EXAM PREPARATION PART-1

Wednesday, April 24, 2024 10:38 AM

Learning Path

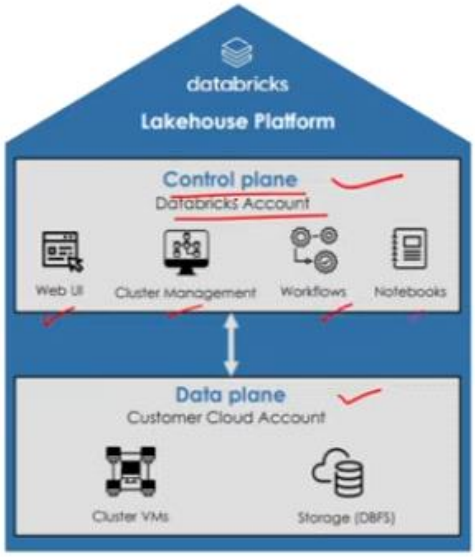
1. Lake house platform
2. ETL with sql and python
3. Incremental data processing - streaming data, auto loader
4. Production pipelines - DLT
5. Data governance - how to access, revoke etc

- Databricks Lakehouse Platform – 24% (11/45)
- ELT with Spark SQL and Python – 29% (13/45)
- Incremental Data Processing – 22% (10/45)
- Production Pipelines – 16% (7/45)
- Data Governance – 9% (4/45)

DATABRICKS LAKE HOUSE PLATFORM

Lake house = Data Lake + Data Warehouse

- 2 Questions on Databricks Architecture



Sample Questions:

Question 1:

Data warehou
+ Data lakehou

Which of the following locations hosts the Databricks web application?

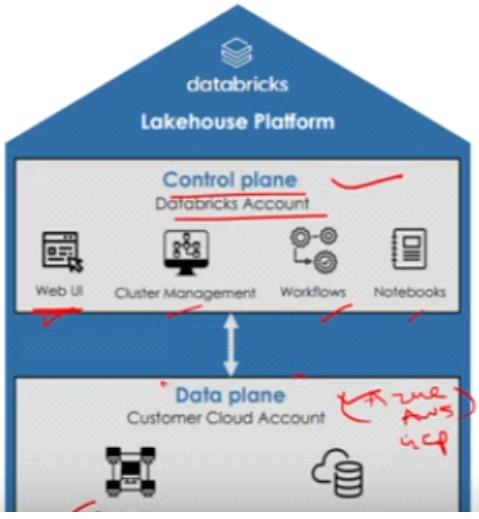
A. Data plane

B. Control plane **(Answer)**

C. Databricks Filesystem

D. Databricks-managed cluster

E. Customer Cloud Account



Question 2:

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

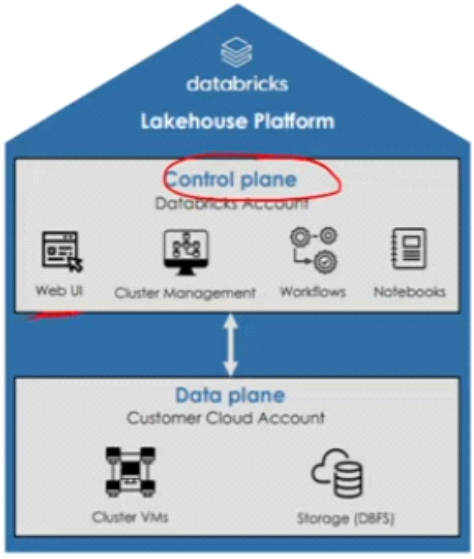
A. Worker node

B. JDBC data source

C. Databricks web application

D. Databricks Filesystem

E. Driver node



Question 3:

According to the Databricks Lakehouse architecture, which of the following is located in the customer's cloud account?

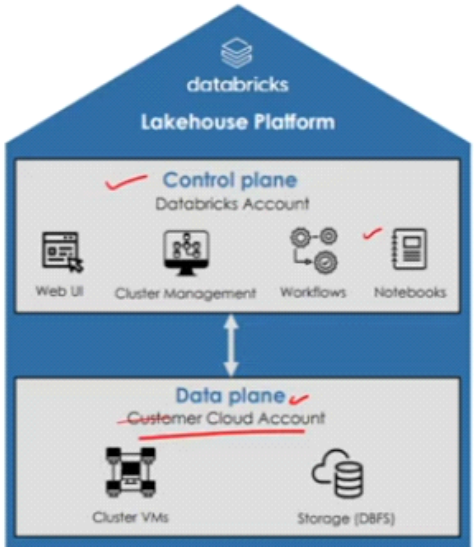
A. Databricks web application

B. Notebooks

C. Repos

D. Cluster virtual machines

E. Workflows



Question 4: (Answer)

Which of the following best describes Databricks Lakehouse?

DW + DL

A. Single, flexible, high-performance system that supports data, analytics, and machine learning workloads.

B. Reliable data management system with transactional guarantees for organization's structured data.

C. Platform that helps reduce the costs of storing organization's open-format data files in the cloud.

D. Platform for developing increasingly complex machine learning workloads using a simple, SQL-based solution.

E. Platform that scales data lake workloads for organizations without investing on-premises hardware.

Question 5:

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

Question 6:

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

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.

Over traditional data lake - ACID properties
Over traditional data warehouse - storage for both structured and unstructured data

Question 7:
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

Question 8:
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.

Lakehouse

Collection
→ Parquet
delta log →

Delta tables are always store in collection of files not a single file

Question 9:

If the default notebook language is SQL, which of the following options a data engineer can use to run a Python code in this SQL Notebook ?

A. They need first to import the python module in a cell

B. This is not possible! They need to change the default language of the notebook to Python

C. Databricks detects cells language automatically, so they can write Python syntax in any cell

D. They can add %language magic command at the start of a cell to force language detection.

E. They can add %python at the start of a cell.

Question 10:

In Delta Lake tables, which of the following is the primary format for the data files?

transaction log
delta log

A. Delta

B. Parquet

C. JSON

D. Hive-specific format

E. Both, Parquet and JSON

Primary format of data files - parquet
Primary format of Transaction log or delta logs - json

We know that in delta log the data is stored as json and crc (binary).

Question 11:

Which of the following commands can a data engineer use to compact small data files of a Delta table into larger ones ?

A. PARTITION BY

B. ZORDER BY

C. COMPACT

D. VACUUM

E. OPTIMIZE (Answer)

create table

insert

2

3

4

delete id=1

time travel

delta log

1 parquet

2

3

4

optimize

Question 12:

select * from table version as of 2

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.

vacuum

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 nun on the table

E. The HISTORY command was run on the table

Question 13:

Which of the following statements is Not true about Delta Lake?

A. Delta Lake provides ACID transaction guarantees

B. Delta Lake provides scalable data and metadata handling

C. Delta Lake provides audit history and time travel

D. Delta Lake builds upon standard data formats: Parquet + XML

E. Delta Lake supports unified streaming and batch data processing

Parquet + xml = ✗

Parquet + Json = ☑

Question 14:

A data engineer is trying to use Delta time travel to rollback a table to a previous version, but the data engineer received an error that the data files are no longer present.

Which of the following commands was run on the table that caused deleting the data files?

A. VACUUM

B. OPTIMIZE

(Answer)

C. ZORDER BY

D. DEEP CLONE

E. DELETE

Question 15:

How long is the default retention period of the VACUUM command ?

A. 0 Days

B. 7 Days (Answer) (168 hours)

C. 30 Days

D. 90 Days

E. 365 Days

Retention Period for VACUUM = 7 days or 168 hours

Databricks Repo - 2 questions

Question 16:

In Databricks Repos, which of the following operations a data engineer can use to update the local version of a repo from its remote Git repository?

A. Clone

B. Commit

C. Merge

D. Push

E. Pull **(Answer)**

User wants to update databricks repo from his github repository - we can just pull

Question 17:

Which of the following tasks is not supported by Databricks Repos, and must be performed in your Git provider?

A. Clone, push to, or pull from a remote Git repository.

B. Create and manage branches for development work.

C. Create notebooks, and edit notebooks and other files.

D. Visually compare differences upon commit.

E. Merge, delete, or rebase branches **(Answer)**

Documentation - https://learn.microsoft.com/en-us/azure/databricks/repos/

What can you do with Databricks Git folders?

Databricks Git folders provides source control for data and AI projects by integrating with Git providers.

In Databricks Git folders, 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.

Can be done outside databricks repos

For following tasks, work in your Git provider:

- Create a pull request.
- Resolve merge conflicts.
- Merge or delete branches.
- Rebase a branch.

Question 18:

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

A. Commit

B. Pull

C. Push

D. Clone

E. Merge **(Answer)**

Question 19:

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 **(Answer)**

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