DataBricks - Associate Data Engineer

Q1. A data engineer needs to develop integration tests for an ETL process and deploy a version-controlled, packaged workflow into production using an external job scheduler. Which tool should the data engineer use for this job?

- A. Databricks Software Development Kit
- B. Databricks Connect
- C. Databricks Asset Bundles
- D. Databricks Command Line Interface

✓ Databricks Asset Bundles

Q2. A data engineer streams customer orders into a Kafka topic (orders_topic) and is currently riting the ingestion script of a DLT pipeline. The data engineer needs to ingest the data from Kafka brokers to DLT using Databricks. What is the correct code for ingesting the data?

A.

```
import dlt @dlt.table
( name = "orders_raw")
def orders_raw():
    return (
    spark.readStream.format("kafka").option("kafka.bootstrap.servers", "broker:9092")
    .option("subscribe", "orders_topic").option("startingOffsets", "earliest").load())
```

B.

```
CREATE STREAMING LIVE TABLE orders_raw AS SELECT value_order_id AS
```

```
order_id, value.customer_id AS customer_id, value.amount AS amount,
value.order_status AS order_status, value.order_timestamp AS order_timestamp
FROM
cloud_files("kafka://broker:9092/orders_topic", "json");
```

```
C.
import dlt @dlt.table( name = "orders_raw")
def orders_raw():
    return (
        spark.readStream.format("cloudFiles").option("cloudFiles.format", "json")
        .option("cloudFiles.schemaLocation", "/schema/location")
        .load("kafka://broker:9092/orders_topic")
        )
```

```
D.
CREATE LIVE TABLE orders_raw AS SELECT CAST(json_data AS STRING) AS
json_data FROM kafka.kafka_broker:9092.orders_topic;
```

✓

```
A.
import dlt @dlt.table
( name = "orders_raw")
def orders_raw():
    return (
    spark.readStream.format("kafka").option("kafka.bootstrap.servers",
"broker:9092")
    .option("subscribe", "orders_topic").option("startingOffsets", "earlie
st").load())
```

Q3) A data engineer is attempting to write Python and SQL in the same command cell and is running into an error. The engineer thought that it was possible to use

a Python variable in a select statement. Why does the command fail?

- A. Databricks supports one language per cell.
- B. Databricks supports language interoperability in the same cell but only between Scala and SQL.
- C. Databricks supports multiple languages but only one per notebook.
- D. Databricks supports language interoperability but only if a special character is used.

> ✓ A. Databricks supports one language per cell

Q4. ***A Databricks single-task workflow fails at the last task due to an error in a notebook. The data engineer fixes the mistake in the notebook. What should the data engineer do to rerun the workflow?***

- A. Repair the task
- B. Rerun the pipeline
- C. Restart the cluster
- D. Switch the cluster

> ✓ A. Repair the task

Q5. A data engineer is processing ingested streaming tables and needs to filter out NULL values in the order_datetime column from the raw streaming table orders_raw and store the results in a new table orders_valid using DLT. Which code snippet should the data engineer use?

● A. CREATE OR REFRESH STREAMING LIVE TABLE orders_valid CONSTRAINT valid_date EXPECT (order_datetime IS NOT NULL) ON VIOLATION DROP ROW A
S
SELECT * FROM orders_raw;

● B. CREATE OR REFRESH STREAMING LIVE TABLE orders_valid EXPECT (order_datetime IS NOT NULL) ON VIOLATION DROP ROW AS SELECT * FROM STREAM orders_raw;

● C. CREATE OR REFRESH STREAMING TABLE orders_valid AS SELECT * FROM STREAM orders_raw WHERE order_datetime IS NOT NULL;

● D. CREATE OR REPLACE STREAMING TABLE orders_valid (FILTER (order_dat
etime
IS NOT NULL)) AS SELECT * FROM STREAM orders_raw;

✓ CREATE OR REFRESH STREAMING LIVE TABLE orders_valid
EXPECT (order_datetime IS NOT NULL) ON VIOLATION DROP ROW
AS SELECT * FROM STREAM orders_raw;

Q6. ***A data engineer needs to provide access to a group named manufacturing-team. The team needs privileges to create tables in the quality schema. Which set of SQL commands will grant a group named manufacturing-team to create tables in a schema named production with the parent catalog named manufacturing with the least privileges?***

● A. GRANT CREATE TABLE ON SCHEMA manufacturing.quality TO
manufacturing-team;
GRANT CREATE SCHEMA ON SCHEMA manufacturing.quality
TO manufacturing-team;

GRANT CREATE CATALOG ON CATALOG manufacturing TO manufacturing-team;

● B. GRANT CREATE TABLE ON SCHEMA manufacturing.quality TO manufacturing-team;
GRANT CREATE SCHEMA ON SCHEMA manufacturing.quality TO manufacturing-team;
GRANT USE CATALOG ON CATALOG manufacturing TO manufacturing-team;

● C. GRANT USE TABLE ON SCHEMA manufacturing.quality TO manufacturing-team;
GRANT USE SCHEMA ON SCHEMA manufacturing.quality TO manufacturing-team;
GRANT USE CATALOG ON CATALOG manufacturing TO manufacturing-team;

● D. GRANT CREATE TABLE ON SCHEMA manufacturing.quality TO manufacturing-team;
GRANT USE SCHEMA ON SCHEMA manufacturing.quality TO manufacturing-team;
GRANT USE CATALOG ON CATALOG manufacturing TO manufacturing-team;

Q7. A Python file is ready to go into production and the client wants to use the cheapest but most efficient type of cluster possible. The workload is quite small, only processing 10GBs of data
with only simple joins and no complex aggregations or wide transformations.
Which cluster meets the requirement?

● A. Interactive cluster
● B. Job cluster with spot instances enabled
● C. Job cluster with spot instances disabled
● D. Job cluster with Photon enabled

✅ **Correct answer: B. Job cluster with spot instances enabled**

Q8. What is the functionality of Auto Loader in Databricks?

● A. Auto Loader automatically ingests and processes new files from cloud storage, handling only streaming data with no support for schema evolution.
● B. Auto Loader automatically ingests and processes new files from cloud storage, handling batch data with support for schema evolution.
● C. Auto Loader automatically ingests and processes new files from cloud storage, handling both batch and streaming data with no support for schema evolution.
● D. Auto Loader automatically ingests and processes new files from cloud storage, handling both batch and streaming data with support for schema evolution.

✅ **Correct answer: D**

Q9. An organization has implemented a data pipeline in Databricks and needs to ensure it can scale automatically based on varying workloads without manual cluster management. The goal
is to meet the company's Service Level Agreements (SLAs), which require high availability and minimal downtime, while Databricks automatically handles resource allocation and optimization.
Which approach fulfills these requirements?

● A. Use Serverless compute in Databricks to automatically scale and provision resources
with minimal manual intervention.
● B. Use Interactive Clusters in Databricks, adjusting cluster sizes manually based on
workload demands.
● C. Use Spot Instances to allocate resources dynamically while minimizing costs, with
potential interruptions.
● D. Deploy Job Clusters with fixed configurations, dedicated to specific tasks, without
automatic scaling.

✅ **Correct answer: A. Use Serverless compute in Databricks**

Q10. A data engineer is reviewing the documentation on audit logs in Databricks for compliance purposes and needs to understand the format in which audit logs output events. How are events formatted in Databricks audit logs?

- A. In Databricks, audit logs output events in an XML format.
- B. In Databricks, audit logs output events in a JSON format.
- C. In Databricks, audit logs output events in a CSV format.
- D. In Databricks, audit logs output events in a plain text format.

✅ **B. JSON** → Correct.

Q11. An organization needs to share a dataset stored in its Databricks Unity Catalog with an
external partner who uses a different data platform that is not Databricks. The goal is to maintain data security and ensure the partner can access the data efficiently. Which method should the
data engineer use to securely share the dataset with the external partner?

- A. Using Delta Sharing with the open sharing protocol
- B. Exporting data as CSV files and emailing them
- C. Using a third-party API to access the Delta table
- D. Databricks-to-Databricks Sharing

✅ **Correct answer: A. Using Delta Sharing with the open sharing protocol**

Q12. A data engineer wants to create an external table in Databricks that references data stored in an Azure Data Lake Storage (ADLS) location. The goal is to enable Databricks to access and query this external data without moving it into the Databricks-managed storage. Which step should the data engineer take to successfully create the external table?

- A. Use the CREATE TABLE statement and specify the LOCATION clause with the path to the external data.

● B. Use the CREATE MANAGED TABLE statement and specify the LOCATION clause
with the path to the external data.

● C. CREATE UNMANAGED TABLE statement without specifying a LOCATION clause.

● D. CREATE EXTERNAL TABLE statement without specifying a LOCATION clause.

✅ **Correct answer: A. Use the CREATE TABLE statement and specify the LOCATION clause with the path to the external data.**

Q13. A data engineer is designing an ETL pipeline to process both streaming and batch data from multiple sources. The pipeline must ensure data quality, handle schema evolution, and provide easy maintenance. The team is considering using Delta Live Tables (DLT) in Databricks to achieve these goals. They want to understand the key features and benefits of DLT that make it suitable for this use case. Why is Delta Live Tables (DLT) an appropriate choice?

● A. Automatic data quality checks, built-in support for schema evolution, and declarative
pipeline development

● B. Manual schema enforcement, high operational overhead, and limited scalability

● C. Requires custom code for data quality checks, no support for streaming data, and
complex pipeline maintenance

● D. Supports only batch processing, no data versioning, and high infrastructure costs

✅ **Correct answer: A. Automatic data quality checks, built-in support for schema evolution, and declarative pipeline development**

Q14. Which TWO items are characteristics of the Gold Layer? (Choose 2 answers)

● A. Historical lineage

● B. Normalised
● C. Read-optimized
● D. De-normalised
● E. Raw Data

✅ Correct answers: **C. Read-optimized** and **D. De-normalised**

Q15. A data engineer is managing a data pipeline in Databricks, where multiple Delta tables are used for various transformations. The team wants to track how data flows through the pipeline, including identifying dependencies between Delta tables, notebooks, jobs, and dashboards. The data engineer is utilizing the Unity Catalog lineage feature to monitor this process. How does Unity Catalog's data lineage feature support the visualization of relationships between Delta tables, notebooks, jobs, and dashboards?

● A. Unity Catalog provides an interactive graph that visualizes the dependencies between
Delta tables, notebooks, jobs, and dashboards, while also supporting column-level tracking of data transformations.
● B. Unity Catalog lineage only supports visualizing relationships at the table level and
does not extend to notebooks, jobs, or dashboards.
● C. Unity Catalog lineage provides an interactive graph that tracks dependencies between tables and notebooks but excludes any job-related dependencies or dashboard
visualizations.
● D. Unity Catalog lineage visualizes dependencies between Delta tables, notebooks, and
jobs, but does not provide column-level tracing or relationships with dashboards.

✅ **Correct answer: A.**

Q16. A data engineer is developing a small proof of concept in a notebook. When running the entire notebook, the Cluster usage spikes. The data engineer wants

to keep the development requirements and get real-time results. Which Cluster meets these requirements?

- A. Job Cluster with autoscaling enabled
- B. All-Purpose Cluster with a large fixed memory size
- C. All-Purpose Cluster with autoscaling
- D. Job cluster with Photon enabled and autoscaling

✅ **Correct answer: C. All-Purpose Cluster with autoscaling**

Q17. A data engineering team is using Kafka to capture event data and then ingest it into Databricks. The team wants to be able to see these historical events. Medallion architecture is already in place. The team wants to be mindful of costs. Where should this historical event data be stored?

- A. Gold
- B. Silver
- C. Bronze
- D. Raw layer

✅ **Correct answer: C. Bronze**

Q18. A data engineer at a company that uses Databricks with Unity Catalog needs to share a collection of tables with an external partner who also uses a Databricks workspace enabled for Unity Catalog. The data engineer decides to use Delta Sharing to accomplish this. What is the first piece of information the data engineer should request from the external partner to set up Delta Sharing?

- A. The sharing identifier of their Unity Catalog metastore
- B. Their Databricks account password

● C. The name of their Databricks cluster
● D. The IP address of their Databricks workspace

✅ **Correct answer: A**

Q19. A company is collaborating with a partner that does not use Databricks but needs access to a large historical dataset stored in Delta format. The data engineer needs to ensure that the partner can access the data securely, without the need for them to set up an account, and with read-only access. How should the data be shared?

● A. Share the dataset using Delta Sharing, which allows your partner to access the data
using a secure, read-only URL without requiring a Databricks account, ensuring that they
cannot modify the data.
● B. Share the dataset using Unity Catalog, ensuring that both teams have full write access to the data within the same organization.
● C. Share the dataset by exporting it to a CSV file and manually transferring the file to the
partner's system.
● D. Grant your partner access to your Databricks workspace and assign them full write
permissions to the Delta table, enabling them to modify the dataset.

✅ **Correct answer: A**

Q20. A data engineer needs to parse only png files in a directory that contains files with different
suffixes. Which code should the data engineer use to achieve this task?

● A. df = spark.readStream.format("cloudFiles") .option("cloudFiles.format", "binaryFile")
.load("*/.png")*
● *B. df = spark.readStream.format("cloudFiles") .option("cloudFiles.format", "binaryFile")*

.option("pathGlobFilter", ".png") .load("<base-path>")

● C. df = spark.readStream.format("cloudFiles") .option("cloudFiles.format", "binaryFile")
.append("/.png")

● D. df = spark.readStream.format("cloudFiles") .option("cloudFiles.format", "binaryFile")
.option("pathGlobFilter", ".png") .append()

✅ **Correct answer: B.**

***Q21. A company uses Delta Sharing to collaborate with partners across different cloud providers
and geographic regions. What will result in additional costs due to cross-region or egress fees?***

● A. Accessing Delta Sharing data using a VPN within the same data center

● B. Transferring data via Delta Sharing across clouds and across different geographic
regions

● C. Sharing data within the same cloud provider and region

● D. Utilizing Delta Sharing for internal data analytics within a single cloud environment

✅ **Correct answer: B.**

***Q22. A data engineer has been provided a PySpark DataFrame named df with columns product and revenue. The data engineer needs to compute complex aggregations to determine each product's total revenue, average revenue, and transaction count. Which code snippet should the
data engineer use?***

● A. from pyspark.sql import functions as F aggregated_
df = df.groupBy("product").agg(F.sum("revenue").alias("total_revenue"),
F.avg("revenue").alias("avg_revenue"), F.count("*").alias("transaction_count"))

● B. aggregated_df = df.groupBy("product").agg(sum("revenue"), "avg(revenue)", "count(revenue)")
● C. from pyspark.sql import functions as F aggregated_df = df.select("product", "revenue").groupBy("product").agg(F.sum("revenue"), F.mean("revenue"))
● D. aggregated_df = df.groupBy("product").agg(f"revenue": "sum", "revenue": "avg",
"revenue": "count"))

✅ **Correct answer: A**

Q23. A data engineer needs to ingest from both streaming and batch sources for a firm that relies on highly accurate data. Occasionally, some of the data picked up by the sensors that provide a streaming input are outside the expected parameters. If this occurs, the data must be dropped, but the stream should not fail. Which feature of Delta Live Tables meets this requirement?

● A. Change Data Capture
● B. Error Handling
● C. Expectations
● D. Monitoring

✅ **Correct answer: C. Expectations**

Q24. A data engineer is inspecting an ETL pipeline based on a PySpark job that consistently encounters performance bottlenecks. Based on developer feedback, the data engineer assumes the job is low on compute resources. To pinpoint the issue, the data engineer observes the Spark UI and finds out the job has a high CPU time vs Task time. Which course of action should the data engineer take?

● A. High CPU time vs Task time means a CPU over-utilized job. The data engineer may
need to consider executor and core tuning or resizing the cluster.
● B. High CPU time vs Task time means an under-utilized cluster. The data engineer may

need to repartition data to spread the jobs more evenly throughout the cluster.
● C. High CPU time vs Task time means over-utilized memory and the need to increase
parallelism
● D. High CPU time vs Task time means efficient use of cluster and no change needed

✅ **Correct answer: A**

Q25. An organization has data stored across multiple external systems, including MySQL, Amazon Redshift, and Google BigQuery. The data engineer wants to perform analytics without ingesting directly into Databricks, ensuring unified governance and minimizing data duplication. Which feature of Databricks enables querying these external data sources while maintaining centralized governance?

● A. Databricks Connect
● B. MLflow
● C. Delta Lake
● D. Lakehouse Federation

✅ **Correct answer: D. Lakehouse Federation**

Q26. An organization is looking for an optimized storage layer that supports ACID transactions and schema enforcement. Which technology should the organization use?

● A. Delta Lake
● B. Unity Catalog
● C. Data lake
● D. Cloud File Storage

✅ **Correct answer: A. Delta Lake**

Q27. A data engineer is writing a script that is meant to ingest new data from cloud storage. In the event of the Schema change, the ingestion should fail. It should fail until the changes downstream source can be found and verified as intended changes. Which command will meet the requirements?

● A. addNewColumns

● B. rescue

● C. failOnNewColumns

● D. none

✅ **Correct answer: C. failOnNewColumns**

Q28. A data engineer is working on a personal laptop and needs to perform complex transformations on data stored in a Delta Lake on cloud storage. The engineer decides to use Databricks Connect to interact with Databricks clusters and work in their local IDE. How does Databricks Connect enable the engineer to develop, test, and debug code seamlessly on their
local machine while interacting with Databricks clusters?

● A. By providing a local environment that mimics the Databricks runtime, enabling the
engineer to develop, test, and debug code using a specific tool that is required by Databricks

● B. By providing a local environment that mimics the Databricks runtime, enabling the
engineer to develop, test, and debug code only through Databricks web interface

● C. By allowing direct execution of Spark jobs from the local machine without needing a
network connection

● D. By providing a local environment that mimics the Databricks runtime, enabling the
engineer to develop, test, and debug code using their preferred IDE

✅ **Correct answer: D**

Q29. What is the maximum output supported by a job cluster to ensure a notebook does not fail?

- A. 15MBs
- B. 10MBs
- C. 30MBs
- D. 25MBs

✅ **Correct answer: D. 25MBs** → 30

Q30. A global retail company sells products across multiple categories (e.g., Electronics, Clothing)
and regions (e.g., North, South, East, West). The sales team has provided a PySpark dataframe
named sales_df and the data engineer wants to analyze the sales data to help make strategic
decisions. Given the sales_df: sales_df | product_id | category | sales_amount | region | ...
Calculate the total sales amount for each product category and store the results in a new
dataframe called category_sales. What will generate the expected result of category_sales?
category_sales | category | total_sales_amount | ... | Electronics | 500 | | Clothing | 900 |

- A. category_sales = sales_df.groupBy("category").agg(sum("sales_amount").alias("total_sales_amount"))
- B. category_sales = sales_df.sum("sales_amount").groupBy("category").alias("total_sales_amount")
- C. category_sales = sales_df.agg(sum("sales_amount").groupBy("category").alias("total_sales_amount")
- D. category_sales = sales_df.groupBy("region").agg(sum("sales_amount").alias("total_sales_amount"))

✅ **Correct answer: A**

Q31. A Databricks workflow fails at the last stage due to an error in a notebook. This workflow runs daily. The data engineer fixes the mistake and wants to rerun the pipeline in order to minimize downtime and cost? Which action should the data engineer do?

- A. Re-run the entire workflow
- B. Switch to another cluster
- C. Restart the cluster
- D. Repair run

✅ **Correct answer: D**

Q33. Which compute option should be chosen in a scenario where small-scale ad-hoc Python scripts need to be run at high frequency and should wind down quickly after these queries have finished running?

- A. All-Purpose Cluster
- B. Job Cluster
- C. Serverless Compute
- D. SQL Warehouse

✅ **Correct answer: B**

Q34. A data engineer is developing an ETL process based on Spark SQL. The execution fails. The data engineer checks the Spark UI and can see the ERRORS as follows: "java.lang.OutOfMemoryError: Java heap space" Which two corrective actions should the data engineer perform to resolve this issue? (Choose 2 answers)

- A. Narrow the filters in order to collect less data in the query
- B. Upsize the worker nodes and activate autoshuffle partitions
- C. Upsize the driver node and deactivate autoshuffle partitions

- D. Cache the dataset in order to boost the query performance
- E. Fix the shuffle partitions to 50 to ensure the allocation

✅ **Correct answers: A and B**

Q35. Which SQL code snippet will correctly demonstrate a Data Definition Language (DDL)
operation used to create a table?

- A. CREATE TABLE employees (id INT, name STRING, salary DECIMAL(10,2));
- B. ALTER TABLE employees ADD COLUMN new_column INT;
- C. INSERT INTO employees VALUES (1, 'Alice');
- D. SELECT * FROM employees;

✅ **Correct answer: A**

Q36. A data engineer has written a function in a Databricks Notebook to calculate the population of bacteria in a given medium. Analysts use this function in the notebook and sometimes provide input arguments of the wrong data type, which can cause errors during execution. Which Databricks feature will help the data engineer quickly identify if an incorrect data type has
been provided as input?

- A. The Spark User interface has a debug tab that contains the variables that are used in
this session.
- B. The Databricks debugger enables the use of a variable explorer to see at a glance the
value of the variables.
- C. The Data Engineer should add print statements to find out what the variable is.
- D. The Databricks debugger enables breakpoints that will raise an error if the wrong data
type is submitted

✅ **Correct answer: B. The Databricks debugger enables the use of a variable explorer to see at a glance the value of the variables**

Q37. A data engineer is debugging a Python notebook in Databricks that processes a dataset
using PySpark. The notebook fails with an error during a DataFrame transformation. The engineer wants to inspect the state of variables, such as the input DataFrame and intermediate results, to identify where the error occurs. Which tool should the engineer use to debug the notebook and inspect the values of variables like DataFrames?

● A. Use the Spark UI to analyze the execution plan and identify stages where the job
failed
● B. Use the Databricks CLI to download and analyze driver logs for detailed error messages
● C. Use the Ganglia UI to monitor cluster resource usage and identify hardware issues
● D. Use the Python Notebook Interactive Debugger to set breakpoints and inspect variable values in real-time

✅ **Correct answer: D**

Q38. A data engineer needs to conduct Exploratory Analysis on data residing in a database that is within the company's custom-defined network in the cloud. The data engineer is using SQL for this task. Which type of SQL Warehouse will enable the data engineer to process large numbers of queries quickly and cost-effectively?

● A. Pro SQL Warehouse
● B. Classic SQL Warehouse
● C. Serverless SQL Warehouse
● D. Serverless compute for notebooks

✅ **Correct answer: A. Pro SQL Warehouse**

Q39. What are the transformations typically included in building the Bronze layer?

● A. Business rules and transformations
● B. Include columns Load date/time, process ID
● C. Aggregate data from multiple sources
● D. Perform extensive data cleansing

✅ **Correct answer: B.**

Q40. A data engineer is working on a Databricks project that utilizes cloud storage. The data engineer wants to load several JSON files from containers on a storage account as soon as the file arrives within the storage account. Which syntax should the data engineer follow to first load the files into a dataframe and check that it is working as expected using Python?

● A. df = spark.read.json("/input/path")
● B. df = spark.readStream.format("cloudFiles").option("cloudFiles.format", "json").load("/input/path")
● C. df = spark.readStream.format("cloudFiles").option("json").load("/input/path")
● D. df = spark.read.format("cloud").option("json").load("/input/path")

✅ **Correct answer: B**

Q41. ***A data engineer is configuring Unity Catalog in Databricks and needs to assign a role to a user who should have the ability to grant and revoke privileges on various data objects within a***
specific schema, but should not have read/write access over the schema or its objects. Which
role should the data engineer assign to this user?

- A. Table Owner
- B. USE catalog/schema privilege on the schema
- C. Schema Owner
- D. Catalog Owner

✅ **Correct answer: C. Schema Owner**

Q42. A Data Engineer is building a simple data pipeline using Delta Live Tables (DLT) to ingest customer data. The raw customer data is stored in a cloud storage location in JSON format. The
task is to create a DLT pipeline that reads the raw JSON data and writes it into a Delta table for
further processing. Which code snippet will correctly ingest the raw JSON data and create a
Delta table using DLT?

- A. import dlt @dlt.view def raw_customers(): return spark.format.json("s3://my-bucket/raw-customers/")
- B. import dlt @dlt.table def raw_customers(): return spark.read.format("parquet").load("s3://my-bucket/raw-customers")
- C. import dlt @dlt.table def raw_customers(): return spark.read.format("csv").load("s3://my-bucket/raw-customers/")
- D. import dlt @dlt.table
def raw_customers():
return spark.read.json("s3://my-bucket/raw-customers/")

✅ **Correct answer: D**

Q43. A data engineer wants to reduce costs and optimize cloud spending. The data engineer has
decided to use Databricks Serverless for lowering cloud costs while maintaining existing SLAs.
What is the first step in migrating to Databricks Serverless?

● A. Low frequency BI Dash-boarding and Adhoc SQL Analytics
● B. Legacy Ingestion pipelines that include ingestion from sources APIs, files, JDBC/ODBC connections
● C. A frequently running and efficient Python-based data transformation pipeline compatible with the latest Databricks runtime and Unity Catalog
● D. A frequently running and efficient Scala-based data transformation pipeline compatible with the latest Databricks runtime and Unity Catalog

✅ **Correct answer: A**

***Q44. What Databricks feature can be used to check the data sources and tables used in a
workspace?***

● A. Do not use the lineage feature as it only tracks activity from the last 3 months and will
not provide full details on dependencies.
● B. Use the lineage feature to visualize a graph that highlights where the table is used
only in notebooks.
● C. Use the lineage feature to visualize a graph that highlights where the table is used
only in reports.
● D. Use the lineage feature to visualize a graph that shows all dependencies, including
where the table is used in notebooks, other tables and reports.

✅ **Correct answer: D**

Q45. What is the primary function of the Silver layer in the Databricks medallion architecture?

● A. Store historical data solely for auditing purposes

- B. Digest raw data in its original state
- C. Aggregate and enrich data for business analytics
- D. Validate, clean, and deduplicate data for further processing

✅ **Correct answer: D**

Q46. A data engineer needs to optimize the data layout and query performance for a customers table which is partitioned by "purchase_date", a DATE column which helps with time-based queries but not optimize queries on a "customer_id", a high-cardinality column. The table is usually queried with filters on "customer_id" within specific date ranges, but once the data is spread across each partition, it results in full partition scans and increased runtime and costs. How should the data engineer optimize the Data Layout for efficient reads?

- A. Alter the table implementing liquid clustering by "customer_id" and "purchase_date".
- B. Enable delta caching on the cluster so that frequent reads are cached for performance.
- C. Alter table implementing liquid clustering on "customer_id" while keeping the existing
partitioning.
- D. Alter the table to partition by "customer_id".

✅ **Correct answer: C. Alter table implementing liquid clustering on `customer_id` while keeping the existing partitioning**

Q47. A data engineer needs to combine sales data from an on-premises PostgreSQL database with customer data in Azure Synapse to create a comprehensive report. The goal is to avoid data duplication and ensure up-to-date information. How should the data engineer achieve this using Databricks?

- A. Develop custom ETL pipelines to ingest data into Databricks
- B. Export data from both sources to CSV files and upload them to Databricks

● C. Manually synchronize data from both sources into a single database
● D. Use Lakehouse Federation to query both data sources directly

✅ **Correct answer: D**

Q48. An organization plans to share a large dataset stored in a Databricks workspace on AWS with a partner organization whose Databricks workspace is hosted on Azure. The data engineer wants to minimize data transfer costs while ensuring a secure and efficient data sharing. Which strategy will reduce data egress costs associated with cross-cloud data sharing?

● A. Using Delta Sharing without any additional configurations
● B. Configure VPN connection between AWS and Azure for faster data sharing
● C. Migrating the dataset to Cloudflare R2 object storage before sharing
● D. Sharing data via pre-signed URLs without monitoring egress costs

✅ **Correct answer: C**

Q49. What is the structure of an Asset Bundle?

● A. A YAML configuration file that specifies the artifacts, resources, and configurations for
the project
● B. A single plain text file enumerating the names of assets to be migrated to a new workspace
● C. A compressed archive (ZIP) that solely contains workspace assets without any accompanying metadata
● D. A Docker image containing runtime environments and the source code of the assets

✅ **Correct answer: A**

Q50. A data engineer is setting up access control in Unity Catalog and needs to ensure that a group of data analysts can query tables but not modify data. Which permission should the data engineer grant to the data analysts?

- A. INSERT
- B. MODIFY
- C. ALL PRIVILEGES
- D. SELECT

✅ **Correct answer: D. SELECT**

Q51. A data engineer works for an organization that must meet a stringent Service Level Agreement (SLA) that demands minimal runtime errors and high availability for its data processing pipelines. The data engineer wants to avoid the operational overhead of managing and scaling clusters. Which architectural solution will meet the requirements?

- A. Utilize Databricks serverless compute that automatically optimizes resources and
abstracts cluster management.
- B. Implement a hybrid approach with scheduled batch jobs on custom cloud VMs.
- C. Use an auto-scaling cluster configured and monitored by the user.
- D. Deploy a dedicated, manually managed cluster optimized by in-house IT staff.

✅ **Correct answer: A**

Q52. A data engineer is maintaining an ETL pipeline code with a GitHub repository linked to their Databricks account. The data engineer wants to deploy the ETL pipeline to production as a Databricks workflow. Which approach should the data engineer use?

- A. Manually create and manage the workflow in UI
- B. Maintain workflow_config.json and deploy it using Databricks CLI

● C. Maintain workflow_config.json and deploy it using Terraform
● D. Databricks Asset Bundles (DAB) + Github Integration

✅ **Correct answer: D.**

Q53. A data engineer manages multiple external tables linked to various data sources. The data engineer wants to manage these external tables efficiently and ensure that only the necessary permissions are granted to users for accessing specific external tables. How should the data engineer manage access to these external tables?

● A. Grant permissions on the Databricks workspace level, which will automatically apply
to all external tables.
● B. Create a single user role with full access to all external tables and assign it to all
users.
● C. Use Unity Catalog to manage access controls and permissions for each external table
individually.
● D. Set up Azure Blob Storage permissions at the container level, allowing access to all
external tables

✅ **Correct answer: C**

Q54. A data engineer is getting a partner organization up to speed with Databricks account. Both the teams share same business use cases. The data engineer has to share some of your Unity- catalog managed delta tables and the notebook pies creating those tables with the partner organization. How can the data engineer seamlessly share the required information?

● A. Share access to codebase via Github and allow them to ingest datasets from your

Data lake.

- B. Share required datasets and notebooks via Delta Sharing. Manage permissions via Unity Catalog.
- C. Zip all the code and share via email and allow data ingestion from your data lake
- D. Data and Notebooks can be shared simply using Unity Catalog

✅ **Correct answer: B**

Q55. A data engineering project involves processing large batches of data on a daily schedule using ETL. The jobs are resource-intensive and vary in size, requiring a scalable, cost-efficient compute solution that can automatically scale based on the workload. Which compute approach will satisfy the needs described?

- A. Databricks SQL Serverless
- B. All-Purpose Cluster
- C. Job Cluster
- D. Dedicated Cluster

✅ **Correct answer: C. Job Cluster**

Q56. A data engineer is working in a Python notebook on Databricks to process data, but notices that the output is not as expected. The data engineer wants to investigate the issue by stepping through the code and checking the values of certain variables during execution. Which tool should the data engineer use to inspect the code execution and variables in real-time?

- A. Python Notebook Interactive Debugger
- B. SQL Analytics

- C. Job Execution Dashboard
- D. Cluster Logs

✅ **Correct answer: A.**