set. It exists for the duration of a Spark session and is not persisted to storage. Similar to a regular view, a temporary view allows the data engineer to define a logical schema by pulling and combining data from multiple tables using SQL queries, but it does not store any physical data on disk. Temporary views are suitable when there's no need for long-term storage of the combined data and are helpful for immediate analysis or processing within the current session without incurring storage costs.

upvoted 2 times

  **AndreFR** 6 months ago

Selected Answer: D

<https://docs.databricks.com/en/sql/language-manual/sql-ref-syntax-ddl-create-view.html>

Should be a view or temporary view to avoid copying and storing data.



Does not need to be used by other data engineers in other sessions. So this view should be temporary. TEMPORARY views are visible only to session that created them and are dropped when the session ends.

upvoted 1 times

  **Huroye** 7 months ago

Answer is D. key phrase is "...does not need to be used by other data engineers in other sessions..."

upvoted 2 times

  **kishore1980** 7 months, 3 weeks ago

Selected Answer: D

D is right option

upvoted 1 times

Question #60

Topic 1

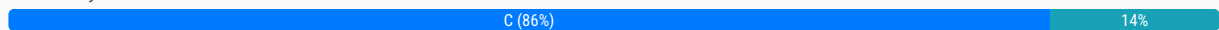
A data analyst has developed a query that runs against Delta table. They want help from the data engineering team to implement a series of tests to ensure the data returned by the query is clean. However, the data engineering team uses Python for its tests rather than SQL.


Which of the following operations could the data engineering team use to run the query and operate with the results in PySpark?

- A. `SELECT * FROM sales`
- B. `spark.delta.table`
- C. `spark.sql`**
- D. There is no way to share data between PySpark and SQL.
- E. `spark.table`

**Correct Answer:** C

Community vote distribution



  **kishanu** Highly Voted 8 months ago

**Selected Answer: C**

spark.sql() should be used to execute a SQL query with Pyspark  
spark.table() can only be used to load a table and not run a query.  
upvoted 6 times

  **benni\_ale** Most Recent 1 month, 3 weeks ago

**Selected Answer: E**

I am not sure whether it is C or E . I see majority went for E but you can still query your data with spark.table by using purely pyspark syntax . I don't see any part of the question specifying you HAVE to use SQL syntax.  
upvoted 1 times

  **meow\_akk** 8 months ago

C is correct

EG :

```
from pyspark.sql import SparkSession
```

```
spark = SparkSession.builder.getOrCreate()
```

```
df = spark.sql("SELECT * FROM sales")
```

```
print(df.count())
```

upvoted 3 times





Which of the following commands will return the number of null values in the member\_id column?

- A. `SELECT count(member_id) FROM my_table;`
- B. `SELECT count(member_id) - count_null(member_id) FROM my_table;`
- C. `SELECT count_if(member_id IS NULL) FROM my_table;`**
- D. `SELECT null(member_id) FROM my_table;`
- E. `SELECT count_null(member_id) FROM my_table;`

**Correct Answer: C**

Community vote distribution



C (100%)

  **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: C**

C is correct



upvoted 1 times

  **bartfto** 5 months, 1 week ago

**Selected Answer: C**

C: There are no 'null' and 'count\_null' functions in SparkSQL

upvoted 1 times

  **55f31c8** 6 months, 3 weeks ago

**Selected Answer: C**

[https://docs.databricks.com/en/sql/language-manual/functions/count\\_if.html](https://docs.databricks.com/en/sql/language-manual/functions/count_if.html)

upvoted 4 times

  **meow\_akk** 8 months ago

Ans C :

<https://docs.databricks.com/en/sql/language-manual/functions/count.html>

Returns

A BIGINT.

If \* is specified also counts row containing NULL values.

If expr are specified counts only rows for which all expr are not NULL.

If DISTINCT duplicate rows are not counted.

upvoted 3 times

  **kishanu** 8 months ago

**Selected Answer: C**

count\_if() can be used in this scenario

upvoted 3 times

A data engineer needs to apply custom logic to identify employees with more than 5 years of experience in array column employees in table stores. The custom logic should create a new column exp\_employees that is an array of all of the employees with more than 5 years of experience for each row. In order to apply this custom logic at scale, the data engineer wants to use the FILTER higher-order function.

Which of the following code blocks successfully completes this task?

- A. `SELECT  
 store_id,  
 employees,  
 FILTER (employees, i -> i.years_exp > 5) AS exp_employees  
FROM stores;`
- B. `SELECT  
 store_id,  
 employees,  
 FILTER (exp_employees, years_exp > 5) AS exp_employees  
FROM stores;`
- C. `SELECT  
 store_id,  
 employees,  
 FILTER (employees, years_exp > 5) AS exp_employees  
FROM stores;`
- D. `SELECT  
 store_id,  
 employees,  
 CASE WHEN employees.years_exp > 5 THEN employees  
 ELSE NULL  
 END AS exp_employees  
FROM stores;`
- E. `SELECT  
 store_id,  
 employees,  
 FILTER (exp_employees, i -> i.years_exp > 5) AS exp_employees  
FROM stores;`

**Correct Answer: A**

*Community vote distribution*

A (100%)

🗄️ 👤 **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: A**

A is correct

upvoted 1 times

🗄️ 👤 **AndreFR** 6 months ago

**Selected Answer: A**

B & E incorrect : source is employees not exp\_employees

D incorrect : does not use FILTER higher-order function)

C incorrect : syntax error

A : correct by elimination & based on <https://docs.databricks.com/en/sql/language-manual/functions/filter.html#examples>

upvoted 1 times

🗄️ 👤 **kz\_data** 6 months, 2 weeks ago

**Selected Answer: A**

A is correct

upvoted 2 times

🗄️ 👤 **55f31c8** 6 months, 3 weeks ago

**Selected Answer: A**

<https://docs.databricks.com/en/sql/language-manual/functions/filter.html>

upvoted 3 times

🗄️ 👤 **meow\_akk** 8 months ago

A is correct.

upvoted 4 times

A data engineer has a Python variable `table_name` that they would like to use in a SQL query. They want to construct a Python code block that will run the query using `table_name`.

They have the following incomplete code block:

```
____(f"SELECT customer_id, spend FROM {table_name}")
```

Which of the following can be used to fill in the blank to successfully complete the task?

- A. `spark.delta.sql`
- B. `spark.delta.table`
- C. `spark.table`
- D. `dbutils.sql`
- E. `spark.sql`**

**Correct Answer: E**

*Community vote distribution*

E (100%)

 **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: E**

E is correct

upvoted 1 times

 **azure\_bimonster** 5 months ago

**Selected Answer: E**

E is correct

upvoted 1 times

 **meow\_akk** 8 months ago

E is correct you use `spark.sql` to execute python comamands

upvoted 3 times

A data engineer has created a new database using the following command:

```
CREATE DATABASE IF NOT EXISTS customer360;
```

In which of the following locations will the `customer360` database be located?

- A. `dbfs:/user/hive/database/customer360`
- B. `dbfs:/user/hive/warehouse`
- C. `dbfs:/user/hive/customer360`
- D. More information is needed to determine the correct response**

E. dbfs:/user/hive/database

**Correct Answer: D**

Community vote distribution

B (83%)


D (17%)

 **kbaba101** Highly Voted 8 months ago

B

B. dbfs:/user/hive/warehouse Thereby showing "dbfs:/user/hive/warehouse/customer360.db

upvoted 8 times

 **aspix82** Most Recent 3 weeks, 6 days ago

B. B is "default"

upvoted 1 times

 **jskibick** 1 month ago

Selected Answer: D

D is correct. We do not know if this is a Unity Catalog enabled database. If so it would be created in default location of catalog as managed table. Therefore too little info to answer.


upvoted 2 times

 **benni\_ale** 1 month, 3 weeks ago

Selected Answer: B

B is correct

upvoted 1 times

 **Bob123456** 3 months, 3 weeks ago

While usage schema and database is interchangeable, schema is preferred. Option B is correct

upvoted 1 times

 **UGOTCOOKIES** 4 months, 3 weeks ago

Selected Answer: B

Creating tables without using the LOCATION keyword to specify a location will create the table (a managed table) in the default directory which is dbfs:/user/hive/warehouse

<https://docs.databricks.com/en/dbfs/root-locations.html>

upvoted 1 times

 **Garyn** 5 months, 3 weeks ago

Selected Answer: B


B. dbfs:/user/hive/warehouse

Explanation:

In Databricks, the default location for databases created in the Hive Metastore is often under the warehouse directory. The CREATE DATABASE command usually creates the metadata entry for the database in the Hive Metastore, but it doesn't directly create the physical database directory within DBFS (Databricks File System).

The exact path structure may differ based on configuration or settings in the Databricks environment, but generally, the warehouse directory is where Hive databases' metadata resides. The physical data within the database will be stored in DBFS, but the metadata for the customer360 database should be within the warehouse directory in Hive Metastore.

upvoted 1 times

 **kz\_data** 6 months, 2 weeks ago

Selected Answer: B

B is correct

upvoted 1 times

 **Huroye** 7 months ago

correct answer is B. dbfs:/user/hive/warehouse. All managed objects are stored in the default location unless specified.

upvoted 1 times

🗨️ **anandpsg101** 7 months, 4 weeks ago

**Selected Answer: B**

b is correct  
upvoted 1 times

🗨️ **SD5713** 7 months, 4 weeks ago

**Selected Answer: B**

dbfs:/user/hive/warehouse - which is the default location  
upvoted 2 times

🗨️ **meow\_akk** 8 months ago

Ans A : <https://community.databricks.com/t5/data-engineering/database-within-a-database-in-databricks/td-p/20731#:~:text=The%20default%20location%20of%20a, and%20Table%20location%20are%20independent.>  
The default location of a database will be in the /user/hive/warehouse/<dbname>. Irrespective of the location of the database the tables in the database can have different locations and they can be specified at the time of creation. Database location and Table location are independent.  
upvoted 1 times

🗨️ **SD5713** 7 months, 4 weeks ago

dbfs:/user/hive/warehouse - which is the default location  
upvoted 3 times

🗨️ **kishanu** 8 months ago

**Selected Answer: B**

dbfs:/user/hive/warehouse - which is the default location of any object created  
upvoted 3 times

## Question #65

Topic 1

A data engineer is attempting to drop a Spark SQL table my\_table and runs the following command:

```
DROP TABLE IF EXISTS my_table;
```

After running this command, the engineer notices that the data files and metadata files have been deleted from the file system.

Which of the following describes why all of these files were deleted?

**A. The table was managed**

B. The table's data was smaller than 10 GB

C. The table's data was larger than 10 GB



D. The table was external

E. The table did not have a location

**Correct Answer: A**

Community vote distribution

A (100%)

  **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: A**

A is correct

upvoted 1 times

  **UGOTCOOKIES** 4 months, 3 weeks ago



**Selected Answer: A**

Two types of tables, managed and external. Both table types are treated the same, except when the table is dropped.

For a managed table the data is stored in the managed storage location that is configured to the meta store. By default this is dbfs:/user/hive/warehouse. When the table is dropped the meta data and the underlying data is deleted.

For external tables the data is stored in a cloud storage location outside of the managed storage location. The underlying data is retained when an external table is dropped, only the metadata is dropped.

upvoted 1 times

  **Garyn** 5 months, 3 weeks ago

**Selected Answer: A**

A. The table was managed.

Explanation:



In Spark SQL, when a table is managed (or internal), both the metadata that contains information about the table and the actual data files associated with the table are managed by the SQL engine.

The DROP TABLE command, when used on a managed table, deletes not only the metadata but also the underlying data files associated with that table from the file system.

When a managed table is dropped, it removes all information about the table, including metadata and data files, leading to the deletion of both the metadata and data files from the file system.

Options B, C, D, and E don't specifically relate to why the data files and metadata files were deleted. The fact that the table was managed (or internal) is the reason for the removal of both the metadata and data files when the table was dropped using the DROP TABLE command.

upvoted 1 times

  **kz\_data** 6 months, 2 weeks ago

**Selected Answer: A**

A is correct

upvoted 2 times

  **meow\_akk** 8 months ago

A is correct, managed tables files and metadata are managed by metastore and will be deleted when the table is dropped. while external table the metadata is stored in an external location. hence when an external table is dropped you clear off only the metadata and the files (data) remain

upvoted 4 times

A data engineer that is new to using Python needs to create a Python function to add two integers together and return the sum?

Which of the following code blocks can the data engineer use to complete this task?

- A. 

```
function add_integers(x, y):  
    return x + y
```
- B. 

```
function add_integers(x, y):  
    x + y
```
- C. 

```
def add_integers(x, y):  
    print(x + y)
```
- D. 

```
def add_integers(x, y):  
    return x + y
```**
- E. 

```
def add_integers(x, y):  
    x + y
```

**Correct Answer:** D

Community vote distribution

D (100%)

 **UGOTCOOKIES** 4 months, 3 weeks ago

**Selected Answer: D**

Python functions start with the def keyword followed by the function name. Function also ends with the return keyword.


upvoted 1 times

 **azure\_bimonster** 5 months ago

**Selected Answer: D**

D is to choose here


upvoted 1 times

 **kz\_data** 6 months, 2 weeks ago

**Selected Answer: D**

D is correct

upvoted 2 times

 **55f31c8** 6 months, 3 weeks ago

**Selected Answer: D**

D : <https://www.geeksforgeeks.org/python-functions/>

upvoted 3 times

 **Syd** 7 months, 3 weeks ago

D is correct.

[https://www.w3schools.com/python/python\\_functions.asp](https://www.w3schools.com/python/python_functions.asp)

upvoted 2 times

 **meow\_akk** 8 months ago

D is correct. if you get this answer wrong you need to learn the basics of python.

upvoted 3 times



Question #67

Topic 1


In which of the following scenarios should a data engineer use the MERGE INTO command instead of the INSERT INTO command?

- A. When the location of the data needs to be changed
- B. When the target table is an external table
- C. When the source table can be deleted
- D. When the target table cannot contain duplicate records**
- E. When the source is not a Delta table

**Correct Answer:** D

*Community vote distribution*

D (100%)

 **fifirifi** 3 months, 1 week ago

**Selected Answer: D**

correct answer: D

explanation: The MERGE INTO command is used when you need to perform both insertions and updates (or deletes) in one operation based on whether a match exists. It is particularly useful for maintaining up-to-date data and ensuring there are no duplicate records in the target table. This is often referred to as an "upsert" operation (update + insert). When the target table needs to be kept free of duplicate records, and there need to update existing records or insert new ones based on some matching condition, MERGE INTO is the appropriate command. The INSERT INTO command, on the other hand, is used to add new records to a table without regard for whether they duplicate existing records. Options B, C, and E do not specifically require the use of MERGE INTO. Therefore, D is the correct answer.

upvoted 1 times

 **UGOTCOOKIES** 4 months, 3 weeks ago

**Selected Answer: D**

MERGE INTO you can upsert (update insert) data from a source table, view or dataframe into the target table. Merge operation allows updates, inserts and deletes to be completed in a single atomic transaction. The main benefit of using the MERGE INTO is to avoid duplicates but does not inherently remove duplicates.


upvoted 1 times

 **azure\_bimonster** 5 months ago

**Selected Answer: D**

D is answer here

upvoted 1 times

 **kz\_data** 6 months, 2 weeks ago

**Selected Answer: D**

D is correct

upvoted 1 times

 **meow\_akk** 8 months ago

Ans D : With merge , you can avoid inserting the duplicate records. The dataset containing the new logs needs to be deduplicated within itself. By the SQL semantics of merge, it matches and deduplicates the new data with the existing data in the table, but if there is duplicate data within the new dataset, it is inserted.

<https://docs.databricks.com/en/delta/merge.html#:~:text=With%20merge%20you%20can%20avoid%20inserting%20the%20duplicate%20records.&text=The%20dataset%20containing%20the%20new,new%20dataset%20it%20is%20inserted.>

upvoted 1 times

## Question #68

## Topic 1

A data engineer is working with two tables. Each of these tables is displayed below in its entirety.

### sales

| customer_id | spend  | units |
|-------------|--------|-------|
| a1          | 28.94  | 7     |
| a3          | 874.12 | 23    |
| a4          | 8.99   | 1     |

### favorite\_stores

| customer_id | store_id |
|-------------|----------|
| a1          | s1       |
| a2          | s1       |
| a4          | s2       |

The data engineer runs the following query to join these tables together:

```

SELECT
    sales.customer_id,
    sales.spend,
    favorite_stores.store_id
FROM sales
LEFT JOIN favorite_stores
ON sales.customer_id = favorite_stores.customer_id;

```

Which of the following will be returned by the above query?

A.

| customer_id | spend | store_id |
|-------------|-------|----------|
| a1          | 28.94 | s1       |
| a4          | 8.99  | s2       |

B.

| customer_id | spend | units | store_id |
|-------------|-------|-------|----------|
| a1          | 28.94 | 7     | s1       |
| a4          | 8.99  | 1     | s2       |

C.

| customer_id | spend  | store_id |
|-------------|--------|----------|
| a1          | 28.94  | s1       |
| a3          | 874.12 | NULL     |
| a4          | 8.99   | s2       |

D.

| customer_id | spend  | store_id |
|-------------|--------|----------|
| a1          | 28.94  | s1       |
| a2          | NULL   | s1       |
| a3          | 874.12 | NULL     |
| a4          | 8.99   | s2       |



E.

| customer_id | spend | store_id |
|-------------|-------|----------|
| a1          | 28.94 | s1       |
| a2          | NULL  | s1       |
| a4          | 8.99  | s2       |

**Correct Answer: D**

Community vote distribution



C (100%)

  **kz\_data** 6 months, 2 weeks ago

**Selected Answer: C**

C is correct



upvoted 2 times

  **nedlo** 6 months, 2 weeks ago

**Selected Answer: C**

C is correct answer

upvoted 1 times

  **55f31c8** 6 months, 3 weeks ago

**Selected Answer: C**

The LEFT JOIN keyword returns all records from the left table (table1), and the matching records from the right table (table2). The result is 0 records from the right side, if there is no match.

upvoted 3 times

  **Blacknight99** 7 months ago

**Selected Answer: C**

C is the correct answer

upvoted 1 times

  **Huroye** 7 months ago

The answer is C. this is a Left Join. In this case you show everything on the left side regardless of whether they appear on the right. When it doesn't appear on the right you represent that with a Null. So, for a3, store id is null.

upvoted 2 times

  **dev\_soumya369** 7 months, 1 week ago

C is the correct answer.

In a LEFT JOIN, all the records from the left table are included, and only the matching records from the right table are added. In this case, "a1" and "a4" from the left table (favorite\_stores) match with "a1" and "a4" from the right table (sales). So, these matching records are fetched. Additionally, all the records from the left table, including "a3," are included. Since "a3" has no corresponding store\_id in the right table, the store\_id for "a3" will be NULL. Therefore, after the LEFT JOIN, the result will include "a1," "a3" (with a NULL store\_id), and "a4."

upvoted 1 times

  **mokrani** 7 months, 2 weeks ago

C is correct.

please refer to this simple blog if any confusion regarding JOINS

<https://sql.sh/cours/jointures>

upvoted 1 times

  **anandpsg101** 7 months, 4 weeks ago

**Selected Answer: C**

c is corret

upvoted 2 times

  **meow\_akk** 8 months ago

Ans C: Left join only keeps left recs and only the matching recs from Right table.

in other words : the left table is preserved as is.

upvoted 1 times

  **kishanu** 8 months ago

**Selected Answer: C**

A typical LEFT JOIN scenario

upvoted 1 times

  **[Removed]** 8 months ago

**Selected Answer: C**

The LEFT JOIN keyword returns all records from the left table, even if there are no matches in the right table.

upvoted 3 times

A data engineer needs to create a table in Databricks using data from a CSV file at location /path/to/csv.

They run the following command:

```
CREATE TABLE new_table  
  
_____  
OPTIONS (  
  header = "true",  
  delimiter = "|" )  
LOCATION "path/to/csv"
```



Which of the following lines of code fills in the above blank to successfully complete the task?

- A. None of these lines of code are needed to successfully complete the task
- B. USING CSV**
- C. FROM CSV
- D. USING DELTA
- E. FROM "path/to/csv"

**Correct Answer:** B

*Community vote distribution*

B (100%)



  **fifirifi** 3 months, 1 week ago

**Selected Answer: B**

correct answer: B

explanation: To create a table in Databricks using data from a CSV file, the correct syntax after specifying the table name and schema (if applicable) would be to use the USING CSV clause to define the format of the source data. This clause tells Databricks that the data source format is CSV. The command would typically look



upvoted 2 times

  **Bob123456** 3 months, 4 weeks ago

I have a question

Why can option using delta



upvoted 1 times

  **kz\_data** 6 months, 2 weeks ago

**Selected Answer: B**

B is correct

upvoted 1 times

  **55f31c8** 6 months, 3 weeks ago

**Selected Answer: B**

<https://docs.databricks.com/en/sql/language-manual/sql-ref-syntax-ddl-create-table-using.html#parameters>

upvoted 3 times

  **meow\_akk** 8 months ago

Ans B : Using csv is correct. that is the correct syntax

upvoted 3 times

  **richary** 8 months ago

## Question #70

Topic 1

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.readStream
  .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 process all of the available data in as many batches as required, which of the following lines of code should the data engineer use to fill in the blank?

- A. processingTime(1)
- B. trigger(availableNow=True)**
- C. trigger(parallelBatch=True)
- D. trigger(processingTime="once")
- E. trigger(continuous="once")

**Correct Answer: B**

Community vote distribution

B (100%)

🗳️ **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: B**

b is ok

upvoted 1 times

🗳️ **fifirifi** 3 months, 1 week ago

**Selected Answer: B**

correct answer: B

explanation: In Structured Streaming, if a data engineer wants to process all the available data in as many batches as required without any explicit trigger interval, they can use the option `trigger(availableNow=True)`. This feature, `availableNow`, is used to specify that the query should process all the data that is available at the moment and not wait for more data to arrive.

upvoted 2 times

🗳️ **AndreFR** 6 months ago

**Selected Answer: B**

it's the only answer with a correct syntax

upvoted 1 times

🗳️ **55f31c8** 6 months, 3 weeks ago

**Selected Answer: B**

<https://spark.apache.org/docs/latest/api/python/reference/pyspark.ss/api/pyspark.sql.streaming.DataStreamWriter.trigger.html>

upvoted 2 times

🗳️ **kbaba101** 7 months, 4 weeks ago

B

`availableNow` bool, optional

if set to True, set a trigger that processes all available data in multiple batches then terminates the query. Only one trigger can be set.

upvoted 4 times

🗳️ **meow\_akk** 8 months ago

sorry Ans is B : <https://stackoverflow.com/questions/71061809/trigger-availablenow-for-delta-source-streaming-queries-in-pyspark-databrick>

for batch we use available now

upvoted 4 times

🗳️ **meow\_akk** 8 months ago

Correct Ans is D :

`%python`

```
spark.readStream.format("delta").load("<delta_table_path>")
.writeStream
.format("delta")
.trigger(processingTime='5 seconds') #Added line of code that defines .trigger processing time.
.outputMode("append")
.option("checkpointLocation", "<checkpoint_path>")
.options(**writeConfig)
.start()
```

<https://kb.databricks.com/streaming/optimize-streaming-transactions-with-trigger>

upvoted 1 times

A data engineer has developed a data pipeline to ingest data from a JSON source using Auto Loader, but the engineer has not provided any type inference or schema hints in their pipeline. Upon reviewing the data, the data engineer has noticed that all of the columns in the target table are of the string type despite some of the fields only including float or boolean values.

Which of the following describes why Auto Loader inferred all of the columns to be of the string type?

- A. There was a type mismatch between the specific schema and the inferred schema
- B. JSON data is a text-based format**
- C. Auto Loader only works with string data
- D. All of the fields had at least one null value
- E. Auto Loader cannot infer the schema of ingested data

**Correct Answer:** B

*Community vote distribution*

B (100%)



AndreFR 6 months ago

Selected Answer: B

<https://docs.databricks.com/en/ingestion/auto-loader/schema.html#how-does-auto-loader-schema-inference-work>

By default, Auto Loader schema inference seeks to avoid schema evolution issues due to type mismatches. For formats that don't encode data types (JSON and CSV), Auto Loader infers all columns as strings (including nested fields in JSON files).

upvoted 1 times

nedlo 6 months, 1 week ago

## Question #72

Topic 1

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.

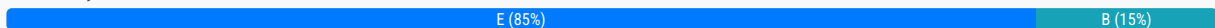
The table is configured to run in Development mode using the Continuous Pipeline Mode.

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 persist until the pipeline is shut down.
- C. All datasets will be updated once and the pipeline will persist without any processing. The compute resources will persist but go unused.
- 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 at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing.**

Correct Answer: B

Community vote distribution



meow\_akk Highly Voted 8 months ago

Ans E : Development and production modes

You can optimize pipeline execution by switching between development and production modes. Use the Delta Live Tables Environment Toggle buttons in the Pipelines UI to switch between these two modes. By default, pipelines run in development mode.

When you run your pipeline in development mode, the Delta Live Tables system does the following:

Reuses a cluster to avoid the overhead of restarts. By default, clusters run for two hours when development mode is enabled. You can change this with the `pipelines.clusterShutdown.delay` setting in the Configure your compute settings.

Disables pipeline retries so you can immediately detect and fix errors.

In production mode, the Delta Live Tables system does the following:

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.

<https://docs.databricks.com/en/delta-live-tables/updates.html#optimize-execution>



upvoted 9 times

benni\_ale Most Recent 1 month, 3 weeks ago

Selected Answer: E

because the cluster actually persists differently from B

upvoted 1 times

  **Garyn** 5 months, 3 weeks ago

**Selected Answer: E**

E. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing



Explanation:

In Development mode, Delta Live Tables persistently updates datasets at set intervals. The pipeline continuously processes incoming data until manually stopped or shut down.

Compute resources, including the cluster used for processing, persist without automatic restarts or retries (as it is the behavior in Developer mode). This persistence allows for ongoing processing of data, enabling additional testing or continued data processing until the pipeline is manually shut down.

Therefore, option E accurately captures the behavior expected in Development mode, emphasizing the continuous update of datasets and the persistence of compute resources until the pipeline is manually terminated.

upvoted 2 times

  **kz\_data** 6 months, 2 weeks ago

**Selected Answer: E**

E seems the correct answer

upvoted 2 times

  **nedlo** 6 months, 2 weeks ago

**Selected Answer: B**

Why E? It persists with same functionality as was before, not for "additional testing"?

upvoted 2 times

  **AndreFR** 6 months ago



because "The table is configured to run in Development mode" when tables are set in dev mode, "The compute resources will persist to allow for additional testing."

upvoted 1 times

  **AndreFR** 6 months ago

So correct answer is E

upvoted 1 times



  **55f31c8** 6 months, 3 weeks ago

**Selected Answer: E**

<https://docs.databricks.com/en/delta-live-tables/updates.html#continuous-vs-triggered-pipeline-execution>

<https://docs.databricks.com/en/delta-live-tables/testing.html#use-development-mode-to-run-pipeline-updates>



upvoted 2 times

  **anandpsg101** 7 months, 4 weeks ago

**Selected Answer: E**

E is correct

upvoted 2 times

  **SD5713** 8 months ago

**Selected Answer: E**

E. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing

upvoted 2 times

Which of the following data workloads will utilize a Gold table as its source?

- A. A job that enriches data by parsing its timestamps into a human-readable format
- B. A job that aggregates uncleaned data to create standard summary statistics
- C. A job that cleans data by removing malformed records
- D. A job that queries aggregated data designed to feed into a dashboard**
- E. A job that ingests raw data from a streaming source into the Lakehouse

**Correct Answer: D**

Community vote distribution

D (100%)

☐ **55f31c8** 6 months, 3 weeks ago

**Selected Answer: D**

<https://docs.databricks.com/en/lakehouse/medallion.html#power-analytics-with-the-gold-layer>

upvoted 1 times

☐ **meow\_akk** 8 months ago

D is correct : std medallion arch

upvoted 3 times

Which of the following must be specified when creating a new Delta Live Tables pipeline?

- A. A key-value pair configuration
- B. The preferred DBU/hour cost
- C. A path to cloud storage location for the written data
- D. A location of a target database for the written data
- E. At least one notebook library to be executed**

**Correct Answer: E**

Community vote distribution

E (63%)

C (37%)

☐ **Shinigami76** 1 week, 1 day ago

C, just tested on databricks DLT

upvoted 1 times

☐ **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: E**

tbh C is correct as well but the question is probably hinting for E

upvoted 1 times

🗨️ **BigMF** 3 months ago

**Selected Answer: C**

Per Databricks documentation (see below), you need to select a destination for datasets published by the pipeline, either the Hive metastore or Unity Catalog. I think A is incorrect because it uses the term "Notebook Library" and not just "Notebook".  
Databricks doc: <https://docs.databricks.com/en/delta-live-tables/tutorial-pipelines.html>

upvoted 1 times

🗨️ **Stemix** 4 months, 3 weeks ago

**Selected Answer: E**

Correct answer is E. storage location is optional.

"(Optional) Enter a Storage location for output data from the pipeline. The system uses a default location if you leave Storage location empty"

upvoted 4 times

🗨️ **azure\_bimonster** 5 months ago

**Selected Answer: E**

As per Pipeline creating steps, choosing a Notebook is mandatory whereas specifying a location is optional. I would go with answer E

upvoted 1 times

🗨️ **Azure\_2023** 5 months ago

**Selected Answer: E**

<https://docs.databricks.com/en/delta-live-tables/tutorial-pipelines.html>

E. The only non-optional selection is a notebook

upvoted 2 times

🗨️ **Garyn** 5 months, 3 weeks ago

**Selected Answer: E**

E. At least one notebook library to be executed.

Explanation:

<https://docs.databricks.com/en/delta-live-tables/tutorial-pipelines.html>

Delta Live Tables pipelines execute notebook libraries as part of their operations. These notebooks contain the logic, code, or instructions defining the data processing steps, transformations, or actions to be performed within the pipeline.

Specifying at least one notebook library to be executed is crucial when creating a new Delta Live Tables pipeline, as it defines the sequence of operations and the logic to be executed on the data within the pipeline, aligning with the documentation provided.

upvoted 2 times

🗨️ **saaaaaa** 6 months ago

**Selected Answer: E**

This should be E. As per the link <https://docs.databricks.com/en/delta-live-tables/tutorial-pipelines.html>

Create a pipeline

Click Jobs Icon Workflows in the sidebar, click the Delta Live Tables tab, and click Create Pipeline.

Give the pipeline a name and click File Picker Icon to select a notebook.

Select Triggered for Pipeline Mode.

(Optional) Enter a Storage location for output data from the pipeline. The system uses a default location if you leave Storage location empty.

(Optional) Specify a Target schema to publish your dataset to the Hive metastore or a Catalog and a Target schema to publish your dataset to Unity Catalog. See Publish datasets.

(Optional) Click Add notification to configure one or more email addresses to receive notifications for pipeline events. See Add email notifications for pipeline events.

Click Create.

upvoted 2 times

🗨️ **55f31c8** 6 months, 3 weeks ago

**Selected Answer: C**

<https://docs.databricks.com/en/delta-live-tables/index.html#what-is-a-delta-live-tables-pipeline>

upvoted 1 times

🗨️ **Huroye** 7 months ago

The correct answer is E. DLT tables needs a notebook where you have to specify the processing info  
upvoted 3 times

🗨️ **kishore1980** 7 months, 3 weeks ago

**Selected Answer: C**

storage location is required to be specified to control the object storage location for data written by the pipeline.  
upvoted 2 times

🗨️ **meow\_akk** 8 months ago

Ans E : i think it might be E - <https://docs.databricks.com/en/delta-live-tables/settings.html> - this doc says that target schema and storage location can be optional so it leaves us with E  
upvoted 3 times

🗨️ **Syd** 7 months, 3 weeks ago

Answer is E  
Storage and location are optional.  
<https://docs.databricks.com/en/delta-live-tables/tutorial-pipelines.html>  
upvoted 1 times

🗨️ **kishanu** 8 months ago

**Selected Answer: C**

A path to a cloud storage location for the written data - considering this option is talking about the source data being stored in cloud storage location being ingested to DLT using an autoloader.  
upvoted 3 times

## Question #75

Topic 1

A data engineer has joined an existing project and they see the following query in the project repository:

```
CREATE STREAMING LIVE TABLE loyal_customers AS
```

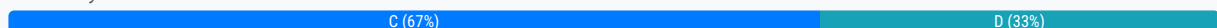
```
SELECT customer_id -  
FROM STREAM(LIVE.customers)  
WHERE loyalty_level = 'high';
```

Which of the following describes why the STREAM function is included in the query?

- A. The STREAM function is not needed and will cause an error.
- B. The table being created is a live table.
- C. The customers table is a streaming live table.**
- D. The customers table is a reference to a Structured Streaming query on a PySpark DataFrame.
- E. The data in the customers table has been updated since its last run.

**Correct Answer: C**

Community vote distribution



  **meow\_akk** Highly Voted 8 months ago

Ans C is correct :

<https://docs.databricks.com/en/sql/load-data-streaming-table.html>

Load data into a streaming table

To create a streaming table from data in cloud object storage, paste the following into the query editor, and then click Run:

SQL

Copy to clipboardCopy

/\* Load data from a volume \*/

CREATE OR REFRESH STREAMING TABLE <table-name> AS

SELECT \* FROM STREAM read\_files('/Volumes/<catalog>/<schema>/<volume>/<path>/<folder>')

/\* Load data from an external location \*/

CREATE OR REFRESH STREAMING TABLE <table-name> AS



SELECT \* FROM STREAM read\_files('s3://<bucket>/<path>/<folder>')

upvoted 6 times

  **benni\_ale** Most Recent 1 month, 3 weeks ago

c is correct . about D: it can be correct but it is not given the fact it comes from pyspark ; sql supports (at least in databricks) the creation of streaming live table as well so it is not necessasarily from pyspark



upvoted 1 times

  **benni\_ale** 1 month, 3 weeks ago

Selected Answer: C

c is ok

upvoted 1 times

  **OfficeSaracus** 1 month, 3 weeks ago

Selected Answer: D

Option E, specifying "at least one notebook library to be executed," is not a requirement for setting up a Delta Live Tables pipeline. Delta Live Tables are built on top of Databricks and use notebooks to define the pipeline's logic, but the actual requirement when setting up the pipeline typically the location where the data will be written to, like a target database or a path to cloud storage. While notebooks may contain the business logic for the transformations and actions within the pipeline, the fundamental requirement for setting up a pipeline is knowing where data will reside after processing, hence why the location of the target database for the written data is crucial.

upvoted 1 times

  **THC1138** 3 weeks, 4 days ago

Wrong question, that's for #73

upvoted 1 times

  **THC1138** 3 weeks, 4 days ago

I mean question #74

upvoted 1 times

  **azure\_bimonster** 5 months ago

Selected Answer: C

C is correct

upvoted 1 times



  **cxw23** 5 months, 2 weeks ago

Ans is A.

CREATE STREAMING LIVE TABLE syntax is does not exist.

It should be CREATE LIVE TABLE AS SELECT \* FROM STREAM.

upvoted 1 times

  **bartfto** 5 months, 1 week ago

LIVE references schema name

customer\_table references table name

upvoted 1 times

Which of the following describes the type of workloads that are always compatible with Auto Loader?

**A. Streaming workloads**

B. Machine learning workloads

C. Serverless workloads

D. Batch workloads

E. Dashboard workloads




**Correct Answer: A**

Community vote distribution

A (100%)

  **meow\_akk**  8 months ago

A is correct Structured streaming for autoloader  
upvoted 5 times

  **benni\_ale**  1 month, 3 weeks ago

**Selected Answer: A**

A is ok  
upvoted 1 times

  **azure\_bimonster** 5 months ago

**Selected Answer: A**

A is correct here  
upvoted 1 times

  **AndreFR** 6 months ago

**Selected Answer: A**

<https://docs.databricks.com/en/ingestion/auto-loader/unity-catalog.html#using-auto-loader-with-unity-catalog>

Auto Loader relies on Structured Streaming for incremental processing  
upvoted 2 times

A data engineer and data analyst are working together on a data pipeline. The data engineer is working on the raw, bronze, and silver layers of the pipeline using Python, and the data analyst is working on the gold layer of the pipeline using SQL. The raw source of the pipeline is a streaming input. They now want to migrate their pipeline to use Delta Live Tables.

Which of the following changes will need to be made to the pipeline when migrating to Delta Live Tables?

A. None of these changes will need to be made

B. The pipeline will need to stop using the medallion-based multi-hop architecture

C. The pipeline will need to be written entirely in SQL

**D. The pipeline will need to use a batch source in place of a streaming source**

E. The pipeline will need to be written entirely in Python

**Correct Answer: B**

Community vote distribution

A (82%)

D (18%)

  **hussamAlHunaiti** 1 week, 1 day ago

**Selected Answer: D**

I had the exam today and option A & B weren't exist, correct answer is D.

upvoted 2 times

  **jaromarg** 1 week, 2 days ago

D:

Delta Live Tables is primarily designed to work with batch processing rather than streaming. This means that when migrating a pipeline to Delta Live Tables, any streaming sources used in the original pipeline will need to be replaced with batch sources.

In the scenario described, where the raw source of the pipeline is a streaming input, the data engineer and data analyst will need to modify the pipeline to read data from a batch source instead. This could involve changing the way data is ingested and processed to align with batch processing paradigms rather than streaming.

Additionally, Delta Live Tables enables the integration of both SQL and Python code within a pipeline, so there's no strict requirement to write pipeline entirely in SQL or Python. Both the data engineer's Python code for the raw, bronze, and silver layers and the data analyst's SQL code for the gold layer can still be used within the Delta Live Tables environment.

Overall, the key change needed when migrating to Delta Live Tables in this scenario is transitioning from a streaming input source to a batch source to align with the batch processing nature of Delta Live Tables.

upvoted 1 times

  **jaromarg** 1 week, 2 days ago

Yes It must be A:

Language Support: DLT allows the use of both SQL and Python, so you can integrate the existing Python and SQL code within the DLT framework.



upvoted 1 times

  **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: A**

A is correct

upvoted 1 times

  **Arunava05** 2 months ago

Cleared the exam today . Option A and B were not available in the exam . There was a different option which was correct.

upvoted 3 times

  **AndreFR** 6 months ago

**Selected Answer: A**

B - DLT support medallion architecture (see example in : <https://docs.databricks.com/en/delta-live-tables/transform.html#combine-streaming-tables-and-materialized-views-in-a-single-pipeline>)



C - DLT can mix Python and SQL using multiple notebooks (according to <https://docs.databricks.com/en/delta-live-tables/tutorial-python.html>) You cannot mix languages within a Delta Live Tables source code file. You can use multiple notebooks or files with different languages in a pipeline)

D - DLT manage streaming sources using streaming tables (ex : <https://docs.databricks.com/en/delta-live-tables/load.html#load-data-from-a-message-bus>)

E - DLT support python and sql (<https://docs.databricks.com/en/delta-live-tables/tutorial-python.html> and <https://docs.databricks.com/en/delta-live-tables/tutorial-sql.html>)

Correct answer is A by elimination

upvoted 4 times



  **kz\_data** 6 months, 2 weeks ago

**Selected Answer: A**

I think the answer is A

upvoted 1 times



  **nedlo** 6 months, 2 weeks ago

**Selected Answer: A**



It should be A. Medallion architecture can be used in DLT pipeline <https://www.databricks.com/glossary/medallion-architecture> "Databricks provides tools like Delta Live Tables (DLT) that allow users to instantly build data pipelines with Bronze, Silver and Gold tables from just a few lines of code."

upvoted 2 times

  **Huroye** 7 months ago

the correct answer is A. DLT needs a notebook where you specify the processing

upvoted 3 times

  **mokrani** 7 months, 2 weeks ago

**Selected Answer: A**

Response A: They have to adapt their notebook's code to be able to declare the DLT pipeline. However, this option is not proposed in the answers so I think it might be A

upvoted 1 times

  **hsks** 7 months, 3 weeks ago

Answer should be A.

upvoted 2 times

  **kbaba101** 7 months, 4 weeks ago

In my opinion, this should be A. Assuming they were working on the same notebook, and weren't declaring the Streaming or Live keywords during development, they would probably need to do so before adding to the DLT workflow. and that is not in the option.

upvoted 2 times

  **meow\_akk** 8 months ago

i think its A ;

upvoted 4 times

A data engineer is using the following code block as part of a batch ingestion pipeline to read from a composable table:

```
transactions_df = (spark.read
    .schema(schema)
    .format("delta")
    .table("transactions")
)
```

Which of the following changes needs to be made so this code block will work when the transactions table is a stream source?

- A. Replace predict with a stream-friendly prediction function
- B. Replace schema(schema) with option ("maxFilesPerTrigger", 1)
- C. Replace "transactions" with the path to the location of the Delta table
- D. Replace format("delta") with format("stream")
- E. Replace spark.read with spark.readStream**

**Correct Answer: E**

Community vote distribution

E (100%)

☐ **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: E**

E is ok

upvoted 1 times

☐ **AndreFR** 6 months ago

**Selected Answer: E**

<https://docs.databricks.com/en/structured-streaming/tutorial.html#use-auto-loader-to-read-streaming-data-from-object-storage>

upvoted 2 times

☐ **55f31c8** 6 months, 3 weeks ago

**Selected Answer: E**

Example from <https://docs.databricks.com/en/structured-streaming/delta-lake.html>

```
spark.readStream.table("table_name")
```

```
spark.readStream.load("/path/to/table")
```

upvoted 3 times

☐ **in89\_io\_90** 6 months, 2 weeks ago

have you cleared the exam

upvoted 1 times

☐ **meow\_akk** 8 months ago

Ans E; for streaming source you use readstream.

<https://docs.databricks.com/en/structured-streaming/delta-lake.html>

upvoted 3 times

Which of the following queries is performing a streaming hop from raw data to a Bronze table?

- A. 

```
(spark.table("sales")
  .groupBy("store")
  .agg(sum("sales"))
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("complete")
  .table("newSales")
)
```
- B. 

```
(spark.table("sales")
  .filter(col("units") > 0)
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("append")
  .table("newSales")
)
```
- C. 

```
(spark.table("sales")
  .withColumn("avgPrice", col("sales") / col("units"))
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("append")
  .table("newSales")
)
```
- D. 

```
(spark.read.load(rawSalesLocation)
  .write
  .mode("append")
  .table("newSales")
)
```
- E. 

```
(spark.readStream.load(rawSalesLocation)
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("append")
  .table("newSales")
)
```

**Correct Answer:** E

Community vote distribution

E (100%)

  **mokrani** Highly Voted 7 months, 2 weeks ago



**Selected Answer: E**

answer E: Raw to Bronze is simply an integration of source data in the lakehouse without any schema needed nor extra operations (e.g. filter, aggregation, joins etc..)

Please refer to this Medallion Architecture article

<https://www.databricks.com/glossary/medallion-architecture>

upvoted 5 times

  **benni\_ale** Most Recent 1 month, 3 weeks ago

**Selected Answer: E**



E is ok, all others are incorrect

upvoted 1 times

  **AndreFR** 6 months ago

sourcename is "rawSalesLocation" (bronze tables contain raw data) and code includes "readStream" to indicate that it is a streaming hop



upvoted 2 times

  **55f31c8** 6 months, 3 weeks ago

**Selected Answer: E**

<https://docs.databricks.com/en/lakehouse/medallion.html#ingest-raw-data-to-the-bronze-layer>

upvoted 2 times

  **sodere** 7 months, 2 weeks ago

Answer is B

upvoted 1 times

  **hsk** 7 months, 2 weeks ago

Answer should be E. Filtering and cleaning usually happens from bronze to silver layer

upvoted 4 times

Question #80

Topic 1

A dataset has been defined using Delta Live Tables and includes an expectations clause:

CONSTRAINT valid\_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION FAIL UPDATE


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

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

**Correct Answer: B**

Community vote distribution



B (100%)

  **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: B**

b is ok

upvoted 1 times

  **55f31c8** 6 months, 3 weeks ago

**Selected Answer: B**

<https://docs.databricks.com/en/delta-live-tables/sql-ref.html#sql-properties>


ON VIOLATION

Optional action to take for failed rows:

FAIL UPDATE: Immediately stop pipeline execution.

DROP ROW: Drop the record and continue processing.

upvoted 2 times

  **Bakhtiyor** 7 months, 2 weeks ago

ON VIOLATION

FAIL UPDATE: Immediately stop pipeline execution.

DROP ROW: Drop the record and continue processing.

upvoted 2 times

  **meow\_akk** 8 months ago

Ans B : delta live tables data quality expectations . - <https://docs.databricks.com/en/delta-live-tables/expectations.html>

Action

Result

warn (default)

Invalid records are written to the target; failure is reported as a metric for the dataset.

drop

Invalid records are dropped before data is written to the target; failure is reported as a metrics for the dataset.

fail

Invalid records prevent the update from succeeding. Manual intervention is required before re-processing.

upvoted 4 times

## Question #81

Topic 1



Which of the following statements regarding the relationship between Silver tables and Bronze tables is always true?

- A. Silver tables contain a less refined, less clean view of data than Bronze data.
- B. Silver tables contain aggregates while Bronze data is unaggregated.
- C. Silver tables contain more data than Bronze tables.
- D. Silver tables contain a more refined and cleaner view of data than Bronze tables**
- E. Silver tables contain less data than Bronze tables.

Correct Answer: D

Community vote distribution

D (100%)

  **benni\_ale** 1 month, 3 weeks ago

**Selected Answer: D**

d is ok

upvoted 1 times

  **azure\_bimonster** 5 months ago

**Selected Answer: D**

D is the right answer

upvoted 2 times

  **meow\_akk** 8 months ago

Ans D : medallion arch databricks

<https://www.databricks.com/glossary/medallion-architecture>

upvoted 2 times

## Question #82

Topic 1

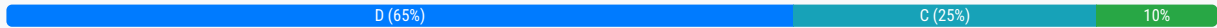
A data engineering team has noticed that their Databricks SQL queries are running too slowly when they are submitted to a non-running SQL endpoint. The data engineering team wants this issue to be resolved.

Which of the following approaches can the team use to reduce the time it takes to return results in this scenario?

- A. They can turn on the Serverless feature for the SQL endpoint and change the Spot Instance Policy to "Reliability Optimized."
- B. They can turn on the Auto Stop feature for the SQL endpoint.
- C. They can increase the cluster size of the SQL endpoint.
- D. They can turn on the Serverless feature for the SQL endpoint**
- E. They can increase the maximum bound of the SQL endpoint's scaling range.

**Correct Answer: D**

Community vote distribution



**carpa\_jo** Highly Voted 5 months, 2 weeks ago

**Selected Answer: D**

The important point of this scenario is "when they are submitted to a non-running SQL endpoint". So its not about increasing the instance size or the amount of instances to improve the query performance, but its about reducing the start-up time.

A: Not possible, serverless can't be combined with spot instance policies, see <https://docs.databricks.com/en/compute/sql-warehouse/serverless.html#limitations>

B: Auto Stop is about terminating a SQL warehouse after x minutes of being idle.

C: Increasing the cluster size provides more capacities for running queries, but doesn't reduce start-up time.

D: Serverless reduces start-up time from minutes to seconds. Jackpot!

E: Increasing the max bound of the SQL endpoints scaling range will help with lots of sequential queries, which is not the case here.

upvoted 11 times

**azure\_bimonster** Most Recent 5 months ago

**Selected Answer: D**

D is correct. Key phrase is "submitted to a non-running SQL endpoint". Increasing cluster size is not going to help if that's in a state like non-running.

upvoted 1 times

**bartfto** 5 months, 1 week ago

**Selected Answer: D**

"when they are submitted to a non-running SQL endpoint" ANSWER D

upvoted 1 times

**Garyn** 5 months, 3 weeks ago

**Selected Answer: C**

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

Explanation:

Increasing the cluster size of the SQL endpoint can enhance query performance by providing more computational resources to execute queries. This can potentially speed up query processing by allowing more parallelism, handling larger workloads, and reducing the time taken for query execution.

upvoted 1 times

 **AndreFR** 6 months ago

key word, “non-running SQL endpoint” which implies that the query is slow because the cluster needs time to get started.

I suggest answer D because :

A : Serverless & spot instances cannot be mixed ?


B : autostop means that jobs are submitted to non-running SQL endpoints

C : increasing the clustersize can compensate for slow startup time

D : serverless is able to start and scale faster than non-running SQL endpoints (seconds instead of minutes)

E : increasing maximum bound will help only if there are simultaneous queries


<https://docs.gcp.databricks.com/en/lakehouse-architecture/cost-optimization/best-practices.html#use-serverless-for-your-workloads>  
upvoted 4 times

 **olaru** 6 months, 1 week ago

**Selected Answer: E**


maximum bound of the SQL endpoint's scaling range

upvoted 2 times

 **nedlo** 6 months, 2 weeks ago

**Selected Answer: C**

D is wrong - its already Serverless (non running SQL endpoint) how would turning Serverless ON help? They also says C here  
<https://community.databricks.com/t5/data-engineering/when-to-increase-maximum-bound-vs-when-to-increase-cluster-size/td-p/27880> . E is  
only true for autoscaling clusters  
upvoted 2 times

 **msengupta** 6 months, 2 weeks ago

**Selected Answer: C**

<https://community.databricks.com/t5/data-engineering/sql-query-takes-too-long-to-run/td-p/21884>  
upvoted 2 times

 **Syd** 7 months, 3 weeks ago

Answer E:

<https://www.databricks.com/blog/2022/03/10/top-5-databricks-performance-tips.html>  
upvoted 2 times

 **Syd** 7 months, 3 weeks ago

I mean answer C

upvoted 1 times

 **meow\_akk** 8 months ago

Ans E : you re welcome :)

<https://community.databricks.com/t5/data-engineering/when-to-increase-maximum-bound-vs-when-to-increase-cluster-size/td-p/27880>  
upvoted 1 times

 **mike\_stewart** 7 months, 3 weeks ago

I don't agree. Your answer is only valid when 'sequential' is mentioned, which is not the case here.

upvoted 1 times



A data engineer has a Job that has a complex run schedule, and they want to transfer that schedule to other Jobs.



Rather than manually selecting each value in the scheduling form in Databricks, which of the following tools can the data engineer use to represent and submit the schedule programmatically?

- A. `pyspark.sql.types.DateType`
- B. `datetime`
- C. `pyspark.sql.types.TimestampType`
- D. Cron syntax**
- E. There is no way to represent and submit this information programmatically

**Correct Answer: D**

*Community vote distribution*

D (100%)

  **55f31c8** 6 months, 3 weeks ago

**Selected Answer: D**

<https://docs.databricks.com/en/sql/user/queries/schedule-query.html>

upvoted 2 times

  **meow\_akk** 8 months ago

Ans D : Cron Syntax with that you can easily copy all the syntax

upvoted 3 times


Which of the following approaches should be used to send the Databricks Job owner an email in the case that the Job fails?

- A. Manually programming in an alert system in each cell of the Notebook
- B. Setting up an Alert in the Job page**
- C. Setting up an Alert in the Notebook
- D. There is no way to notify the Job owner in the case of Job failure
- E. MLflow Model Registry Webhooks

**Correct Answer: B**

Community vote distribution

B (100%)

 **Lavpak** 6 months, 2 weeks ago

**Selected Answer: B**

Setting up an alert in Jobs page  
upvoted 2 times

 **meow\_akk** 8 months ago

Ans B : <https://docs.databricks.com/en/workflows/jobs/job-notifications.html>  
upvoted 3 times

An engineering manager uses a Databricks SQL query to monitor ingestion latency for each data source. The manager checks the results of the query every day, but they are manually rerunning the query each day and waiting for the results.

Which of the following approaches can the manager use to ensure the results of the query are updated each day?

- A. They can schedule the query to refresh every 1 day from the SQL endpoint's page in Databricks SQL.
- B. They can schedule the query to refresh every 12 hours from the SQL endpoint's page in Databricks SQL.
- C. They can schedule the query to refresh every 1 day from the query's page in Databricks SQL.**
- D. They can schedule the query to run every 1 day from the Jobs UI.
- E. They can schedule the query to run every 12 hours from the Jobs UI.

**Correct Answer: C**

Community vote distribution

C (100%)

🗨️ 👤 **Garyn** 5 months, 3 weeks ago

**Selected Answer: C**

The manager can schedule the query to refresh every 1 day from the query's page in Databricks SQL (Option C). Here are the steps to do this:

- In the Query Editor, click Schedule > Add schedule to open a menu with schedule settings.
- Choose when to run the query. Use the dropdown pickers to specify the frequency, period, starting time, and time zone.
- Click Create.

upvoted 3 times

🗨️ 👤 **AndreFR** 6 months ago

**Selected Answer: C**

has to be every 1 day to run once day. <https://docs.databricks.com/en/sql/user/queries/schedule-query.html>

upvoted 1 times

🗨️ 👤 **kz\_data** 6 months, 2 weeks ago

**Selected Answer: C**

Correct Answer is C

upvoted 1 times

🗨️ 👤 **kishore1980** 7 months, 3 weeks ago

**Selected Answer: C**

From the query editor page we have option to schedule the queries

upvoted 2 times

🗨️ 👤 **meow\_akk** 8 months ago

Ans D : think option A might not be right since we are not doing scheduling in sql end points page

upvoted 1 times

🗨️ 👤 **SD5713** 8 months ago

it is C, Question 41 of Practice Exam Databricks

upvoted 6 times



In which of the following scenarios should a data engineer select a Task in the Depends On field of a new Databricks Job Task?

- A. When another task needs to be replaced by the new task
- B. When another task needs to fail before the new task begins
- C. When another task has the same dependency libraries as the new task
- D. When another task needs to use as little compute resources as possible**
- E. When another task needs to successfully complete before the new task begins

**Correct Answer: E**

Community vote distribution

E (100%)

  **Lavpak** 6 months, 2 weeks ago

**Selected Answer: E**

<https://docs.databricks.com/en/workflows/jobs/conditional-tasks.html>

upvoted 1 times

  **meow\_akk** 8 months ago

Ans E : E is correct since dependency means the dependent job must complete successfully.

upvoted 2 times

A data engineer has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to a data analytics dashboard for a retail use case. The job has a Databricks SQL query that returns the number of store-level records where sales is equal to zero. The data engineer wants their entire team to be notified via a messaging webhook whenever this value is greater than 0.

Which of the following approaches can the data engineer use to notify their entire team via a messaging webhook whenever the number of stores with \$0 in sales is greater than zero?

- A. They can set up an Alert with a custom template.
- B. They can set up an Alert with a new email alert destination.
- C. They can set up an Alert with one-time notifications.
- D. They can set up an Alert with a new webhook alert destination.**
- E. They can set up an Alert without notifications.

**Correct Answer: D**


Community vote distribution

D (100%)

 **azure\_bimonster** 5 months ago

**Selected Answer: D**

D is the right answer here  
upvoted 1 times

 **Lavpak** 6 months, 2 weeks ago

**Selected Answer: D**

Set up an Alert with a new webhook alert destination  
upvoted 2 times

 **meow\_akk** 8 months ago

Ans D : the questions specifically says via a messaging webhook  
upvoted 2 times

A data engineer wants to schedule their Databricks SQL dashboard to refresh every hour, but they only want the associated SQL endpoint to be running when it is necessary. The dashboard has multiple queries on multiple datasets associated with it. The data that feeds the dashboard is automatically processed using a Databricks Job.

Which of the following approaches can the data engineer use to minimize the total running time of the SQL endpoint used in the refresh schedule of their dashboard?

- A. They can turn on the Auto Stop feature for the SQL endpoint.**
- B. They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.
- C. They can reduce the cluster size of the SQL endpoint.
- D. They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.

E. They can set up the dashboard's SQL endpoint to be serverless.

**Correct Answer: E**

Community vote distribution

A (100%)

🗄️ 👤 **meow\_akk** Highly Voted 8 months ago

Ans A : the keyword in the question is " running only when its necessary "  
upvoted 5 times

🗄️ 👤 **hussamAlHunaiti** Most Recent 1 week, 3 days ago

Selected Answer: A

Answer is A.  
The question about minimize the total running times not enhance the performance.  
upvoted 1 times

🗄️ 👤 **Lavpak** 6 months, 2 weeks ago

Selected Answer: A

They can turn on the Auto Stop feature  
upvoted 2 times

🗄️ 👤 **Huroye** 7 months ago

The correct answer is A. They can use the Auto Stop feature. When the query is idel it will stop the compute  
upvoted 3 times

🗄️ 👤 **SD5713** 8 months ago

Selected Answer: A

A. They can turn on the Auto Stop feature for the SQL endpoint.  
upvoted 3 times

A data engineer needs access to a table new\_table, but they do not have the correct permissions. They can ask the table owner for permission, but they do not know who the table owner is.

Which of the following approaches can be used to identify the owner of new\_table?

- A. Review the Permissions tab in the table's page in Data Explorer
- B. All of these options can be used to identify the owner of the table
- C. Review the Owner field in the table's page in Data Explorer**
- D. Review the Owner field in the table's page in the cloud storage solution
- E. There is no way to identify the owner of the table

**Correct Answer: C**

Community vote distribution

C (100%)

☐ **Lavpak** 6 months, 2 weeks ago

**Selected Answer: C**

Review the Owner field in the table's page in Data Explorer

upvoted 3 times

☐ **Huroye** 7 months ago

Correct answer is C since you are looking for the owner and not the permissions on the table

upvoted 3 times

☐ **meow\_akk** 8 months ago

ANS C : C IS CORRECT ,

upvoted 2 times

A new data engineering team has been assigned to an ELT project. The new data engineering team will need full privileges on the table sales to fully manage the project.

Which of the following commands can be used to grant full permissions on the database to the new data engineering team?

- A. GRANT ALL PRIVILEGES ON TABLE sales TO team;**
- B. GRANT SELECT CREATE MODIFY ON TABLE sales TO team;
- C. GRANT SELECT ON TABLE sales TO team;
- D. GRANT USAGE ON TABLE sales TO team;
- E. GRANT ALL PRIVILEGES ON TABLE team TO sales;

**Correct Answer: A**

Community vote distribution

A (100%)

🗨️ **hussamAlHunaiti** 1 week, 3 days ago

**Selected Answer: A**

Correct answer A.  
upvoted 1 times

🗨️ **azure\_bimonster** 5 months ago

**Selected Answer: A**

A is correct as provided syntax is right  
upvoted 1 times

🗨️ **Lavpak** 6 months, 2 weeks ago

**Selected Answer: A**

GRANT ALL grants all privileges on "Sales" Table  
upvoted 1 times

🗨️ **meow\_akk** 8 months ago

Ans A :  
grant "privilege" on "object" object\_name to <user or group>  
upvoted 3 times



Which data lakehouse feature results in improved data quality over a traditional data lake?

- A. A data lakehouse stores data in open formats.
- B. A data lakehouse allows the use of SQL queries to examine data.
- C. A data lakehouse provides storage solutions for structured and unstructured data.
- D. A data lakehouse supports ACID-compliant transactions.**

**Correct Answer:** D

Community vote distribution

D (100%)

  **31cadd7** 5 days, 5 hours ago

**Selected Answer:** D

it's D

upvoted 1 times

In which scenario will a data team want to utilize cluster pools?

- A. An automated report needs to be version-controlled across multiple collaborators.
- B. An automated report needs to be runnable by all stakeholders.
- C. An automated report needs to be refreshed as quickly as possible.**
- D. An automated report needs to be made reproducible.

**Correct Answer:** C

  **MDWPartners** 3 weeks, 3 days ago

This question is repeated, a cluster pool helps to reduce the times of cold start and change the scaling .

upvoted 1 times

What is hosted completely in the control plane of the classic Databricks architecture?

- A. Worker node
- B. Databricks web application**
- C. Driver node
- D. Databricks Filesystem

Correct Answer: B

A data engineer needs to determine whether to use the built-in Databricks Notebooks versioning or version their project using Databricks Repos.

What is an advantage of using Databricks Repos over the Databricks Notebooks versioning?

- A. Databricks Repos allows users to revert to previous versions of a notebook
- B. Databricks Repos is wholly housed within the Databricks Data Intelligence Platform
- C. Databricks Repos provides the ability to comment on specific changes
- D. Databricks Repos supports the use of multiple branches**

Correct Answer: D

What is a benefit of the Databricks Lakehouse Architecture embracing open source technologies?

- A. Avoiding vendor lock-in**
- B. Simplified governance
- C. Ability to scale workloads
- D. Cloud-specific integrations

Correct Answer: A

A data engineer needs to use a Delta table as part of a data pipeline, but they do not know if they have the appropriate permissions.

In which location can the data engineer review their permissions on the table?

- A. Jobs
- B. Dashboards
- C. Catalog Explorer**
- D. Repos



Correct Answer: C

A data engineer is running code in a Databricks Repo that is cloned from a central Git repository. A colleague of the data engineer informs them that changes have been made and synced to the central Git repository. The data engineer now needs to sync their Databricks Repo to get the changes from the central Git repository.

Which Git operation does the data engineer need to run to accomplish this task?

- A. Clone
- B. Pull**
- C. Merge
- D. Push

Correct Answer: B

  **MDWPartners** 3 weeks, 3 days ago  
Repeated, also correct.  
upvoted 1 times



Which file format is used for storing Delta Lake Table?

- A. CSV
- B. Parquet**
- C. JSON
- D. Delta

Correct Answer: B

  **SannPro** 3 weeks, 4 days ago

B is correct  
upvoted 1 times

  **aspix82** 1 month ago

The answer is D  
upvoted 1 times

  **aspix82** 3 weeks, 6 days ago

No, file format is Parquet! The correct answer is B  
upvoted 1 times

A data architect has determined that a table of the following format is necessary:

```
employeeId startDate avgRating
```

```
a1          2009-01-06  5.5
a2          2018-11-21  7.1
***          ***          ***
```

Which code block is used by SQL DDL command to create an empty Delta table in the above format regardless of whether a table already exists with this name?

- A. CREATE OR REPLACE TABLE table\_name ( employeeId STRING, startDate DATE, avgRating FLOAT )**
- B. CREATE OR REPLACE TABLE table\_name WITH COLUMNS ( employeeId STRING, startDate DATE, avgRating FLOAT ) USING DELTA
- C. CREATE TABLE IF NOT EXISTS table\_name ( employeeId STRING, startDate DATE, avgRating FLOAT )
- D. CREATE TABLE table\_name AS SELECT employeeId STRING, startDate DATE, avgRating FLOAT

Correct Answer: A

  **MDWPartners** 3 weeks, 3 days ago

Repeated, correct.  
upvoted 1 times

A data engineer has been given a new record of data:

```
id STRING = 'a1'  
rank INTEGER = 6  
rating FLOAT = 9.4
```

Which SQL commands can be used to append the new record to an existing Delta table my\_table?



**A. INSERT INTO my\_table VALUES ('a1', 6, 9.4)**

B. INSERT VALUES ('a1', 6, 9.4) INTO my\_table

C. UPDATE my\_table VALUES ('a1', 6, 9.4)

D. UPDATE VALUES ('a1', 6, 9.4) my\_table

Correct Answer: A

  **MDWPartners** 3 weeks, 3 days ago  
Repeated, correct.  
upvoted 1 times

A data engineer has realized that the data files associated with a Delta table are incredibly small. They want to compact the small files to form larger files to improve performance.

Which keyword can be used to compact the small files?



**A. OPTIMIZE**



B. VACUUM

C. COMPACTION

D. REPARTITION

Correct Answer: A

  **kim32** 1 day ago  
The OPTIMIZE command is used to compact small files into larger ones, which helps improve the performance of Delta Lake tables. It consolidates small files into fewer larger files to reduce the overhead associated with having many small files. This process is often referred to "compaction" but the specific keyword in Databricks Delta Lake is OPTIMIZE.  
upvoted 1 times

  **MDWPartners** 3 weeks, 3 days ago  
Repeated, correct.  
upvoted 1 times

A data engineer wants to create a data entity from a couple of tables. The data entity must be used by other data engineers in other sessions. It also must be saved to a physical location.

Which of the following data entities should the data engineer create?

- A. Table**
- B. Function
- C. View
- D. Temporary view

Correct Answer: A

A data engineer runs a statement every day to copy the previous day's sales into the table transactions. Each day's sales are in their own file in the location `"/transactions/raw"`.

Today, the data engineer runs the following command to complete this task:



```
COPY INTO transactions
FROM "/transactions/raw"
FILEFORMAT = PARQUET;
```

After running the command today, the data engineer notices that the number of records in table transactions has not changed.

What explains why the statement might not have copied any new records into the table?

- A. The format of the files to be copied were not included with the `FORMAT_OPTIONS` keyword.
- B. The `COPY INTO` statement requires the table to be refreshed to view the copied rows.
- C. The previous day's file has already been copied into the table.**
- D. The `PARQUET` file format does not support `COPY INTO`.


Correct Answer: C

  **MDWPartners** 3 weeks, 3 days ago  
Repeated, correct.  
upvoted 1 times

Which command can be used to write data into a Delta table while avoiding the writing of duplicate records?

- A. DROP
- B. INSERT
- C. MERGE**
- D. APPEND

Correct Answer: C


 **MDWPartners** 3 weeks, 3 days ago  
Repeated, correct.  
upvoted 1 times

A data analyst has created a Delta table sales that is used by the entire data analysis team. They want help from the data engineering team to implement a series of tests to ensure the data is clean. However, the data engineering team uses Python for its tests rather than SQL.

Which command could the data engineering team use to access sales in PySpark?

- A. `SELECT * FROM sales`
- B. `spark.table("sales")`**
- C. `spark.sql("sales")`
- D. `spark.delta.table("sales")`

Correct Answer: B

 **MDWPartners** 3 weeks, 3 days ago  
Repeated, correct.  
upvoted 1 times



A data engineer has created a new database using the following command:

```
CREATE DATABASE IF NOT EXISTS customer360;
```

In which location will the customer360 database be located?

- A. dbfs:/user/hive/database/customer360
- B. dbfs:/user/hive/warehouse**
- C. dbfs:/user/hive/customer360
- D. dbfs:/user/hive/database

**Correct Answer:** B

  **MDWPartners** 3 weeks, 3 days ago  
Repeated, correct.  
upvoted 1 times



A data engineer is attempting to drop a Spark SQL table `my_table` and runs the following command:

```
DROP TABLE IF EXISTS my_table;
```

After running this command, the engineer notices that the data files and metadata files have been deleted from the file system.

What is the reason behind the deletion of all these files?

**A. The table was managed**

B. The table's data was smaller than 10 GB



C. The table did not have a location

D. The table was external

**Correct Answer: D**

Community vote distribution

A (100%)

  **31cadd7** 5 days, 5 hours ago

**Selected Answer: A**

it's A

upvoted 1 times

  **THC1138** 3 weeks, 4 days ago

**Selected Answer: A**

For D to be correct, the metadata would have been deleted, but the data would still exist. The answer is A

upvoted 1 times

  **PreranaC** 3 weeks, 6 days ago

**Selected Answer: A**


A - Both Data was deleted as well along with Metadata

upvoted 1 times

  **Ivan\_Petrov** 1 month, 1 week ago

Answer should be A as data was deleted table was managed

upvoted 1 times

  **jetplanes** 1 month, 1 week ago

**Selected Answer: A**

The answer should be A --> the table was MANAGED. If the metadata and the underlined files have been deleted, then this is a MANAGED table and not an external table.

upvoted 3 times

A data engineer needs to create a table in Databricks using data from a CSV file at location /path/to/csv.

They run the following command:

```
CREATE TABLE new_table  
  
_____  
OPTIONS (  
  header = "true",  
  delimiter = "|" )  
LOCATION "path/to/csv"
```

Which of the following lines of code fills in the above blank to successfully complete the task?

A. FROM "path/to/csv"

**B. USING CSV**

C. FROM CSV

D. USING DELTA

Correct Answer: B

What is a benefit of creating an external table from Parquet rather than CSV when using a CREATE TABLE AS SELECT statement?

- A. Parquet files can be partitioned
- B. Parquet files will become Delta tables
- C. Parquet files have a well-defined schema**
- D. Parquet files have the ability to be optimized

**Correct Answer:** C

Community vote distribution

C (100%)

 **MDWPartners** 3 weeks, 3 days ago

**Selected Answer: C**

Parquet are columnar and can be optimized. However, I think the key part is "when using a CREATE TABLE AS SELECT statement?", that's C  
upvoted 1 times

 **mgari** 1 month ago

D parquet file are columnar and optimised  
upvoted 1 times

Which SQL keyword can be used to convert a table from a long format to a wide format?

- A. TRANSFORM
- B. PIVOT**
- C. SUM
- D. CONVERT

**Correct Answer:** B

A data engineer has a Python variable `table_name` that they would like to use in a SQL query. They want to construct a Python code block that will run the query using `table_name`.

They have the following incomplete code block:

```
____(f"SELECT customer_id, spend FROM {table_name}")
```

What can be used to fill in the blank to successfully complete the task?

- A. `spark.delta.sql`
- B. `spark.sql`**
- C. `spark.table`
- D. `dbutils.sql`

Correct Answer: B

A data engineer is working with two tables. Each of these tables is displayed below in its entirety.

```
sales
customer_id spend units
a1           28.94  7
a3           874.1223
a4           8.99   1
```

```
favorite_stores
```

```
customer_id store_id
a1           s1
a2           s1
a4           s2
```

The data engineer runs the following query to join these tables together:

```
SELECT
    sales.customer_id,
    sales.spend,
    favorite_stores.store_id
FROM sales
LEFT JOIN favorite_stores
ON sales.customer_id = favorite_stores.customer_id;
```

A. 

|    | customer_id | spend | store_id |
|----|-------------|-------|----------|
| a1 |             | 28.94 | s1       |
| a2 |             | NULL  | s1       |
| a4 |             | 8.99  | s2       |

B. 

|    | customer_id | spend | store_id |
|----|-------------|-------|----------|
| a1 |             | 28.94 | s1       |
| a4 |             | 8.99  | s2       |

C. 

|    | customer_id | spend  | store_id |
|----|-------------|--------|----------|
| a1 |             | 28.94  | s1       |
| a3 |             | 874.12 | NULL     |
| a4 |             | 8.99   | s2       |

D. 

|    | customer_id | spend  | store_id |
|----|-------------|--------|----------|
| a1 |             | 28.94  | s1       |
| a2 |             | NULL   | s1       |
| a3 |             | 874.12 | NULL     |
| a4 |             | 8.99   | s2       |

Correct Answer: C

Community vote distribution

D (100%)

A data engineer needs to apply custom logic to identify employees with more than 5 years of experience in array column employees in table stores. The custom logic should create a new column exp\_employees that is an array of all of the employees with more than 5 years of experience for each row. In order to apply this custom logic at scale, the data engineer wants to use the FILTER higher-order function.

Which code block successfully completes this task?

- ```
SELECT
    store_id,
A.    employees,
    FILTER (employees, i -> i.years_exp > 5) AS exp_employees
FROM stores;

SELECT
    store_id,
B.    employees,
    FILTER (exp_employees, i -> i.years_exp > 5) AS exp_employees
FROM stores;

SELECT
    store_id,
C.    employees,
    FILTER (employees, years_exp > 5) AS exp_employees
FROM stores;

SELECT
    store_id,
    employees,
D.    CASE WHEN employees.years_exp > 5 THEN employees
        ELSE NULL
    END AS exp_employees
FROM stores;
```

Correct Answer: A

A data engineer that is new to using Python needs to create a Python function to add two integers together and return the sum?

Which code block can the data engineer use to complete this task?

- ```
A. function add_integers(x, y):
    return x + y

B. def add_integers(x, y):
    print(x + y)
```

C. 

```
def add_integers(x, y):  
    x + y
```

D. 

```
def add_integers(x, y):  
    return x + y
```

**Correct Answer: C**

Community vote distribution

D (100%)

☐ **hussamAlHunaiti** 1 week, 3 days ago

**Selected Answer: D**

Answer is D.

upvoted 1 times

☐ **nawfalbourass** 3 weeks, 6 days ago

**Selected Answer: D**

RETURN is needed

upvoted 2 times

☐ **PreranaC** 3 weeks, 6 days ago

**Selected Answer: D**

D, RETURN in Python needed

upvoted 1 times

☐ **Kunka** 1 month, 1 week ago

Answer is D.

C is wrong answer, as it missed return key word

upvoted 1 times

☐ **Ivan\_Petrov** 1 month, 1 week ago

Correct answer is D

upvoted 2 times

☐ **helmerpaiva** 1 month, 1 week ago

Correct is D

upvoted 3 times

☐ **jetplanes** 1 month, 1 week ago

**Selected Answer: D**

The correct answer is D because the python function needs to provide a return for the function, so C is incorrect.

upvoted 1 times

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")
  .trigger(
    .table("new_sales")
  )
)
```

Which line of code should the data engineer use to fill in the blank if the data engineer only wants the query to execute a micro-batch to process data every 5 seconds?

- A. trigger("5 seconds")
- B. trigger(continuous="5 seconds")
- C. trigger(once="5 seconds")
- D. trigger(processingTime="5 seconds")**

Correct Answer: D

A data engineer is maintaining a data pipeline. Upon data ingestion, the data engineer notices that the source data is starting to have a lower level of quality. The data engineer would like to automate the process of monitoring the quality level.

Which of the following tools can the data engineer use to solve this problem?

- A. Auto Loader
- B. Unity Catalog
- C. Delta Lake
- D. Delta Live Tables**

Correct Answer: D



A data engineer has three tables in a Delta Live Tables (DLT) pipeline. They have configured the pipeline to drop invalid records at each table. They notice that some data is being dropped due to quality concerns at some point in the DLT pipeline. They would like to determine at which table in their pipeline the data is being dropped.

Which approach can the data engineer take to identify the table that is dropping the records?

- A. They can set up separate expectations for each table when developing their DLT pipeline.
- B. They can navigate to the DLT pipeline page, click on the "Error" button, and review the present errors.
- C. They can set up DLT to notify them via email when records are dropped.
- D. They can navigate to the DLT pipeline page, click on each table, and view the data quality statistics.**

Correct Answer: D

What is used by Spark to record the offset range of the data being processed in each trigger in order for Structured Streaming to reliably track the exact progress of the processing so that it can handle any kind of failure by restarting and/or reprocessing?

- A. Checkpointing and Write-ahead Logs**
- B. Replayable Sources and Idempotent Sinks
- C. Write-ahead Logs and Idempotent Sinks
- D. Checkpointing and Idempotent Sinks

Correct Answer: A

What describes the relationship between Gold tables and Silver tables?

- A. Gold tables are more likely to contain aggregations than Silver tables.**
- B. Gold tables are more likely to contain valuable data than Silver tables.
- C. Gold tables are more likely to contain a less refined view of data than Silver tables.
- D. Gold tables are more likely to contain truthful data than Silver tables.

Correct Answer: A

What describes when to use the CREATE STREAMING LIVE TABLE (formerly CREATE INCREMENTAL LIVE TABLE) syntax over the CREATE LIVE TABLE syntax when creating Delta Live Tables (DLT) tables using SQL?

- A. CREATE STREAMING LIVE TABLE should be used when the subsequent step in the DLT pipeline is static.
- B. CREATE STREAMING LIVE TABLE should be used when data needs to be processed incrementally.**
- C. CREATE STREAMING LIVE TABLE should be used when data needs to be processed through complicated aggregations.
- D. CREATE STREAMING LIVE TABLE should be used when the previous step in the DLT pipeline is static.

Correct Answer: B

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 Production mode using the Continuous Pipeline Mode.

What is the expected outcome after clicking Start to update the pipeline assuming previously unprocessed data exists and all definitions are valid?

- A. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing.
- B. All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.
- C. 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.**
- D. All datasets will be updated once and the pipeline will shut down. The compute resources will be terminated.

Correct Answer: C

Which type of workloads are compatible with Auto Loader?

- A. Streaming workloads**
- B. Machine learning workloads
- C. Serverless workloads
- D. Batch workloads

Correct Answer: A

A data engineer has developed a data pipeline to ingest data from a JSON source using Auto Loader, but the engineer has not provided any type inference or schema hints in their pipeline. Upon reviewing the data, the data engineer has noticed that all of the columns in the target table are of the string type despite some of the fields only including float or boolean values.

Why has Auto Loader inferred all of the columns to be of the string type?

- A. Auto Loader cannot infer the schema of ingested data
- B. JSON data is a text-based format**
- C. Auto Loader only works with string data
- D. All of the fields had at least one null value

Correct Answer: B

Which statement regarding the relationship between Silver tables and Bronze tables is always true?



**A. Silver tables contain a less refined, less clean view of data than Bronze data.**

B. Silver tables contain aggregates while Bronze data is unaggregated.

C. Silver tables contain more data than Bronze tables.

D. Silver tables contain less data than Bronze tables.

Correct Answer: B

  **Ivan\_Petrov** Highly Voted 1 month, 1 week ago

looks like there is no correct answer. Should be like A but Silver and Bronze should be changed in their places  
upvoted 6 times

  **mgari** Most Recent 1 week, 5 days ago

in my opinion it is D Silver has only the data with no error  
upvoted 1 times

  **b79962e** 3 weeks, 6 days ago

I think there is no correct answer  
upvoted 2 times

  **PreranaC** 3 weeks, 6 days ago

Silver tables contain a more refined and cleaner view of data than Bronze tables.  
upvoted 2 times

  **helmerpaiva** 1 month, 1 week ago

Correct is A  
upvoted 1 times

Which query is performing a streaming hop from raw data to a Bronze table?

- A. 

```
(spark.table("sales")
    .groupBy("store")
    .agg(sum("sales"))
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("complete")
    .table("newSales")
)
```
- B. 

```
(spark.read.load(rawSalesLocation)
    .write
    .mode("append")
    .table("newSales")
)
```
- C. 

```
(spark.table("sales")
    .withColumn("avgPrice", col("sales") / col("units"))
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```
- D. 

```
(spark.readStream.load(rawSalesLocation)
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```

Correct Answer: D

A dataset has been defined using Delta Live Tables and includes an expectations clause:

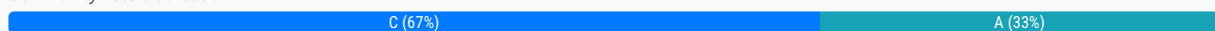
```
CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION DROP ROW
```

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

- A. Records that violate the expectation cause the job to fail.
- B. Records that violate the expectation are added to the target dataset and flagged as invalid in a field added to the target dataset.
- C. Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.**
- D. Records that violate the expectation are added to the target dataset and recorded as invalid in the event log.

**Correct Answer:** C

Community vote distribution



☐ **31cadd7** 5 days, 5 hours ago

**Selected Answer: C**

it's C

upvoted 1 times

☐ **d39c1db** 3 weeks ago

C. Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.

When a constraint defined using the EXPECT clause is violated, Delta Live Tables will drop the records that violate the expectation from the target dataset. Additionally, information about the dropped records and the reason for their exclusion will be recorded in the event log for audit and monitoring purposes. This ensures that only valid data meeting the specified constraints is included in the target dataset.

upvoted 1 times

☐ **PreranaC** 3 weeks, 6 days ago

**Selected Answer: C**

C should be correct, A is for ON VIOLATION FAIL UPDATE

upvoted 1 times

☐ **PreranaC** 3 weeks, 6 days ago

**Selected Answer: A**

A should be correct

upvoted 1 times

☐ **MDWPartners** 3 weeks, 3 days ago

i don't agree, it shouldn't make the job to fail.

upvoted 1 times

A data engineer has a Job with multiple tasks that runs nightly. Each of the tasks runs slowly because the clusters take a long time to start.

Which action can the data engineer perform to improve the start up time for the clusters used for the Job?

- A. They can use endpoints available in Databricks SQL
- B. They can use jobs clusters instead of all-purpose clusters
- C. They can configure the clusters to autoscale for larger data sizes
- D. They can use clusters that are from a cluster pool**

**Correct Answer:** D

A data engineer has a single-task Job that runs each morning before they begin working. After identifying an upstream data issue, they need to set up another task to run a new notebook prior to the original task.

Which approach can the data engineer use to set up the new task?

- A. They can clone the existing task in the existing Job and update it to run the new notebook.
- B. They can create a new task in the existing Job and then add it as a dependency of the original task.**
- C. They can create a new task in the existing Job and then add the original task as a dependency of the new task.
- D. They can create a new job from scratch and add both tasks to run concurrently.

**Correct Answer: C**

*Community vote distribution*

B (100%)

🗳️ **hussamAlHunaiti** 1 week, 3 days ago

**Selected Answer: B**

Answer is B.  
New task is prior than the original task.  
upvoted 1 times

🗳️ **PreranaC** 3 weeks, 6 days ago

**Selected Answer: B**

B is correct  
upvoted 1 times

🗳️ **nmosq** 1 month ago

B is correct, "needs to run prior to the original task"  
upvoted 1 times

🗳️ **BharaniRaj** 1 month ago

**Selected Answer: B**

B is correct  
upvoted 1 times

🗳️ **Kunka** 1 month, 1 week ago

B is correct, as new task runs first  
upvoted 1 times

🗳️ **Ivan\_Petrov** 1 month, 1 week ago

B is correct  
upvoted 1 times



A single Job runs two notebooks as two separate tasks. A data engineer has noticed that one of the notebooks is running slowly in the Job's current run. The data engineer asks a tech lead for help in identifying why this might be the case.

Which approach can the tech lead use to identify why the notebook is running slowly as part of the Job?

- A. They can navigate to the Runs tab in the Jobs UI to immediately review the processing notebook.
- B. They can navigate to the Tasks tab in the Jobs UI and click on the active run to review the processing notebook.
- C. They can navigate to the Runs tab in the Jobs UI and click on the active run to review the processing notebook.**
- D. They can navigate to the Tasks tab in the Jobs UI to immediately review the processing notebook.

Correct Answer: C

A data analysis team has noticed that their Databricks SQL queries are running too slowly when connected to their always-on SQL endpoint. They claim that this issue is present when many members of the team are running small queries simultaneously. They ask the data engineering team for help. The data engineering team notices that each of the team's queries uses the same SQL endpoint.

Which approach can the data engineering team use to improve the latency of the team's queries?

- A. They can increase the cluster size of the SQL endpoint.
- B. They can increase the maximum bound of the SQL endpoint's scaling range.**
- C. They can turn on the Auto Stop feature for the SQL endpoint.
- D. They can turn on the Serverless feature for the SQL endpoint.

Correct Answer: B

A data engineer has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to an ELT job. The ELT job has its Databricks SQL query that returns the number of input records containing unexpected NULL values. The data engineer wants their entire team to be notified via a messaging webhook whenever this value reaches 100.

Which approach can the data engineer use to notify their entire team via a messaging webhook whenever the number of NULL values reaches 100?

- A. They can set up an Alert with a custom template.
- B. They can set up an Alert with a new email alert destination.
- C. They can set up an Alert with a new webhook alert destination.**
- D. They can set up an Alert with one-time notifications.

Correct Answer: C

A data engineer wants to schedule their Databricks SQL dashboard to refresh once per day, but they only want the associated SQL endpoint to be running when it is necessary.

Which approach can the data engineer use to minimize the total running time of the SQL endpoint used in the refresh schedule of their dashboard?

- A. They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.
- B. They can set up the dashboard's SQL endpoint to be serverless.
- C. They can turn on the Auto Stop feature for the SQL endpoint.**
- D. They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.

Correct Answer: C

An engineering manager wants to monitor the performance of a recent project using a Databricks SQL query. For the first week following the project's release, the manager wants the query results to be updated every minute. However, the manager is concerned that the compute resources used for the query will be left running and cost the organization a lot of money beyond the first week of the project's release.

Which approach can the engineering team use to ensure the query does not cost the organization any money beyond the first week of the project's release?

- A. They can set a limit to the number of DBUs that are consumed by the SQL Endpoint.
- B. They can set the query's refresh schedule to end after a certain number of refreshes.
- C. They can set the query's refresh schedule to end on a certain date in the query scheduler.**
- D. They can set a limit to the number of individuals that are able to manage the query's refresh schedule.

**Correct Answer: C**

A new data engineering team has been assigned to work on a project. The team will need access to database customers in order to see what tables already exist. The team has its own group team.

Which command can be used to grant the necessary permission on the entire database to the new team?

- A. GRANT VIEW ON CATALOG customers TO team;
- B. GRANT CREATE ON DATABASE customers TO team;
- C. GRANT USAGE ON CATALOG team TO customers;
- D. GRANT USAGE ON DATABASE customers TO team;**

**Correct Answer: D**

A new data engineering team has been assigned to an ELT project. The new data engineering team will need full privileges on the table sales to fully manage the project.

Which command can be used to grant full permissions on the database to the new data engineering team?

- A. **GRANT ALL PRIVILEGES ON TABLE sales TO team**;
- B. GRANT SELECT CREATE MODIFY ON TABLE sales TO team;
- C. GRANT SELECT ON TABLE sales TO team;
- D. GRANT ALL PRIVILEGES ON TABLE team TO sales;

Correct Answer: A

## Get IT Certification

Unlock free, top-quality video courses on ExamTopics with a simple registration. Elevate your learning journey with our expertly curated content. Register now to access a diverse range of educational resources designed for your success. Start learning today with ExamTopics!

[Start Learning for free](#)