AWS Certified Data Engineer – Associate

DEA – C01

Disclaimer: This content is for educational purposes. Accuracy is attempted but not guaranteed.

No job outcome is promised.

Question: 1
A data engineer needs to optimize the performance of a data pipeline that handles retail orders. Data about the orders is ingested daily into an Amazon S3 bucket.
The data engineer runs queries once each week to extract metrics from the orders data based the order date for multiple date ranges. The data engineer needs an optimization solution that ensures the query performance will not degrade when the volume of data increases.
Which solution will meet this requirement MOST cost-effectively?

Options
A. Partition the data based on order date. Use Amazon Athena to query the data.
B. Partition the data based on order date. Use Amazon Redshift to query the data.
C. Partition the data based on load date. Use Amazon EMR to query the data.
D. Partition the data based on load date. Use Amazon Aurora to query the data.

Answer: A

Explanation:
Partitioning by order date allows Athena to scan only the relevant partitions, improving query performance as data volume grows. Athena is serverless and cost-effective for weekly queries.

- B: Redshift is powerful but overkill and more expensive for weekly queries.
- C: EMR adds unnecessary complexity and cost.
- D: Aurora is not designed for large-scale analytical queries.

Question: 2
A data engineer has two datasets that contain sales information for multiple cities and states. One dataset is named reference, and the other dataset is named primary.
The data engineer needs a solution to determine whether a specific set of values in the city and state columns of the primary dataset exactly match the same specific values in the reference dataset. The data engineer wants to use Data Quality Definition Language (DQDL) rules in an AWS Glue Data Quality job.
Which rule will meet these requirements?

Options
A. DatasetMatch "reference" “city->ref_city, state->ref_state” = 1.0
B. ReferentialIntegrity “city,state” “reference. ref_city,ref_state ” = 1.0
C. DatasetMatch “reference” “city->ref_city, state->ref_state” = 100
D. Referentialintegrity “city,state” "reference. ref_city,ref_state ” = 100

Answer: B

Explanation:
ReferentialIntegrity checks ensure that values in one dataset match values in another dataset, which is the requirement here.

- A, C: DatasetMatch is for comparing entire datasets, not validating specific column referential integrity.
- D: Incorrect syntax (100 instead of 1.0).

Question: 3
A company has an on-premises PostgreSQL database that contains customer data. The company wants to migrate the customer data to an Amazon Redshift data warehouse. The company has established a VPN connection between the on-premises database and AWS. The on-premises database is continuously updated. The company must ensure that the data in Amazon Redshift is updated as quickly as possible.
Which solution will meet these requirements?

Options
A. Use the pg_dump utility to generate a backup of the PostgreSQL database. Use the AWS Schema Conversion Tool (AWS SCT) to upload the backup to Amazon Redshift. Set up a cron job to perform a backup. Upload the backup to Amazon Redshift every night.
B. Create an AWS Database Migration Service (AWS DMS) full-load task. Set Amazon Redshift as the target. Configure the task to use the change data capture (CDC) feature.
C. Use the pg_dump utility to generate a backup of the PostgreSQL database. Upload the backup to an Amazon S3 bucket. Use the COPY command to import the data into Amazon Redshift.
D. Create an AWS Database Migration Service (AWS DMS) full-load task. Set Amazon Redshift as the target. Configure the task to perform a full load of the database to Amazon Redshift every night.

Answer: B

Explanation:
DMS with CDC continuously replicates changes from PostgreSQL to Redshift, ensuring near real-time updates.

- A, C, D: These rely on batch/full loads and do not meet the requirement for continuous updates.

Question: 4
A company has several new datasets in CSV and JSON formats. A data engineer needs to make the data available to a team of data analysts who will analyze the data by using SQL queries.
Which solution will meet these requirements in the MOST cost-effective way?

Options
A. Create an Amazon RDS MySQL cluster. Use AWS Glue to transform and load the CSV and JSON files into database tables. Provide the data analysts access to the MySQL cluster.
B. Create an AWS Glue DataBrew project that contains the new data. Make the DataBrew project available to the data analysts.
C. Store the data in an Amazon S3 bucket. Use an AWS Glue crawler to catalog the S3 bucket as tables. Create an Amazon Athena workgroup that has a data usage threshold. Grant the data analysts access to the Athena workgroup.
D. Load the data into Super-fast, Parallel, In-memory Calculation Engine (SPICE) in Amazon QuickSight. Allow the data analysts to create analyses and dashboards in QuickSight.

Answer: C

Explanation:
Athena with Glue crawler is cost-effective, serverless, and allows direct querying on S3 without heavy infrastructure.

- A: RDS adds cost and management overhead.
- B: DataBrew is for data preparation, not analyst SQL queries.
- D: QuickSight SPICE is for visualization, not raw SQL queries.

Question: 5
A retail company stores order information in an Amazon Aurora table named Orders. The company needs to create operational reports from the Orders table with minimal latency. The Orders table contains billions of rows, and over 100,000 transactions can occur each second. A marketing team needs to join the Orders data with an Amazon Redshift table named Campaigns in the marketing team's data warehouse. The operational Aurora database must not be affected.
Which solution will meet these requirements with the LEAST operational effort?

Options
A. Use AWS Database Migration Service (AWS DMS) Serverless to replicate the Orders table to Amazon Redshift. Create a materialized view in Amazon Redshift to join with the Campaigns table.
B. Use the Aurora zero-ETL integration with Amazon Redshift to replicate the Orders table. Create a materialized view in Amazon Redshift to join with the Campaigns table.
C. Use AWS Glue to replicate the Orders table to Amazon Redshift. Create a materialized view in Amazon Redshift to join with the Campaigns table.
D. Use federated queries to query the Orders table directly from Aurora. Create a materialized view in Amazon Redshift to join with the Campaigns table.

Answer: B

Explanation:
Aurora zero-ETL with Redshift provides near real-time replication with minimal setup and operational effort.

- A: DMS works but requires setup and management, more overhead.
- C: Glue ETL adds latency and overhead.
- D: Federated queries increase Aurora load, violating the requirement.

Question: 6
A company is building a new application that ingests CSV files into Amazon Redshift. The company has developed the frontend for the application.
The files are stored in an Amazon S3 bucket. Files are no larger than 5 MB.
A data engineer is developing the extract, transform, and load (ETL) pipeline for the CSV files. The data engineer configured a Redshift cluster and an AWS Lambda function that copies the data out of the files into the Redshift cluster.
Which additional steps should the data engineer perform to meet these requirements?

Options
A. Configure the bucket to send S3 event notifications to Amazon EventBridge. Configure an EventBridge rule that matches S3 new object created events. Set the Lambda function as the target.
B. Configure the S3 bucket to send S3 event notifications to an Amazon Simple Queue Service (Amazon SQS) queue. Configure the Lambda function to process the queue.
C. Configure AWS Database Migration Service (AWS DMS) to stream new S3 objects to a data stream in Amazon Kinesis Data Streams. Set the Lambda function as the target of the data stream.
D. Configure an Amazon EventBridge rule that matches S3 new object created events. Set an Amazon Simple Queue Service (Amazon SQS) queue as the target of the rule. Configure the Lambda function to process the queue.

Answer: A

Explanation:
Using S3 event notifications with EventBridge triggers Lambda directly on object creation, providing a simple and cost-effective design.

- B: SQS adds an extra layer without benefit here.
- C: DMS is unnecessary for S3 ingestion.
- D: Adds complexity by routing via SQS unnecessarily.

Question: 7
A company stores sensitive data in an Amazon Redshift table. The company needs to give specific users the ability to access the sensitive data. The company must not create duplication in the data.
Customer support users must be able to see the last four characters of the sensitive data. Audit users must be able to see the full value of the sensitive data. No other users can have the ability to access the sensitive information.
Which solution will meet these requirements?

Options
A. Create a dynamic data masking policy to allow access based on each user role. Create IAM roles that have specific access permissions. Attach the masking policy to the column that contains sensitive data.
B. Enable metadata security on the Redshift cluster. Create IAM users and IAM roles for the customer support users and the audit users. Grant the IAM users and IAM roles permissions to view the metadata in the Redshift cluster.
C. Create a row-level security policy to allow access based on each user role. Create IAM roles that have specific access permissions. Attach the security policy to the table.
D. Create an AWS Glue job to redact the sensitive data and to load the data into a new Redshift table.

Answer: A

Explanation:
Dynamic data masking in Redshift allows column-level masking for specific users without duplicating data.

- B: Metadata security does not enforce data-level access control.
- C: Row-level security restricts rows, not column-level sensitive values.
- D: Duplicates data, violating the requirement.

Question: 8
A data engineer uses AWS Lake Formation to manage access to data that is stored in an Amazon S3 bucket. The data engineer configures an AWS Glue crawler to discover data at a specific file location in the bucket, s3://examplepath. The crawler execution fails with the following error: “The S3 location: s3://examplepath is not registered.”
The data engineer needs to resolve the error.
Which solution will meet this requirement?

Options
A. Attach an appropriate IAM policy to the IAM role of the AWS Glue crawler to grant the crawler permission to read the S3 location.
B. Register the S3 location in Lake Formation to allow the crawler to access the data.
C. Create a new AWS Glue database. Assign the correct permissions to the database for the crawler.
D. Configure the S3 bucket policy to allow cross-account access.

Answer: B

Explanation:
Lake Formation requires registering S3 locations before Glue crawlers can access them.

- A: IAM role permission is not enough without registration.
- C: Database permissions don’t solve unregistered location error.
- D: Cross-account access is unrelated here.

Question: 9
A company built a data lake and a data warehouse on AWS. The company wants to implement a data catalog to enhance the current data storage solutions. The company wants to have the capability to add business metadata and glossary information to the data catalog for every asset.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A. Use AWS Glue Catalog. Create a user table for the business glossary. Use the AWS Glue API to change table properties to add business metadata. Create a web application to access the metadata.
B. Use an Apache Hive metastore. Create a user table for the business glossary. Use the ALTER TABLE command to change table properties to add business metadata. Create a web application to access the metadata.
C. Use Amazon DataZone. Create the business glossaries. Create metadata forms. Use the Amazon DataZone data portal to access the metadata.
D. Use Amazon OpenSearch Service. Create an index for the business glossary. Create a second index for the business metadata. Use the OpenSearch Service dashboard to access the metadata.

Answer: C

Explanation:
Amazon DataZone is purpose-built for metadata management, business glossaries, and data catalogs with minimal effort.

- A: Glue Data Catalog doesn’t provide glossary management.
- B: Hive metastore requires high overhead.
- D: OpenSearch isn’t designed for metadata governance.

Question: 10
A data engineer is using an AWS Glue ETL job to remove outdated customer records from a table that contains customer account information. The data engineer is using the following SQL command to remove customers that exist in a table named monthly_accounts_update table from the customer accounts table:

MERGE INTO accounts t USING monthly_accounts_update s

ON t.customer = s.customer -

WHEN MATCHED -

THEN DELETE -

What will happen when the data engineer runs the SQL command?

Options
A. All customer records that exist in both the customer accounts table and the monthly_accounts_update table will be deleted from the accounts table.
B. Only customer records that are present in both tables will be retained in the customer accounts table.
C. The monthly_accounts_update table will be deleted.
D. No records will be deleted because the command syntax is not valid in AWS Glue.

Answer: D

Explanation:
AWS Glue does not support MERGE with THEN DELETE syntax, making the command invalid.

- A, B, C: Incorrect because the query won't execute successfully.

Question: 11
A data engineer finished testing an Amazon Redshift stored procedure that processes and inserts data into a table that is not mission critical. The engineer wants to automatically run the stored procedure on a daily basis. Which solution will meet this requirement in the MOST cost-effective way?

Options
A.Create an AWS Lambda function to schedule a cron job to run the stored procedure.
B.Schedule and run the stored procedure by using the Amazon Redshift Data API in an Amazon EC2 Spot Instance.
C.Use query editor v2 to run the stored procedure on a schedule.
D.Schedule an AWS Glue Python shell job to run the stored procedure.

Answer: C

Explanation:
Query editor v2 supports scheduling SQL queries, including stored procedures, to run directly inside Redshift without external orchestration. This avoids extra services like Lambda, Glue, or EC2, making it the most cost-effective option.

- A: Lambda could run the stored procedure via Data API, but it introduces additional components and cost.
- B: Using an EC2 Spot Instance just to run a query is overkill and adds unnecessary infrastructure management.
- D: Glue Python shell jobs work, but Glue is more expensive and complex than using Redshift's built-in scheduler.

Question: 12
A marketing company collects clickstream data. The company sends the clickstream data to Amazon Kinesis Data Firehose and stores the clickstream data in Amazon S3. The company wants to build a series of dashboards that hundreds of users from multiple departments will use. The company will use Amazon QuickSight to develop the dashboards. The company wants a solution that can scale and provide daily updates about clickstream activity. Which combination of steps will meet these requirements MOST cost-effectively? (Choose two.)

Options
A.Use Amazon Redshift to store and query the clickstream data.
B.Use Amazon Athena to query the clickstream data
C.Use Amazon S3 analytics to query the clickstream data.
D.Access the query data through a QuickSight direct SQL query.
E.Access the query data through QuickSight SPICE (Super-fast, Parallel, In-memory Calculation Engine). Configure a daily refresh for the dataset.

Answer: BE

Explanation:
Athena queries S3 data directly without extra infrastructure. Using SPICE improves scalability and performance by caching results and refreshing daily, which supports many concurrent users at low cost.

- A: Redshift works but introduces cluster costs, which are higher than Athena + SPICE.
- C: S3 analytics provides storage usage insights, not general querying for dashboards.
- D: Direct SQL queries from QuickSight to Athena would repeatedly hit Athena, increasing query cost and reducing performance at scale.

Question: 13
A data engineer is building a data orchestration workflow. The data engineer plans to use a hybrid model that includes some on-premises resources and some resources that are in the cloud. The data engineer wants to prioritize portability and open source resources. Which service should the data engineer use in both the on-premises environment and the cloud-based environment?

Options
A.AWS Data Exchange
B.Amazon Simple Workflow Service (Amazon SWF)
C.Amazon Managed Workflows for Apache Airflow (Amazon MWAA)
D.AWS Glue

Answer: C

Explanation:
Apache Airflow is open-source and portable. MWAA provides managed orchestration in AWS, while the same DAGs and orchestration can run on-premises using open-source Airflow, meeting the hybrid and portability need.

- A: Data Exchange is for sharing datasets, not orchestration.
- B: SWF is an older orchestration service but not portable or open-source.
- D: Glue is a managed ETL service and cannot run on-premises.

Question: 14
A gaming company uses a NoSQL database to store customer information. The company is planning to migrate to AWS. The company needs a fully managed AWS solution that will handle high online transaction processing (OLTP) workload, provide single-digit millisecond performance, and provide high availability around the world. Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Amazon Keyspaces (for Apache Cassandra)
B.Amazon DocumentDB (with MongoDB compatibility)
C.Amazon DynamoDB
D.Amazon Timestream

Answer: C

Explanation:
DynamoDB is fully managed, designed for OLTP workloads with millisecond latency, and supports global tables for multi-Region high availability with minimal operational effort.

- A: Keyspaces supports Cassandra API but does not provide global tables or strong OLTP optimizations like DynamoDB.
- B: DocumentDB is for MongoDB workloads, but its scaling and performance are not as strong as DynamoDB for OLTP.
- D: Timestream is purpose-built for time-series data, not general OLTP workloads.

Question: 15
A data engineer creates an AWS Lambda function that an Amazon EventBridge event will invoke. When the data engineer tries to invoke the Lambda function by using an EventBridge event, an AccessDeniedException message appears. How should the data engineer resolve the exception?

Options
A.Ensure that the trust policy of the Lambda function execution role allows EventBridge to assume the execution role.
B.Ensure that both the IAM role that EventBridge uses and the Lambda function's resource-based policy have the necessary permissions.
C.Ensure that the subnet where the Lambda function is deployed is configured to be a private subnet.
D.Ensure that EventBridge schemas are valid and that the event mapping configuration is correct.

Answer: B

Explanation:
EventBridge needs permission to invoke the Lambda function. This requires both EventBridge having IAM permissions and the Lambda resource-based policy granting EventBridge access.

- A: The trust policy is for assuming roles, not for event invocation.
- C: Subnet configuration doesn’t affect permissions.
- D: Schemas and mapping correctness won’t solve AccessDenied errors.

Question: 16
A company uses a data lake that is based on an Amazon S3 bucket. To comply with regulations, the company must apply two layers of server-side encryption to files that are uploaded to the S3 bucket. The company wants to use an AWS Lambda function to apply the necessary encryption. Which solution will meet these requirements?

Options
A.Use both server-side encryption with AWS KMS keys (SSE-KMS) and the Amazon S3 Encryption Client.
B.Use dual-layer server-side encryption with AWS KMS keys (DSSE-KMS).
C.Use server-side encryption with customer-provided keys (SSE-C) before files are uploaded.
D.Use server-side encryption with AWS KMS keys (SSE-KMS).

Answer: B

Explanation:
DSSE-KMS is a feature of S3 that applies two independent layers of KMS-based server-side encryption, meeting the regulatory requirement.

- A: S3 Encryption Client + SSE-KMS would be client + server-side, not dual-layer server-side.
- C: SSE-C only provides single-layer encryption managed by customers.
- D: SSE-KMS applies only one layer of encryption.

Question: 17
A data engineer notices that Amazon Athena queries are held in a queue before the queries run. How can the data engineer prevent the queries from queueing?

Options
A.Increase the query result limit.
B.Configure provisioned capacity for an existing workgroup.
C.Use federated queries.
D.Allow users who run the Athena queries to an existing workgroup.

Answer: B

Explanation:
Provisioned capacity in Athena reserves dedicated query processing resources, preventing queries from queueing.

- A: Query result limits only affect output size, not queuing.
- C: Federated queries expand sources, not capacity.
- D: Changing workgroup doesn’t solve underlying resource contention.

Question: 18
A data engineer needs to debug an AWS Glue job that reads from Amazon S3 and writes to Amazon Redshift. The data engineer enabled the bookmark feature for the AWS Glue job. The data engineer has set the maximum concurrency for the AWS Glue job to 1. The AWS Glue job is successfully writing the output to Amazon Redshift. However, the Amazon S3 files that were loaded during previous runs of the AWS Glue job are being reprocessed by subsequent runs. What is the likely reason the AWS Glue job is reprocessing the files?

Options
A.The AWS Glue job does not have the s3:GetObjectAcl permission that is required for bookmarks to work correctly.
B.The maximum concurrency for the AWS Glue job is set to 1.
C.The data engineer incorrectly specified an older version of AWS Glue for the Glue job.
D.The AWS Glue job does not have a required commit statement.

Answer: D

Explanation:
Glue bookmarks require commit statements to persist processed state. Without a commit, Glue does not mark files as completed, so they get reprocessed.

- A: s3:GetObjectAcl is not required for bookmarks.
- B: Concurrency affects parallelism, not bookmarking.
- C: Glue version mismatch doesn't inherently cause bookmarking failure.

Question: 19
An ecommerce company wants to use AWS to migrate data pipelines from an on-premises environment into the AWS Cloud. The company currently uses a third-party tool in the on-premises environment to orchestrate data ingestion processes. The company wants a migration solution that does not require the company to manage servers. The solution must be able to orchestrate Python and Bash scripts. The solution must not require the company to refactor any code. Which solution will meet these requirements with the LEAST operational overhead?

Options
A.AWS Lambda
B.Amazon Managed Workflows for Apache Airflow (Amazon MVVAA)
C.AWS Step Functions
D.AWS Glue

Answer: B

Explanation:
MWAA provides managed Airflow orchestration, compatible with existing DAGs, including Python and Bash scripts, avoiding refactoring and server management.

- A: Lambda would require rearchitecting ingestion logic into event-driven functions.
- C: Step Functions orchestrate AWS services but not arbitrary Bash/Python scripts directly without wrappers.
- D: Glue is mainly for ETL and would not replicate general orchestration needs without rewriting.

Question: 20
A retail company stores data from a product lifecycle management (PLM) application in an on-premises MySQL database. The PLM application frequently updates the database when transactions occur. The company wants to gather insights from the PLM application in near real time. The company wants to integrate the insights with other business datasets and to analyze the combined dataset by using an Amazon Redshift data warehouse. The company has already established an AWS Direct Connect connection between the on-premises infrastructure and AWS. Which solution will meet these requirements with the LEAST development effort?

Options
A.Run a scheduled AWS Glue extract, transform, and load (ETL) job to get the MySQL database updates by using a Java Database Connectivity (JDBC) connection. Set Amazon Redshift as the destination for the ETL job.
B.Run a full load plus CDC task in AWS Database Migration Service (AWS DMS) to continuously replicate the MySQL database changes. Set Amazon Redshift as the destination for the task.
C.Use the Amazon AppFlow SDK to build a custom connector for the MySQL database to continuously replicate the database changes. Set Amazon Redshift as the destination for the connector.
D.Run scheduled AWS DataSync tasks to synchronize data from the MySQL database. Set Amazon Redshift as the destination for the tasks.

Answer: B

Explanation:
AWS DMS supports continuous replication with Change Data Capture (CDC) from on-premises MySQL into Redshift, providing near real-time insights with minimal development effort.

- A: Glue jobs run in batches, not continuous near real-time CDC.
- C: AppFlow doesn't natively support custom MySQL connectors, and building SDK-based connectors requires more development.
- D: DataSync is for file transfers, not transactional database replication.

Question: 21
A data engineer needs to use Amazon Neptune to develop graph applications.
Which programming languages should the engineer use to develop the graph applications? (Choose two.)

Options
A.Gremlin
B.SQL
C.ANSI SQL
D.SPARQL
E.Spark SQL

Answer: AD

Explanation:
Amazon Neptune supports Gremlin (for property graphs) and SPARQL (for RDF graphs). These are the correct query languages for graph-based applications.

- B: SQL is relational, not supported in Neptune.
- C: ANSI SQL is also relational and not designed for graph queries.
- E: Spark SQL is for Spark processing, not Neptune.

Question: 22
A mobile gaming company wants to capture data from its gaming app. The company wants to make the data available to three internal consumers of the data. The data records are approximately 20 KB in size.
The company wants to achieve optimal throughput from each device that runs the gaming app. Additionally, the company wants to develop an application to process data streams. The stream-processing application must have dedicated throughput for each internal consumer.
Which solution will meet these requirements?

Options
A.Configure the mobile app to call the PutRecords API operation to send data to Amazon Kinesis Data Streams. Use the enhanced fan-out feature with a stream for each internal consumer.
B.Configure the mobile app to call the PutRecordBatch API operation to send data to Amazon Kinesis Data Firehose. Submit an AWS Support case to turn on dedicated throughput for the company's AWS account. Allow each internal consumer to access the stream.
C.Configure the mobile app to use the Amazon Kinesis Producer Library (KPL) to send data to Amazon Kinesis Data Firehose. Use the enhanced fan-out feature with a stream for each internal consumer.
D.Configure the mobile app to call the PutRecords API operation to send data to Amazon Kinesis Data Streams. Host the stream-processing application for each internal consumer on Amazon EC2 instances. Configure auto scaling for the EC2 instances.

Answer: A

Explanation:
Kinesis Data Streams with enhanced fan-out provides dedicated throughput per consumer with low latency. This ensures each consumer can process data independently at scale.

- B: Firehose doesn't support enhanced fan-out and no AWS Support case enables dedicated throughput.
- C: Firehose does not provide dedicated throughput to multiple consumers.

- D: Consumers on EC2 can work, but without enhanced fan-out, they share throughput, reducing efficiency.

Question: 23
A retail company uses an Amazon Redshift data warehouse and an Amazon S3 bucket. The company ingests retail order data into the S3 bucket every day.
The company stores all order data at a single path within the S3 bucket. The data has more than 100 columns. The company ingests the order data from a third-party application that generates more than 30 files in CSV format every day. Each CSV file is between 50 and 70 MB in size.
The company uses Amazon Redshift Spectrum to run queries that select sets of columns. Users aggregate metrics based on daily orders. Recently, users have reported that the performance of the queries has degraded. A data engineer must resolve the performance issues for the queries.
Which combination of steps will meet this requirement with LEAST developmental effort? (Choose two.)

Options
A.Configure the third-party application to create the files in a columnar format.
B.Develop an AWS Glue ETL job to convert the multiple daily CSV files to one file for each day.
C.Partition the order data in the S3 bucket based on order date.
D.Configure the third-party application to create the files in JSON format.
E.Load the JSON data into the Amazon Redshift table in a SUPER type column.

Answer: AC

Explanation:
Columnar formats (like Parquet) improve query performance by scanning only required columns. Partitioning data on order date allows Spectrum to prune unnecessary files, reducing scan time and cost.

- B: Combining into one file per day reduces file count but doesn't improve column pruning.
- D: JSON is less efficient than columnar formats.
- E: SUPER type ingestion increases complexity and cost, not performance.

Question: 24
A company stores customer records in Amazon S3. The company must not delete or modify the customer record data for 7 years after each record is created. The root user also must not have the ability to delete or modify the data.
A data engineer wants to use S3 Object Lock to secure the data.
Which solution will meet these requirements?

Options
A.Enable governance mode on the S3 bucket. Use a default retention period of 7 years.
B.Enable compliance mode on the S3 bucket. Use a default retention period of 7 years.
C.Place a legal hold on individual objects in the S3 bucket. Set the retention period to 7 years.
D.Set the retention period for individual objects in the S3 bucket to 7 years.

Answer: B

Explanation:
Compliance mode enforces write-once-read-many (WORM) and ensures that even the root user cannot delete or overwrite data until the retention period expires.

- A: Governance mode allows privileged users (like root) to bypass restrictions.
- C: Legal hold can be removed anytime, not sufficient.
- D: Object-level retention still doesn't prevent root from deleting unless compliance mode is enabled.

Question: 25
A data engineer needs to create a new empty table in Amazon Athena that has the same schema as an existing table named old_table.
Which SQL statement should the data engineer use to meet this requirement?

Options
A.CREATE TABLE new_table AS SELECT * FROM old_tables;
B.INSERT INTO new_table SELECT * FROM old_table;
C.CREATE TABLE new_table (LIKE old_table);
D.CREATE TABLE new_table AS (SELECT * FROM old_table) WITH NO DATA;

Answer: C

Explanation:
CREATE TABLE new_table (LIKE old_table) copies the schema only, without copying any data.

- A: CREATE TABLE AS SELECT creates a new table with data.
- B: INSERT copies data, not schema.
- D: Athena does not support WITH NO DATA clause.

Question: 26
A data engineer needs to create an Amazon Athena table based on a subset of data from an existing Athena table named cities_world. The cities_world table contains cities that are located around the world. The data engineer must create a new table named cities_us to contain only the cities from cities_world that are located in the US.
Which SQL statement should the data engineer use to meet this requirement?

Options
A.INSERT INTO cities_usa (city,state) SELECT city, state FROM cities_world WHERE country='usa';
B.MOVE city, state FROM cities_world TO cities_usa WHERE country='usa';
C.INSERT INTO cities_usa SELECT city, state FROM cities_world WHERE country='usa';
D.UPDATE cities_usa SET (city, state) = (SELECT city, state FROM cities_world WHERE country='usa');

Answer: C

Explanation:
INSERT INTO cities_usa SELECT ... WHERE country='usa' loads filtered data into the new table.

- A: Syntax is invalid for Athena.
- B: MOVE is not valid SQL in Athena.
- D: UPDATE is invalid because Athena does not support row-level updates.

Question: 27
A company implements a data mesh that has a central governance account. The company needs to catalog all data in the governance account. The governance account uses AWS Lake Formation to centrally share data and grant access permissions.
The company has created a new data product that includes a group of Amazon Redshift Serverless tables. A data engineer needs to share the data product with a marketing team. The marketing team must have access to only a subset of columns. The data engineer needs to share the same data product with a compliance team. The compliance team must have access to a different subset of columns than the marketing team needs access to.
Which combination of steps should the data engineer take to meet these requirements? (Choose two.)

Options
A.Create views of the tables that need to be shared. Include only the required columns.
B.Create an Amazon Redshift data share that includes the tables that need to be shared.
C.Create an Amazon Redshift managed VPC endpoint in the marketing team's account. Grant the marketing team access to the views.
D.Share the Amazon Redshift data share to the Lake Formation catalog in the governance account.
E.Share the Amazon Redshift data share to the Amazon Redshift Serverless workgroup in the marketing team's account.

Answer: BD

Explanation:
Creating a Redshift data share (B) allows datasets to be shared across accounts. Sharing the data share via Lake Formation (D) ensures governance and access control across teams. Views can then be used to control column-level access.

- A: Views help restrict columns but are not sufficient without sharing mechanism.
- C: VPC endpoints are not needed for cross-account governance with Lake Formation.
- E: Redshift Serverless workgroup sharing is not the governance model required here.

Question: 28
A company has a data lake in Amazon S3. The company uses AWS Glue to catalog data and AWS Glue Studio to implement data extract, transform, and load (ETL) pipelines.
The company needs to ensure that data quality issues are checked every time the pipelines run. A data engineer must enhance the existing pipelines to evaluate data quality rules based on predefined thresholds.
Which solution will meet these requirements with the LEAST implementation effort?

Options
A.Add a new transform that is defined by a SQL query to each Glue ETL job. Use the SQL query to implement a ruleset that includes the data quality rules that need to be evaluated.
B.Add a new Evaluate Data Quality transform to each Glue ETL job. Use Data Quality Definition Language (DQDL) to implement a ruleset that includes the data quality rules that need to be evaluated.
C.Add a new custom transform to each Glue ETL job. Use the PyDeequ library to implement a ruleset that includes the data quality rules that need to be evaluated.
D.Add a new custom transform to each Glue ETL job. Use the Great Expectations library to implement a ruleset that includes the data quality rules that need to be evaluated.

Answer: B

Explanation:
AWS Glue provides a built-in Evaluate Data Quality transform with DQDL, allowing direct integration with Glue jobs and minimal effort.

- A: SQL-based checks are manual and lack Glue-native integration.
- C: PyDeequ requires coding and maintenance.
- D: Great Expectations is external and increases development effort.

Question: 29
A company has an application that uses a microservice architecture. The company hosts the application on an Amazon Elastic Kubernetes Services (Amazon EKS) cluster.
The company wants to set up a robust monitoring system for the application. The company needs to analyze the logs from the EKS cluster and the application. The company needs to correlate the cluster's logs with the application's traces to identify points of failure in the whole application request flow.
Which combination of steps will meet these requirements with the LEAST development effort? (Choose two.)

Options
A.Use FluentBit to collect logs. Use OpenTelemetry to collect traces.
B.Use Amazon CloudWatch to collect logs. Use Amazon Kinesis to collect traces.
C.Use Amazon CloudWatch to collect logs. Use Amazon Managed Streaming for Apache Kafka (Amazon MSK) to collect traces.
D.Use Amazon OpenSearch to correlate the logs and traces.
E.Use AWS Glue to correlate the logs and traces.

Answer: AD

Explanation:
FluentBit (A) integrates with EKS for log collection. OpenTelemetry (A) is the industry standard for distributed tracing. OpenSearch (D) can ingest both logs and traces to provide correlation dashboards with minimal effort.

- B: Kinesis is not meant for tracing.
- C: MSK is overkill for traces.
- E: Glue is ETL, not log/trace correlation.

Question: 30
A company has a gaming application that stores data in Amazon DynamoDB tables. A data engineer needs to ingest the game data into an Amazon OpenSearch Service cluster. Data updates must occur in near real time.
Which solution will meet these requirements?

Options
A.Use AWS Step Functions to periodically export data from the Amazon DynamoDB tables to an Amazon S3 bucket. Use an AWS Lambda function to load the data into Amazon OpenSearch Service.
B.Configure an AWS Glue job to have a source of Amazon DynamoDB and a destination of Amazon OpenSearch Service to transfer data in near real time.
C.Use Amazon DynamoDB Streams to capture table changes. Use an AWS Lambda function to process and update the data in Amazon OpenSearch Service.
D.Use a custom OpenSearch plugin to sync data from the Amazon DynamoDB tables.

Answer: C

Explanation:
DynamoDB Streams capture changes in near real time. Lambda can process these streams and update OpenSearch immediately.

- A: Step Functions batch processing introduces latency.
- B: Glue jobs are not real-time.
- D: Custom plugins add unnecessary development overhead.

Question: 31
A company receives marketing campaign data from a vendor. The company ingests the data into an Amazon S3 bucket every 40 to 60 minutes. The data is in CSV format. File sizes are between 100 KB and 300 KB.
A data engineer needs to set-up an extract, transform, and load (ETL) pipeline to upload the content of each file to Amazon Redshift.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A. Create an AWS Lambda function that connects to Amazon Redshift and runs a COPY command. Use Amazon EventBridge to invoke the Lambda function based on an Amazon S3 upload trigger.
B. Create an Amazon Data Firehose stream. Configure the stream to use an AWS Lambda function as a source to pull data from the S3 bucket. Set Amazon Redshift as the destination.
C. Use Amazon Redshift Spectrum to query the S3 bucket. Configure an AWS Glue Crawler for the S3 bucket to update metadata in an AWS Glue Data Catalog.
D. Creates an AWS Database Migration Service (AWS DMS) task. Specify an appropriate data schema to migrate. Specify the appropriate type of migration to use.

Answer: A

Explanation:
Lambda + COPY triggered by S3 events is lightweight, serverless, and fits small, frequent files.

- B: Firehose does not natively pull from S3.
- C: Redshift Spectrum queries external S3 data, not ingest.
- D: DMS is not designed for S3 → Redshift ingestion.

Question: 32
A company wants to build a dimension table in an Amazon S3 bucket. The bucket contains historical data that includes 10 million records. The historical data is 1 TB in size.
A data engineer needs a solution to update changes for up to 10,000 records in the base table every day.
Which solution will meet this requirement with the LOWEST runtime?

Options
A. Develop an Apache Spark job in Amazon EMR to read the historical data and the new changes into two Spark DataFrames. Use the Spark update method to update the base table.
B. Develop an AWS Glue Python job to read the historical data and new changes into two Pandas DataFrames. Use the Pandas update method to update the base table.
C. Develop an AWS Glue Apache Spark job to read the historical data and new changes into two Spark DataFrames. Use the Spark update method to update the base table.
D. Develop an Amazon EMR job to read new changes into Apache Spark DataFrames. Use the Apache Hudi framework to create the base table in Amazon S3. Use the Spark update method to update the base table.

Answer: D

Explanation:
Apache Hudi supports incremental updates and upserts on S3, which is efficient for daily updates.

- A, C: Would require full table rewrite each time.
- B: Pandas cannot handle 1 TB efficiently.

Question: 33
A data engineer develops an AWS Glue Apache Spark ETL job to perform transformations on a dataset. When the data engineer runs the job, the job returns an error that reads, “No space left on device.” The data engineer needs to identify the source of the error and provide a solution.
Which combinations of steps will meet this requirement MOST cost-effectively? (Choose two.)

Options
A. Scale out the workers vertically to address data skewness.
B. Use the Spark UI and AWS Glue metrics to monitor data skew in the Spark executors.
C. Scale out the number of workers horizontally to address data skewness.
D. Enable the --write-shuffie-files-to-s3 job parameter. Use the salting technique.
E. Use error logs in Amazon CloudWatch to monitor data skew.

Answer: BD

Explanation:

- B: Spark UI and Glue metrics help detect data skew, root cause of disk space issue.
- D: Writing shuffle files to S3 and salting prevents local disk overflow.
- A, C: Scaling is costly and doesn’t fix skew itself.
- E: CloudWatch logs show errors but don’t help identify skew details efficiently.

Question: 34
A company builds a new data pipeline to process data for business intelligence reports. Users have noticed that data is missing from the reports.
A data engineer needs to add a data quality check for columns that contain null values and for referential integrity at a stage before the data is added to storage.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A. Use Amazon SageMaker Data Wrangler to create a Data Quality and Insights report.
B. Use AWS Glue ETL jobs to perform a data quality evaluation transform on the data. Use an IsComplete rule on the requested columns. Use a ReferentialItegrity rule for each join.
C. Use AWS Glue ETL jobs to perform a SQL transform on the data to determine whether requested column contain null values. Use a second SQL transform to check referential integrity.
D. Use Amazon SageMaker Data Wrangler and a custom Python transform to create custom rules to check for null values and referential integrity.

Answer: B

Explanation:
Glue Data Quality transform supports built-in rules like IsComplete and ReferentialIntegrity with minimal setup.

- A, D: Data Wrangler not designed for automated pipelines at scale.
- C: SQL transforms add more development and operational overhead.

Question: 35
A company is setting up a data pipeline in AWS. The pipeline extracts client data from Amazon S3 buckets, performs quality checks, and transforms the data. The pipeline stores the processed data in a relational database. The company will use the processed data for future queries.
Which solution will meet these requirements MOST cost-effectively?

Options
A. Use AWS Glue ETL to extract the data from the S3 buckets and perform the transformations. Use AWS Glue Data Quality to enforce suggested quality rules. Load the data and the quality check results into an Amazon RDS for MySQL instance.
B. Use AWS Glue Studio to extract the data from the S3 buckets. Use AWS Glue DataBrew to perform the transformations and quality checks. Load the processed data into an Amazon RDS for MySQL instance. Load the quality check results into a new S3 bucket.
C. Use AWS Glue ETL to extract the data from the S3 buckets and perform the transformations. Use AWS Glue DataBrew to perform quality checks. Load the processed data and the quality check results into a new S3 bucket.
D. Use AWS Glue Studio to extract the data from the S3 buckets. Use AWS Glue DataBrew to perform the transformations and quality checks. Load the processed data and quality check results into an Amazon RDS for MySQL instance.

Answer: A

Explanation:
Glue ETL + Glue Data Quality integrates natively, and RDS is the relational storage target. This minimizes moving parts.

- B, D: DataBrew is more for ad-hoc data prep, not production ETL.
- C: Puts results into S3 instead of required relational DB.

Question: 36
A company uses Amazon Redshift as a data warehouse solution. One of the datasets that the company stores in Amazon Redshift contains data for a vendor.
Recently, the vendor asked the company to transfer the vendor's data into the vendor's Amazon S3 bucket once each week.
Which solution will meet this requirement?

Options
A. Create an AWS Lambda function to connect to the Redshift data warehouse. Configure the Lambda function to use the Redshift COPY command to copy the required data to the vendor's S3 bucket on a schedule.
B. Create an AWS Glue job to connect to the Redshift data warehouse. Configure the AWS Glue job to use the Redshift UNLOAD command to load the required data to the vendor's S3 bucket on a schedule.
C. Use the Amazon Redshift data sharing feature. Set the vendor's S3 bucket as the destination. Configure the source to be as a custom SQL query that selects the required data.
D. Configure Amazon Redshift Spectrum to use the vendor's S3 bucket a destination, Enable data querying in both directions.

Answer: B

Explanation:
The Redshift UNLOAD command writes data from Redshift to S3 efficiently, and Glue can orchestrate scheduling.

- A: COPY is for loading into Redshift, not unloading.
- C, D: Data sharing and Spectrum do not export data into vendor's S3 bucket.

Question: 37
A company uses an Amazon Redshift cluster as a data warehouse that is shared across two departments. To comply with a security policy, each department must have unique access permissions.
Department A must have access to tables and views for Department A. Department B must have access to tables and views for Department B.
The company often runs SQL queries that use objects from both departments in one query.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A. Group tables and views for each department into dedicated schemas. Manage permissions at the schema level.
B. Group tables and views for each department into dedicated databases. Manage permissions at the database level.
C. Update the names of the tables and views to follow a naming convention that contains the department names. Manage permissions based on the new naming convention.
D. Create an IAM user group for each department. Use identity-based IAM policies to grant table and view permissions based on the IAM user group.

Answer: A

Explanation:
Schemas provide a simple way to logically group and control permissions while still allowing cross-schema queries.

- B: Splitting into databases complicates cross-department queries.
- C: Naming conventions don't enforce permissions.
- D: Redshift access is managed internally, not via IAM groups directly.

Question: 38
A company wants to ingest streaming data into an Amazon Redshift data warehouse from an Amazon Managed Streaming for Apache Kafka (Amazon MSK) cluster. A data engineer needs to develop a solution that provides low data access time and that optimizes storage costs.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A. Create an external schema that maps to the MSK cluster. Create a materialized view that references the external schema to consume the streaming data from the MSK topic.
B. Develop an AWS Glue streaming extract, transform, and load (ETL) job to process the incoming data from Amazon MSK. Load the data into Amazon S3. Use Amazon Redshift Spectrum to read the data from Amazon S3.
C. Create an external schema that maps to the streaming data source. Create a new Amazon Redshift table that references the external schema.
D. Create an Amazon S3 bucket. Ingest the data from Amazon MSK. Create an event-driven AWS Lambda function to load the data from the S3 bucket to a new Amazon Redshift table.

Answer: B

Explanation:
Glue streaming job can process MSK data and write Parquet/optimized format to S3. Redshift Spectrum then queries efficiently with low storage cost.

- A, C: Redshift doesn't directly map external schemas to MSK.
- D: Adds latency and complexity with Lambda.

Question: 39
A sales company uses AWS Glue ETL to collect, process, and ingest data into an Amazon S3 bucket. The AWS Glue pipeline creates a new file in the S3 bucket every hour. File sizes vary from 200 KB to 300 KB. The company wants to build a sales prediction model by using data from the previous 5 years. The historic data includes 44,000 files.
The company builds a second AWS Glue ETL pipeline by using the smallest worker type. The second pipeline retrieves the historic files from the S3 bucket and processes the files for downstream analysis. The company notices significant performance issues with the second ETL pipeline.
The company needs to improve the performance of the second pipeline.
Which solution will meet this requirement MOST cost-effectively?

Options
A. Use a larger worker type.
B. Increase the number of workers in the AWS Glue ETL jobs.
C. Use the AWS Glue DynamicFrame grouping option.
D. Enable AWS Glue auto scaling.

Answer: C

Explanation:
DynamicFrame grouping combines many small files into fewer larger files, improving Glue performance without scaling costs.

- A, B, D: These improve performance but at higher cost compared to grouping optimization.

Question: 40
A company wants to combine data from multiple software as a service (SaaS) applications for analysis.
A data engineering team needs to use Amazon QuickSight to perform the analysis and build dashboards. A data engineer needs to extract the data from the SaaS applications and make the data available for QuickSight queries.
Which solution will meet these requirements in the MOST operationally efficient way?

Options
A. Create AWS Lambda functions that call the required APIs to extract the data from the applications. Store the data in an Amazon S3 bucket. Use AWS Glue to catalog the data in the S3 bucket. Create a data source and a dataset in QuickSight.
B. Use AWS Lambda functions as Amazon Athena data source connectors to run federated queries against the SaaS applications. Create an Athena data source and a dataset in QuickSight.
C. Use Amazon AppFlow to create a flow for each SaaS application. Set an Amazon S3 bucket as the destination. Schedule the flows to extract the data to the bucket. Use AWS Glue to catalog the data in the S3 bucket. Create a data source and a dataset in QuickSight.
D. Export data the from the SaaS applications as Microsoft Excel files. Create a data source and a dataset in QuickSight by uploading the Excel files.

Answer: C

Explanation:
Amazon AppFlow provides a managed, scalable way to pull data from SaaS apps into S3, with minimal development effort.

- A, B: Require custom Lambda functions, increasing maintenance.
- D: Manual and not scalable.

Question: 41
company analyzes data in a data lake every quarter to perform inventory assessments. A data engineer uses AWS Glue DataBrew to detect any personally identifiable formation (PII) about customers within the data. The company's privacy policy considers some custom categories of information to be PII. However, the categories are not included in standard DataBrew data quality rules.
The data engineer needs to modify the current process to scan for the custom PII categories across multiple datasets within the data lake.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Manually review the data for custom PII categories.
B.Implement custom data quality rules in DataBrew. Apply the custom rules across datasets.
C.Develop custom Python scripts to detect the custom PII categories. Call the scripts from DataBrew.
D.Implement regex patterns to extract PII information from fields during extract transform, and load (ETL) operations into the data lake.

Answer: B

Explanation:
Custom rules in DataBrew are the most efficient and low-maintenance way to extend PII detection with company-specific categories. They can be applied at scale across datasets.

- A: Manual review is labor-intensive and error-prone.
- C: Custom scripts increase development/maintenance.
- D: Regex in ETL adds complexity and does not leverage DataBrew's managed rules.

Question: 42
A company receives a data file from a partner each day in an Amazon S3 bucket. The company uses a daily AWS Glue extract, transform, and load (ETL) pipeline to clean and transform each data file. The output of the ETL pipeline is written to a CSV file named Daily.csv in a second S3 bucket.
Occasionally, the daily data file is empty or is missing values for required fields. When the file is missing data, the company can use the previous day’s CSV file.
A data engineer needs to ensure that the previous day's data file is overwritten only if the new daily file is complete and valid.
Which solution will meet these requirements with the LEAST effort?

Options
A.Invoke an AWS Lambda function to check the file for missing data and to fill in missing values in required fields.
B.Configure the AWS Glue ETL pipeline to use AWS Glue Data Quality rules. Develop rules in Data Quality Definition Language (DQDL) to check for missing values in required fields and empty files.
C.Use AWS Glue Studio to change the code in the ETL pipeline to fill in any missing values in the required fields with the most common values for each field.
D.Run a SQL query in Amazon Athena to read the CSV file and drop missing rows. Copy the corrected CSV file to the second S3 bucket.

Answer: B

Explanation:
AWS Glue Data Quality rules (with DQDL) can validate files for completeness and missing values automatically, preventing bad data from overwriting the existing file.

- A: Lambda validation requires extra scripting/management.
- C: Filling missing values is not acceptable for this scenario.
- D: Athena queries add unnecessary steps and still overwrite with incomplete data.

Question: 43
A marketing company uses Amazon S3 to store marketing data. The company uses versioning in some buckets. The company runs several jobs to read and load data into the buckets.
To help cost-optimize its storage, the company wants to gather information about incomplete multipart uploads and outdated versions that are present in the S3 buckets.
Which solution will meet these requirements with the LEAST operational effort?

Options
A.Use AWS CLI to gather the information.
B.Use Amazon S3 Inventory configurations reports to gather the information.
C.Use the Amazon S3 Storage Lens dashboard to gather the information.
D.Use AWS usage reports for Amazon S3 to gather the information.

Answer: C

Explanation:
Amazon S3 Storage Lens provides visibility into incomplete multipart uploads and object versions across accounts and buckets in a dashboard, with minimal configuration.

- A: CLI requires manual scripts and repeated execution.
- B: Inventory gives metadata but not metrics about incomplete uploads.
- D: Usage reports provide billing data, not operational metrics on versions/uploads.

Question: 44
A gaming company uses Amazon Kinesis Data Streams to collect clickstream data. The company uses Amazon Data Firehose delivery streams to store the data in JSON format in Amazon S3. Data scientists at the company use Amazon Athena to query the most recent data to obtain business insights.
The company wants to reduce Athena costs but does not want to recreate the data pipeline. Which solution will meet these requirements with the LEAST management effort?

Options
A.Change the Firehose output format to Apache Parquet. Provide a custom S3 object YYYYMMDD prefix expression and specify a large buffer size. For the existing data, create an AWS Glue extract, transform, and load (ETL) job. Configure the ETL job to combine small JSON files, convert the JSON files to large Parquet files, and add the YYYYMMDD prefix. Use the ALTER TABLE ADD PARTITION statement to reflect the partition on the existing Athena table.
B.Create an Apache Spark job that combines JSON files and converts the JSON files to Apache Parquet files. Launch an Amazon EMR ephemeral cluster every day to run the Spark job to create new Parquet files in a different S3 location. Use the ALTER TABLE SET LOCATION statement to reflect the new S3 location on the existing Athena table.
C.Create a Kinesis data stream as a delivery destination for Firehose. Use Amazon Managed Service for Apache Flink (previously known as Amazon Kinesis Data Analytics) to run Apache Flink on the Kinesis data stream. Use Flink to aggregate the data and save the data to Amazon S3 in Apache Parquet format with a custom S3 object YYYYMMDD prefix. Use the ALTER TABLE ADD PARTITION statement to reflect the partition on the existing Athena table.
D.Integrate an AWS Lambda function with Firehose to convert source records to Apache Parquet and write them to Amazon S3. In parallel, run an AWS Glue extract, transform, and load (ETL) job to combine the JSON files and convert the JSON files to large Parquet files. Create a custom S3 object YYYYMMDD prefix. Use the ALTER TABLE ADD PARTITION statement to reflect the partition on the existing Athena table.

Answer: A

Explanation:
Firehose natively supports Parquet conversion, reducing Athena scan costs. For historical JSON, a one-time Glue ETL can backfill into Parquet with partitioning. This minimizes operational work.

- B: EMR jobs require cluster management.
- C: Flink requires new stream processing pipelines.
- D: Lambda conversions add complexity and parallel Glue job increases ops.

Question: 45
A company needs a solution to manage costs for an existing Amazon DynamoDB table. The company also needs to control the size of the table. The solution must not disrupt any ongoing read or write operations. The company wants to use a solution that automatically deletes data from the table after 1 month.
Which solution will meet these requirements with the LEAST ongoing maintenance?

Options
A.Use the DynamoDB TTL feature to automatically expire data based on timestamps.
B.Configure a scheduled Amazon EventBridge rule to invoke an AWS Lambda function to check for data that is older than 1 month. Configure the Lambda function to delete old data.
C.Configure a stream on the DynamoDB table to invoke an AWS Lambda function. Configure the Lambda function to delete data in the table that is older than 1 month.
D.Use an AWS Lambda function to periodically scan the DynamoDB table for data that is older than 1 month. Configure the Lambda function to delete old data.

Answer: A

Explanation:
DynamoDB TTL is a fully managed feature that automatically deletes expired items without affecting ongoing reads/writes and requires no extra maintenance.

- B: EventBridge + Lambda adds custom code and scheduling overhead.
- C: Streams + Lambda adds unnecessary processing for TTL-like behavior.
- D: Lambda scans are costly and inefficient for large tables.

Question: 46
A company uses Amazon S3 to store data and Amazon QuickSight to create visualizations,
The company has an S3 bucket in an AWS account named Hub-Account. The S3 bucket is encrypted by an AWS Key Management Service (AWS KMS) key. The company's QuickSight instance is in a separate account named BI-Account.
The company updates the S3 bucket policy to grant access to the QuickSight service role. The company wants to enable cross-account access to allow QuickSight to interact with the S3 bucket.
Which combination of steps will meet this requirement? (Choose two.)

Options
A.Use the existing AWS KMS key to encrypt connections from QuickSight to the S3 bucket.
B.Add the S3 bucket as a resource that the QuickSight service role can access.
C.Use AWS Resource Access Manager (AWS RAM) to share the S3 bucket with the BI-Account account.
D.Add an IAM policy to the QuickSight service role to give QuickSight access to the KMS key that encrypts the S3 bucket.
E.Add the KMS key as a resource that the QuickSight service role can access.

Answer: BE

Explanation:
QuickSight must have both S3 bucket access (B) and permissions to use the KMS key (E) to read encrypted objects across accounts.

- A: KMS keys are not used to "encrypt connections"; they encrypt objects.
- C: S3 sharing is not done through RAM.
- D: Adding an IAM policy alone doesn’t grant access unless the KMS key policy allows it.

Question: 47
A car sales company maintains data about cars that are listed for sale in an area. The company receives data about new car listings from vendors who upload the data daily as compressed files into Amazon S3. The compressed files are up to 5 KB in size. The company wants to see the most up-to-date listings as soon as the data is uploaded to Amazon S3.
A data engineer must automate and orchestrate the data processing workflow of the listings to feed a dashboard. The data engineer must also provide the ability to perform one-time queries and analytical reporting. The query solution must be scalable.
Which solution will meet these requirements MOST cost-effectively?

Options
A.Use an Amazon EMR cluster to process incoming data. Use AWS Step Functions to orchestrate workflows. Use Apache Hive for one-time queries and analytical reporting. Use Amazon OpenSearch Service to bulk ingest the data into compute optimized instances. Use OpenSearch Dashboards in OpenSearch Service for the dashboard.
B.Use a provisioned Amazon EMR cluster to process incoming data. Use AWS Step Functions to orchestrate workflows. Use Amazon Athena for one-time queries and analytical reporting. Use Amazon QuickSight for the dashboard.
C.Use AWS Glue to process incoming data. Use AWS Step Functions to orchestrate workflows. Use Amazon Redshift Spectrum for one-time queries and analytical reporting. Use OpenSearch Dashboards in Amazon OpenSearch Service for the dashboard.
D.Use AWS Glue to process incoming data. Use AWS Lambda and S3 Event Notifications to orchestrate workflows. Use Amazon Athena for one-time queries and analytical reporting. Use Amazon QuickSight for the dashboard.

Answer: D

Explanation:
Glue handles processing, Lambda with S3 events orchestrates near real-time updates, Athena provides serverless ad-hoc querying, and QuickSight gives dashboards — all cost-effective and low maintenance.

- A: EMR + OpenSearch adds unnecessary cost and complexity.
- B: Provisioned EMR adds cost for small files.
- C: Redshift Spectrum is overkill for small-scale queries and not cost-optimal.

Question: 48
A company has AWS resources in multiple AWS Regions. The company has an Amazon EFS file system in each Region where the company operates. The company's data science team operates within only a single Region. The data that the data science team works with must remain within the team's Region.
A data engineer needs to create a single dataset by processing files that are in each of the company's Regional EFS file systems. The data engineer wants to use an AWS Step Functions state machine to orchestrate AWS Lambda functions to process the data.
Which solution will meet these requirements with the LEAST effort?

Options
A.Peer the VPCs that host the EFS file systems in each Region with the VPC that is in the data science team's Region. Enable EFS file locking. Configure the Lambda functions in the data science team's Region to mount each of the Region specific file systems. Use the Lambda functions to process the data.
B.Configure each of the Regional EFS file systems to replicate data to the data science team's Region. In the data science team's Region, configure the Lambda functions to mount the replica file systems. Use the Lambda functions to process the data.
C.Deploy the Lambda functions to each Region. Mount the Regional EFS file systems to the Lambda functions. Use the Lambda functions to process the data. Store the output in an Amazon S3 bucket in the data science team's Region.
D.Use AWS DataSync to transfer files from each of the Regional EFS files systems to the file system that is in the data science team's Region. Configure the Lambda functions in the data science team's Region to mount the file system that is in the same Region. Use the Lambda functions to process the data.

Answer: A

Explanation:
VPC peering with EFS mounts allows the central Lambda functions to access all Regional EFS file systems without replication or data transfer, keeping data in-region.

- B: Cross-region replication adds cost and effort.
- C: Deploying Lambda in multiple Regions complicates orchestration.
- D: DataSync transfers duplicate data, adding extra ops and cost.

Question: 49
A company hosts its applications on Amazon EC2 instances. The company must use SSL/TLS connections that encrypt data in transit to communicate securely with AWS infrastructure that is managed by a customer.
A data engineer needs to implement a solution to simplify the generation, distribution, and rotation of digital certificates. The solution must automatically renew and deploy SSL/TLS certificates.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Store self-managed certificates on the EC2 instances.
B.Use AWS Certificate Manager (ACM).
C.Implement custom automation scripts in AWS Secrets Manager.
D.Use Amazon Elastic Container Service (Amazon ECS) Service Connect.

Answer: B

Explanation:
AWS Certificate Manager (ACM) provides managed SSL/TLS certificates with automatic renewal and easy integration into AWS services, minimizing operational overhead.

- A: Self-managed certs require manual rotation.
- C: Secrets Manager isn't designed for cert lifecycle management.
- D: ECS Service Connect does not manage SSL/TLS certificates.

Question: 50
A company saves customer data to an Amazon S3 bucket. The company uses server-side encryption with AWS KMS keys (SSE-KMS) to encrypt the bucket. The dataset includes personally identifiable information (PII) such as social security numbers and account details.
Data that is tagged as PII must be masked before the company uses customer data for analysis.
Some users must have secure access to the PII data during the pre-processing phase. The company needs a low-maintenance solution to mask and secure the PII data throughout the entire engineering pipeline.
Which combination of solutions will meet these requirements? (Choose two.)

Options
A.Use AWS Glue DataBrew to perform extract, transform, and load (ETL) tasks that mask the PII data before analysis.
B.Use Amazon GuardDuty to monitor access patterns for the PII data that is used in the engineering pipeline.
C.Configure an Amazon Macie discovery job for the S3 bucket.
D.Use AWS Identity and Access Management (IAM) to manage permissions and to control access to the PII data.
E.Write custom scripts in an application to mask the PII data and to control access.

Answer: AD

Explanation:
Glue DataBrew (A) can mask sensitive fields efficiently during preprocessing. IAM (D) ensures only authorized users can access sensitive PII data. This combination provides masking + secure access with minimal ops.

- B: GuardDuty is for threat detection, not masking or access control.
- C: Macie detects sensitive data but doesn't mask it.
- E: Custom scripts increase dev/maintenance effort.

Question: 51
A data engineer needs Amazon Athena queries to finish faster. The data engineer notices that all the files the Athena queries use are currently stored in uncompressed .csv format. The data engineer also notices that users perform most queries by selecting a specific column. Which solution will MOST speed up the Athena query performance?

Options
A.Change the data format from .csv to JSON format. Apply Snappy compression.
B.Compress the .csv files by using Snappy compression.
C.Change the data format from .csv to Apache Parquet. Apply Snappy compression.
D.Compress the .csv files by using gzip compression.

Answer: C

Explanation:
Columnar formats like Parquet are optimized for analytics queries that scan specific columns. Combined with compression such as Snappy, Athena can read less data, significantly improving query speed.

- A is incorrect because JSON is row-based, not columnar, and is less efficient than Parquet.
- B is incorrect because compressing CSV reduces storage size but does not improve columnar read efficiency.
- D is incorrect because gzip compression on CSV still leaves it as row-based, so column pruning is not possible.

Question: 52
A manufacturing company collects sensor data from its factory floor to monitor and enhance operational efficiency. The company uses Amazon Kinesis Data Streams to publish the data that the sensors collect to a data stream. Then Amazon Kinesis Data Firehose writes the data to an Amazon S3 bucket. The company needs to display a real-time view of operational efficiency on a large screen in the manufacturing facility. Which solution will meet these requirements with the LOWEST latency?

Options
A.Use Amazon Managed Service for Apache Flink (previously known as Amazon Kinesis Data Analytics) to process the sensor data. Use a connector for Apache Flink to write data to an Amazon Timestream database. Use the Timestream database as a source to create a Grafana dashboard.
B.Configure the S3 bucket to send a notification to an AWS Lambda function when any new object is created. Use the Lambda function to publish the data to Amazon Aurora. Use Aurora as a source to create an Amazon QuickSight dashboard.
C.Use Amazon Managed Service for Apache Flink (previously known as Amazon Kinesis Data Analytics) to process the sensor data. Create a new Data Firehose delivery stream to publish data directly to an Amazon Timestream database. Use the Timestream database as a source to create an Amazon QuickSight dashboard.
D.Use AWS Glue bookmarks to read sensor data from the S3 bucket in real time. Publish the data to an Amazon Timestream database. Use the Timestream database as a source to create a Grafana dashboard.

Answer: A

Explanation:
Managed Service for Apache Flink provides near real-time stream processing. Writing directly to Amazon Timestream and visualizing with Grafana ensures very low latency.

- B is incorrect because relying on S3 and Lambda introduces delays and is not real-time.
- C is incorrect because Firehose buffers data before delivery, which adds latency.
- D is incorrect because Glue is batch-oriented, not real-time.

Question: 53
A company stores daily records of the financial performance of investment portfolios in .csv format in an Amazon S3 bucket. A data engineer uses AWS Glue crawlers to crawl the S3 data. The data engineer must make the S3 data accessible daily in the AWS Glue Data Catalog. Which solution will meet these requirements?

Options
A.Create an IAM role that includes the AmazonS3FullAccess policy. Associate the role with the crawler. Specify the S3 bucket path of the source data as the crawler's data store. Create a daily schedule to run the crawler. Configure the output destination to a new path in the existing S3 bucket.
B.Create an IAM role that includes the AWSGlueServiceRole policy. Associate the role with the crawler. Specify the S3 bucket path of the source data as the crawler's data store. Create a daily schedule to run the crawler. Specify a database name for the output.
C.Create an IAM role that includes the AmazonS3FullAccess policy. Associate the role with the crawler. Specify the S3 bucket path of the source data as the crawler's data store. Allocate data processing units (DPUs) to run the crawler every day. Specify a database name for the output.
D.Create an IAM role that includes the AWSGlueServiceRole policy. Associate the role with the crawler. Specify the S3 bucket path of the source data as the crawler's data store. Allocate data processing units (DPUs) to run the crawler every day. Configure the output destination to a new path in the existing S3 bucket.

Answer: B

Explanation:
Glue crawlers need the AWSGlueServiceRole IAM policy to run properly. By scheduling the crawler daily and specifying a database name, the data is kept updated in the Data Catalog automatically.

- A is incorrect because AmazonS3FullAccess alone is insufficient; the crawler also needs Glue permissions.
- C is incorrect because allocating DPUs is not necessary for crawlers, which run as serverless.
- D is incorrect because output destinations are not relevant for crawlers; they populate the Data Catalog, not write data.

Question: 54
A company loads transaction data for each day into Amazon Redshift tables at the end of each day. The company wants to have the ability to track which tables have been loaded and which tables still need to be loaded. A data engineer wants to store the load statuses of Redshift tables in an Amazon DynamoDB table. The data engineer creates an AWS Lambda function to publish the details of the load statuses to DynamoDB. How should the data engineer invoke the Lambda function to write load statuses to the DynamoDB table?

Options
A.Use a second Lambda function to invoke the first Lambda function based on Amazon CloudWatch events.
B.Use the Amazon Redshift Data API to publish an event to Amazon EventBridge. Configure an EventBridge rule to invoke the Lambda function.
C.Use the Amazon Redshift Data API to publish a message to an Amazon Simple Queue Service (Amazon SQS) queue. Configure the SQS queue to invoke the Lambda function.
D.Use a second Lambda function to invoke the first Lambda function based on AWS CloudTrail events.

Answer: B

Explanation:
The Redshift Data API can publish events to EventBridge. EventBridge rules can then invoke Lambda functions to update DynamoDB, enabling automated and event-driven status updates.

- A is incorrect because chaining Lambdas adds complexity without integrating with Redshift.
- C is incorrect because Redshift Data API does not directly publish to SQS.
- D is incorrect because CloudTrail tracks API calls, not table load statuses in this case.

Question: 55
A data engineer needs to securely transfer 5 TB of data from an on-premises data center to an Amazon S3 bucket. Approximately 5% of the data changes every day. Updates to the data need to be regularly proliferated to the S3 bucket. The data includes files that are in multiple formats. The data engineer needs to automate the transfer process and must schedule the process to run periodically. Which AWS service should the data engineer use to transfer the data in the MOST operationally efficient way?

Options
A.AWS DataSync
B.AWS Glue
C.AWS Direct Connect
D.Amazon S3 Transfer Acceleration

Answer: A

Explanation:
AWS DataSync is designed for efficient, automated, and scheduled transfers of large amounts of on-premises data to AWS storage services like S3, with built-in incremental syncs.

- B is incorrect because Glue is for ETL, not file transfer.
- C is incorrect because Direct Connect provides a dedicated link, but does not automate or schedule transfers.
- D is incorrect because S3 Transfer Acceleration speeds uploads but does not handle scheduling or incremental sync.

Question: 56
A company uses an on-premises Microsoft SQL Server database to store financial transaction data. The company migrates the transaction data from the on-premises database to AWS at the end of each month. The company has noticed that the cost to migrate data from the on-premises database to an Amazon RDS for SQL Server database has increased recently. The company requires a cost-effective solution to migrate the data to AWS. The solution must cause minimal downtime for the applications that access the database. Which AWS service should the company use to meet these requirements?

Options
A.AWS Lambda
B.AWS Database Migration Service (AWS DMS)
C.AWS Direct Connect
D.AWS DataSync

Answer: B

Explanation:
AWS DMS is designed for database migrations with minimal downtime. It can continuously replicate changes and is cost-effective compared to manual or repeated full transfers.

- A is incorrect because Lambda is not a migration tool.
- C is incorrect because Direct Connect is a network connection, not a migration service.
- D is incorrect because DataSync is for files, not structured database migrations.

Question: 57
A data engineer is building a data pipeline on AWS by using AWS Glue extract, transform, and load (ETL) jobs. The data engineer needs to process data from Amazon RDS and MongoDB, perform transformations, and load the transformed data into Amazon Redshift for analytics. The data updates must occur every hour. Which combination of tasks will meet these requirements with the LEAST operational overhead? (Choose two.)

Options
A.Configure AWS Glue triggers to run the ETL jobs every hour.
B.Use AWS Glue DataBrew to clean and prepare the data for analytics.
C.Use AWS Lambda functions to schedule and run the ETL jobs every hour.
D.Use AWS Glue connections to establish connectivity between the data sources and Amazon Redshift.
E.Use the Redshift Data API to load transformed data into Amazon Redshift.

Answer: AD

Explanation:
Glue triggers automate job scheduling at set intervals (hourly). Glue connections provide managed connectivity to sources like RDS and MongoDB and to Redshift as a target, reducing operational burden.

- B is incorrect because DataBrew is for interactive data preparation, not automated hourly ETL pipelines.
- C is incorrect because Lambda-based scheduling adds unnecessary complexity compared to Glue triggers.
- E is incorrect because Glue ETL jobs natively support Redshift loading; using the Data API is unnecessary here.

Question: 58
A company uses an Amazon Redshift cluster that runs on RA3 nodes. The company wants to scale read and write capacity to meet demand. A data engineer needs to identify a solution that will turn on concurrency scaling. Which solution will meet this requirement?

Options
A.Turn on concurrency scaling in workload management (WLM) for Redshift Serverless workgroups.
B.Turn on concurrency scaling at the workload management (WLM) queue level in the Redshift cluster.
C.Turn on concurrency scaling in the settings during the creation of any new Redshift cluster.
D.Turn on concurrency scaling for the daily usage quota for the Redshift cluster.

Answer: B

Explanation:
In provisioned Redshift clusters, concurrency scaling is enabled at the WLM queue level. This allows additional clusters to spin up automatically when queries exceed capacity.

- A is incorrect because Serverless does not use WLM queues in the same way.
- C is incorrect because concurrency scaling is not enabled only at creation; it is configured at WLM level.
- D is incorrect because daily quotas apply to usage, not enabling concurrency scaling.

Question: 59
A data engineer must orchestrate a series of Amazon Athena queries that will run every day. Each query can run for more than 15 minutes. Which combination of steps will meet these requirements MOST cost-effectively? (Choose two.)

Options
A.Use an AWS Lambda function and the Athena Boto3 client start_query_execution API call to invoke the Athena queries programmatically.
B.Create an AWS Step Functions workflow and add two states. Add the first state before the Lambda function. Configure the second state as a Wait state to periodically check whether the Athena query has finished using the Athena Boto3 get_query_execution API call. Configure the workflow to invoke the next query when the current query has finished running.
C.Use an AWS Glue Python shell job and the Athena Boto3 client start_query_execution API call to invoke the Athena queries programmatically.
D.Use an AWS Glue Python shell script to run a sleep timer that checks every 5 minutes to determine whether the current Athena query has finished running successfully. Configure the Python shell script to invoke the next query when the current query has finished running.
E.Use Amazon Managed Workflows for Apache Airflow (Amazon MWAA) to orchestrate the Athena queries in AWS Batch.

Answer: AB

Explanation:
Lambda with Boto3 start_query_execution triggers queries programmatically. Step Functions provides orchestration and waiting for completion with minimal cost. Together they handle long queries cost-effectively.

- C is incorrect because using Glue Python shell jobs adds cost compared to Lambda.
- D is incorrect because polling with sleep timers is inefficient and costly.
- E is incorrect because MWAA introduces unnecessary overhead and cost for this simple orchestration.

Question: 60
A company is migrating on-premises workloads to AWS. The company wants to reduce overall operational overhead. The company also wants to explore serverless options. The company's current workloads use Apache Pig, Apache Oozie, Apache Spark, Apache Hbase, and Apache Flink. The on-premises workloads process petabytes of data in seconds. The company must maintain similar or better performance after the migration to AWS. Which extract, transform, and load (ETL) service will meet these requirements?

Options
A.AWS Glue
B.Amazon EMR
C.AWS Lambda
D.Amazon Redshift

Answer: B

Explanation:
Amazon EMR is best for migrating large-scale big data workloads that use frameworks like Spark, Hive, HBase, and Flink. It provides performance at scale and supports all the required frameworks.

- A is incorrect because Glue supports Spark but not all of Pig, Oozie, and HBase.
- C is incorrect because Lambda is for lightweight serverless functions, not petabyte-scale processing.
- D is incorrect because Redshift is a data warehouse, not a general-purpose ETL platform for these frameworks.

Question: 61
A company currently uses a provisioned Amazon EMR cluster that includes general purpose Amazon EC2 instances. The EMR cluster uses EMR managed scaling between one to five task nodes for the company’s long-running Apache Spark extract, transform, and load (ETL) job. The company runs the ETL job every day.
When the company runs the ETL job, the EMR cluster quickly scales up to five nodes. The EMR cluster often reaches maximum CPU usage, but the memory usage remains under 30%.
The company wants to modify the EMR cluster configuration to reduce the EMR costs to run the daily ETL job.
Which solution will meet these requirements MOST cost-effectively?

Options
A.Increase the maximum number of task nodes for EMR managed scaling to 10.
B.Change the task node type from general purpose EC2 instances to memory optimized EC2 instances.
C.Switch the task node type from general purpose EC2 instances to compute optimized EC2 instances.
D.Reduce the scaling cooldown period for the provisioned EMR cluster.

Answer: C

Explanation:
Since the workload is CPU-bound (high CPU, low memory usage), switching to compute-optimized instances reduces cost by aligning instance type with workload requirements.

- A: Adding more general purpose nodes increases cost without fixing CPU-bound inefficiency.
- B: Memory-optimized instances are unnecessary since memory usage is low.
- D: Shorter cooldown affects scale timing but doesn't address CPU bottleneck.

Question: 62
A company uploads .csv files to an Amazon S3 bucket. The company's data platform team has set up an AWS Glue crawler to perform data discovery and to create the tables and schemas.
An AWS Glue job writes processed data from the tables to an Amazon Redshift database. The AWS Glue job handles column mapping and creates the Amazon Redshift tables in the Redshift database appropriately.
If the company reruns the AWS Glue job for any reason, duplicate records are introduced into the Amazon Redshift tables. The company needs a solution that will update the Redshift tables without duplicates.
Which solution will meet these requirements?

Options
A.Modify the AWS Glue job to copy the rows into a staging Redshift table. Add SQL commands to update the existing rows with new values from the staging Redshift table.
B.Modify the AWS Glue job to load the previously inserted data into a MySQL database. Perform an upsert operation in the MySQL database. Copy the results to the Amazon Redshift tables.
C.Use Apache Spark's DataFrame dropDuplicates() API to eliminate duplicates. Write the data to the Redshift tables.
D.Use the AWS Glue ResolveChoice built-in transform to select the value of the column from the most recent record.

Answer: A

Explanation:
The proper solution is staging + upsert (MERGE pattern). Staging allows controlled deduplication before final insert into production Redshift tables.

- B: Introducing MySQL adds unnecessary complexity and cost.
- C: dropDuplicates() prevents duplicates within one job run but does not handle reruns across multiple loads.
- D: ResolveChoice resolves schema conflicts, not duplicate data issues.

Question: 63
A company is using Amazon Redshift to build a data warehouse solution. The company is loading hundreds of files into a fact table that is in a Redshift cluster.
The company wants the data warehouse solution to achieve the greatest possible throughput. The solution must use cluster resources optimally when the company loads data into the fact table.
Which solution will meet these requirements?

Options
A.Use multiple COPY commands to load the data into the Redshift cluster.
B.Use S3DistCp to load multiple files into Hadoop Distributed File System (HDFS). Use an HDFS connector to ingest the data into the Redshift cluster.
C.Use a number of INSERT statements equal to the number of Redshift cluster nodes. Load the data in parallel into each node.
D.Use a single COPY command to load the data into the Redshift cluster.

Answer: D

Explanation:
COPY is optimized for high throughput bulk loads into Redshift. It parallelizes across cluster nodes automatically, giving maximum performance.

- A: Multiple COPY commands are less efficient than a single COPY.
- B: HDFS staging introduces unnecessary complexity.
- C: INSERT is not optimal for large-scale loads and will be very slow.

Question: 64
A company ingests data from multiple data sources and stores the data in an Amazon S3 bucket. An AWS Glue extract, transform, and load (ETL) job transforms the data and writes the transformed data to an Amazon S3 based data lake. The company uses Amazon Athena to query the data that is in the data lake.
The company needs to identify matching records even when the records do not have a common unique identifier.
Which solution will meet this requirement?

Options
A.Use Amazon Macie pattern matching as part of the ETL job.
B.Train and use the AWS Glue PySpark Filter class in the ETL job.
C.Partition tables and use the ETL job to partition the data on a unique identifier.
D.Train and use the AWS Lake Formation FindMatches transform in the ETL job.

Answer: D

Explanation:
Lake Formation FindMatches is specifically built for record deduplication/matching without unique identifiers, using ML-based fuzzy matching.

- A: Macie is for data security and sensitive data discovery, not record matching.
- B: PySpark Filter cannot perform probabilistic matching without identifiers.
- C: Partitioning improves query efficiency but doesn't solve matching.

Question: 65
A data engineer is using an AWS Glue crawler to catalog data that is in an Amazon S3 bucket. The S3 bucket contains both .csv and json files. The data engineer configured the crawler to exclude the .json files from the catalog.
When the data engineer runs queries in Amazon Athena, the queries also process the excluded .json files. The data engineer wants to resolve this issue. The data engineer needs a solution that will not affect access requirements for the .csv files in the source S3 bucket.
Which solution will meet this requirement with the SHORTEST query times?

Options
A.Adjust the AWS Glue crawler settings to ensure that the AWS Glue crawler also excludes .json files.
B.Use the Athena console to ensure the Athena queries also exclude the .json files.
C.Relocate the .json files to a different path within the S3 bucket.
D.Use S3 bucket policies to block access to the .json files.

Answer: C

Explanation:
Moving JSON files to a different S3 path ensures Athena table points only to CSV files, preventing unnecessary scan costs and improving performance.

- A: Already configured; Athena still sees them if they share the same prefix.
- B: Manual exclusion in queries adds complexity and longer query times.
- D: Bucket policies block access entirely, which is not desired.

Question: 66
A data engineer set up an AWS Lambda function to read an object that is stored in an Amazon S3 bucket. The object is encrypted by an AWS KMS key.
The data engineer configured the Lambda function's execution role to access the S3 bucket. However, the Lambda function encountered an error and failed to retrieve the content of the object.
What is the likely cause of the error?

Options
A.The data engineer misconfigured the permissions of the S3 bucket. The Lambda function could not access the object.
B.The Lambda function is using an outdated SDK version, which caused the read failure.
C.The S3 bucket is located in a different AWS Region than the Region where the data engineer works. Latency issues caused the Lambda function to encounter an error.
D.The Lambda function's execution role does not have the necessary permissions to access the KMS key that can decrypt the S3 object.

Answer: D

Explanation:
To access KMS-encrypted objects, the Lambda role needs both S3 object permissions and KMS Decrypt permission. Missing KMS access causes read failure.

- A: S3 access is already configured.
- B: SDK version is not relevant here.
- C: Region differences cause access errors, not latency issues.

Question: 67
A data engineer has implemented data quality rules in 1,000 AWS Glue Data Catalog tables. Because of a recent change in business requirements, the data engineer must edit the data quality rules.
How should the data engineer meet this requirement with the LEAST operational overhead?

Options
A.Create a pipeline in AWS Glue ETL to edit the rules for each of the 1,000 Data Catalog tables. Use an AWS Lambda function to call the corresponding AWS Glue job for each Data Catalog table.
B.Create an AWS Lambda function that makes an API call to AWS Glue Data Quality to make the edits.
C.Create an Amazon EMR cluster. Run a pipeline on Amazon EMR that edits the rules for each Data Catalog table. Use an AWS Lambda function to run the EMR pipeline.
D.Use the AWS Management Console to edit the rules within the Data Catalog.

Answer: B

Explanation:
Using the Glue Data Quality API via Lambda automates the updates across 1,000 tables with minimal effort and no manual intervention.

- A: Running 1,000 ETL jobs introduces high overhead.
- C: EMR adds unnecessary complexity and management burden.
- D: Manually editing 1,000 tables is inefficient.

Question: 68
Two developers are working on separate application releases. The developers have created feature branches named Branch A and Branch B by using a GitHub repository's master branch as the source.
The developer for Branch A deployed code to the production system. The code for Branch B will merge into a master branch in the following week's scheduled application release.
Which command should the developer for Branch B run before the developer raises a pull request to the master branch?

Options
A.git diff branchB master git commit -m
B.git pull master
C.git rebase master
D.git fetch -b master

Answer: C

Explanation:
git rebase master ensures Branch B is updated with the latest master changes before creating the pull request, preventing conflicts.

- A: git diff only shows differences, does not sync code.
- B: git pull master is not a valid command; git pull origin master would be, but rebasing is cleaner.
- D: git fetch -b master is not a valid command.

Question: 69
A company stores employee data in Amazon Redshift. A table named Employee uses columns named Region ID, Department ID, and Role ID as a compound sort key.
Which queries will MOST increase the speed of query by using a compound sort key of the table? (Choose two.)

Options
A.Select *from Employee where Region ID='North America';
B.Select *from Employee where Region ID='North America' and Department ID=20;
C.Select *from Employee where Department ID=20 and Region ID='North America';
D.Select *from Employee where Role ID=50;
E.Select *from Employee where Region ID='North America' and Role ID=50;

Answer: BE

Explanation:
Compound sort keys optimize queries when the first column(s) in the sort key (Region ID, then Department ID) are included. Queries using Region ID and Department ID (B) or Region ID plus Role ID (E, still starting with Region ID) are fastest.

- A: Region ID alone helps, but adding more key columns improves further.
- C: Department ID first without Region ID does not leverage the compound key.
- D: Role ID alone is not useful for the compound key.

Question: 70
A company receives test results from testing facilities that are located around the world. The company stores the test results in millions of 1 KB JSON files in an Amazon S3 bucket. A data engineer needs to process the files, convert them into Apache Parquet format, and load them into Amazon Redshift tables. The data engineer uses AWS Glue to process the files, AWS Step Functions to orchestrate the processes, and Amazon EventBridge to schedule jobs.
The company recently added more testing facilities. The time required to process files is increasing. The data engineer must reduce the data processing time.
Which solution will MOST reduce the data processing time?

Options
A.Use AWS Lambda to group the raw input files into larger files. Write the larger files back to Amazon S3. Use AWS Glue to process the files. Load the files into the Amazon Redshift tables.
B.Use the AWS Glue dynamic frame file-grouping option to ingest the raw input files. Process the files. Load the files into the Amazon Redshift tables.
C.Use the Amazon Redshift COPY command to move the raw input files from Amazon S3 directly into the Amazon Redshift tables. Process the files in Amazon Redshift.
D.Use Amazon EMR instead of AWS Glue to group the raw input files. Process the files in Amazon EMR. Load the files into the Amazon Redshift tables.

Answer: B

Explanation:
AWS Glue's dynamic frame file-grouping option combines small files efficiently during ETL processing, reducing overhead and speeding up transformation and load.

- A: Lambda is not suited for large-scale file grouping.
- C: COPY command with millions of small files is inefficient.
- D: EMR could work but adds more cost and complexity than Glue's built-in grouping.

Question: 71
A data engineer maintains custom Python scripts that perform a data formatting process that many AWS Lambda functions use. When the data engineer needs to modify the Python scripts, the data engineer must manually update all the Lambda functions.
The data engineer requires a less manual way to update the Lambda functions.
Which solution will meet this requirement?

Options
A. Store the custom Python scripts in a shared Amazon S3 bucket. Store a pointer to the custom scripts in the execution context object.
B. Package the custom Python scripts into Lambda layers. Apply the Lambda layers to the Lambda functions.
C. Store the custom Python scripts in a shared Amazon S3 bucket. Store a pointer to the customer scripts in environment variables.
D. Assign the same alias to each Lambda function. Call each Lambda function by specifying the function's alias.

Answer: B

Explanation:
Using Lambda layers allows packaging of shared code once and attaching to multiple functions. Updating the layer updates all functions, minimizing manual changes.

- A: Would require each function to fetch from S3 manually.
- C: Still requires custom fetch logic per function.
- D: Aliases don’t address shared code updates.

Question: 72
A company stores customer data in an Amazon S3 bucket. Multiple teams in the company want to use the customer data for downstream analysis. The company needs to ensure that the teams do not have access to personally identifiable information (PII) about the customers.
Which solution will meet this requirement with LEAST operational overhead?

Options
A. Use Amazon Macie to create and run a sensitive data discovery job to detect and remove PII.
B. Use S3 Object Lambda to access the data, and use Amazon Comprehend to detect and remove PII.
C. Use Amazon Data Firehose and Amazon Comprehend to detect and remove PII.
D. Use an AWS Glue DataBrew job to store the PII data in a second S3 bucket. Perform analysis on the data that remains in the original S3 bucket.

Answer: A

Explanation:
Amazon Macie automatically scans and detects PII in S3 with minimal manual setup, making it the lowest overhead option.

- B: Object Lambda + Comprehend adds complexity.
- C: Kinesis Firehose not required since data is already in S3.
- D: Glue DataBrew would require extra transformations.

Question: 73
A company stores its processed data in an S3 bucket. The company has a strict data access policy. The company uses IAM roles to grant teams within the company different levels of access to the S3 bucket.
The company wants to receive notifications when a user violates the data access policy. Each notification must include the username of the user who violated the policy.
Which solution will meet these requirements?

Options
A. Use AWS Config rules to detect violations of the data access policy. Set up compliance alarms.
B. Use Amazon CloudWatch metrics to gather object-level metrics. Set up CloudWatch alarms.
C. Use AWS CloudTrail to track object-level events for the S3 bucket. Forward events to Amazon CloudWatch to set up CloudWatch alarms.
D. Use Amazon S3 server access logs to monitor access to the bucket. Forward the access logs to an Amazon CloudWatch log group. Use metric filters on the log group to set up CloudWatch alarms.

Answer: C

Explanation:
CloudTrail provides object-level event logging, including usernames, which can be forwarded to CloudWatch alarms for notifications.

- A: Config checks compliance with resource configuration, not object-level access.
- B: CloudWatch metrics don't provide user identity at object level.
- D: S3 access logs work but require parsing logs; higher overhead.

Question: 74
A company needs to load customer data that comes from a third party into an Amazon Redshift data warehouse. The company stores order data and product data in the same data warehouse. The company wants to use the combined dataset to identify potential new customers.
A data engineer notices that one of the fields in the source data includes values that are in JSON format.
How should the data engineer load the JSON data into the data warehouse with the LEAST effort?

Options
A. Use the SUPER data type to store the data in the Amazon Redshift table.
B. Use AWS Glue to flatten the JSON data and ingest it into the Amazon Redshift table.
C. Use Amazon S3 to store the JSON data. Use Amazon Athena to query the data.
D. Use an AWS Lambda function to flatten the JSON data. Store the data in Amazon S3.

Answer: A

Explanation:
Redshift's SUPER data type supports semi-structured JSON data directly, eliminating the need for flattening or preprocessing.

- B: Glue flattening adds extra ETL effort.
- C: Athena queries do not load into Redshift.
- D: Lambda transformation adds unnecessary complexity.

Question: 75
A company wants to analyze sales records that the company stores in a MySQL database. The company wants to correlate the records with sales opportunities identified by Salesforce.
The company receives 2 GB of sales records every day. The company has 100 GB of identified sales opportunities.
A data engineer needs to develop a process that will analyze and correlate sales records and sales opportunities. The process must run once each night.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A. Use Amazon Managed Workflows for Apache Airflow (Amazon MWAA) to fetch both datasets. Use AWS Lambda functions to correlate the datasets. Use AWS Step Functions to orchestrate the process.
B. Use Amazon AppFlow to fetch sales opportunities from Salesforce. Use AWS Glue to fetch sales records from the MySQL database. Correlate the sales records with the sales opportunities. Use Amazon Managed Workflows for Apache Airflow (Amazon MWAA) to orchestrate the process.
C. Use Amazon AppFlow to fetch sales opportunities from Salesforce. Use AWS Glue to fetch sales records from the MySQL database. Correlate the sales records with sales opportunities. Use AWS Step Functions to orchestrate the process.
D. Use Amazon AppFlow to fetch sales opportunities from Salesforce. Use Amazon Kinesis Data Streams to fetch sales records from the MySQL database. Use Amazon Managed Service for Apache Flink to correlate the datasets. Use AWS Step Functions to orchestrate the process.

Answer: C

Explanation:
AppFlow provides a managed way to fetch Salesforce data, Glue can extract MySQL data, and Step Functions can orchestrate — all serverless, with low operational overhead.

- A: MWAA adds more management than needed.
- B: MWAA orchestration is overkill compared to Step Functions.
- D: Kinesis + Flink is for real-time streams, not nightly batch jobs.

Question: 76
A company stores server logs in an Amazon S3 bucket. The company needs to keep the logs for 1 year. The logs are not required after 1 year.
A data engineer needs a solution to automatically delete logs that are older than 1 year.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A. Define an S3 Lifecycle configuration to delete the logs after 1 year.
B. Create an AWS Lambda function to delete the logs after 1 year.
C. Schedule a cron job on an Amazon EC2 instance to delete the logs after 1 year.
D. Configure an AWS Step Functions state machine to delete the logs after 1 year.

Answer: A

Explanation:
S3 Lifecycle policies can automatically delete objects after a set period (1 year) without any extra management.

- B: Lambda requires additional scheduling.
- C: EC2 cron adds cost and overhead.
- D: Step Functions is unnecessary complexity.

Question: 77
A company is designing a serverless data processing workflow in AWS Step Functions that involves multiple steps. The processing workflow ingests data from an external API, transforms the data by using multiple AWS Lambda functions, and loads the transformed data into Amazon DynamoDB.
The company needs the workflow to perform specific steps based on the content of the incoming data.
Which Step Functions state type should the company use to meet this requirement?

Options
A. Parallel
B. Choice
C. Task
D. Map

Answer: B

Explanation:
The Choice state in Step Functions lets workflows branch based on conditions in the input, enabling different steps for different data.

- A: Parallel runs branches simultaneously, not conditionally.
- C: Task performs a single unit of work, no branching.
- D: Map iterates over items but doesn't handle branching.

Question: 78
A data engineer created a table named cloudtrail_logs in Amazon Athena to query AWS CloudTrail logs and prepare data for audits. The data engineer needs to write a query to display errors with error codes that have occurred since the beginning of 2024. The query must return the 10 most recent errors.
Which query will meet these requirements?

Options
A. select count (*) as TotalEvents, eventname, errorcode, errormessage from cloudtrail_logs where errorcode is not null and eventtime >= '2024-01-01T00:00:00Z' group by eventname, errorcode, errormessage order by TotalEvents desc limit 10;*
*B. select count (*) as TotalEvents, eventname, errorcode, errormessage from cloudtrail_logs where eventtime >= '2024-01-01T00:00:00Z' group by eventname, errorcode, errormessage order by TotalEvents desc limit 10;
C. select count (*) as TotalEvents, eventname, errorcode, errormessage from cloudtrail_logs where eventtime >= '2024-01-01T00:00:00Z' group by eventname, errorcode, errormessage order by eventname asc limit 10;*
*D. select count (*) as TotalEvents, eventname, errorcode, errormessage from cloudtrail_logs where errorcode is not null and eventtime >= '2024-01-01T00:00:00Z' group by eventname, errorcode, errormessage limit 10;

Answer: A

Explanation:
Option A filters for errors (errorcode not null), applies the date filter, groups appropriately, and orders by TotalEvents to get the most recent errors.

- B: Misses errorcode filter.
- C: Orders alphabetically, not by recency or count.
- D: Lacks ordering, so results may not be the most recent.

Question: 79
An online retailer uses multiple delivery partners to deliver products to customers. The delivery partners send order summaries to the retailer. The retailer stores the order summaries in Amazon S3.
Some of the order summaries contain personally identifiable information (PII) about customers. A data engineer needs to detect PII in the order summaries so the company can redact the PII.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A. Amazon Textract
B. Amazon S3 Storage Lens
C. Amazon Macie
D. Amazon SageMaker Data Wrangler

Answer: C

Explanation:
Amazon Macie provides automated PII detection in S3 objects, making it the simplest and most managed solution.

- A: Textract is for extracting text from documents, not PII detection in files.
- B: Storage Lens monitors usage/storage metrics, not PII.
- D: Data Wrangler prepares ML datasets, not for detecting PII.

Question: 80
A company has an Amazon Redshift data warehouse that users access by using a variety of IAM roles. More than 100 users access the data warehouse every day.
The company wants to control user access to the objects based on each user's job role, permissions, and how sensitive the data is.
Which solution will meet these requirements?

Options
A. Use the role-based access control (RBAC) feature of Amazon Redshift.
B. Use the row-level security (RLS) feature of Amazon Redshift.
C. Use the column-level security (CLS) feature of Amazon Redshift.
D. Use dynamic data masking policies in Amazon Redshift.

Answer: A

Explanation:
RBAC allows defining roles and assigning permissions based on job functions, ideal for managing large user bases with varied access needs.

- B: RLS restricts row access but not role-based permissions across many users.
- C: CLS restricts column visibility but doesn’t manage broad role-based control.
- D: Data masking protects sensitive fields but doesn’t fully manage user access by role.

Question: 81
A company uses Amazon DataZone as a data governance and business catalog solution. The company stores data in an Amazon S3 data lake. The company uses AWS Glue with an AWS Glue Data Catalog.
A data engineer needs to publish AWS Glue Data Quality scores to the Amazon DataZone portal.
Which solution will meet this requirement?

Options
A. Create a data quality ruleset with Data Quality Definition language (DQDL) rules that apply to a specific AWS Glue table. Schedule the ruleset to run daily. Configure the Amazon DataZone project to have an Amazon Redshift data source. Enable the data quality configuration for the data source.
B. Configure AWS Glue ETL jobs to use an Evaluate Data Quality transform. Define a data quality ruleset inside the jobs. Configure the Amazon DataZone project to have an AWS Glue data source. Enable the data quality configuration for the data source.
C. Create a data quality ruleset with Data Quality Definition language (DQDL) rules that apply to a specific AWS Glue table. Schedule the ruleset to run daily. Configure the Amazon DataZone project to have an AWS Glue data source. Enable the data quality configuration for the data source.
D. Configure AWS Glue ETL jobs to use an Evaluate Data Quality transform. Define a data quality ruleset inside the jobs. Configure the Amazon DataZone project

Answer: C

Explanation:
Publishing Glue Data Quality scores to DataZone requires DQDL rulesets applied to Glue tables and scheduling them to run, with DataZone configured to use Glue as a source.

- A: Uses Redshift as the data source, not relevant.
- B: Requires embedding quality rules in ETL jobs, higher overhead.
- D: Incomplete and lacks full requirement coverage.

Question: 82
A company has a data warehouse in Amazon Redshift. To comply with security regulations, the company needs to log and store all user activities and connection activities for the data warehouse.
Which solution will meet these requirements?

Options
A. Create an Amazon S3 bucket. Enable logging for the Amazon Redshift cluster. Specify the S3 bucket in the logging configuration to store the logs.
B. Create an Amazon Elastic File System (Amazon EFS) file system. Enable logging for the Amazon Redshift cluster. Write logs to the EFS file system.
C. Create an Amazon Aurora MySQL database. Enable logging for the Amazon Redshift cluster. Write the logs to a table in the Aurora MySQL database.
D. Create an Amazon Elastic Block Store (Amazon EBS) volume. Enable logging for the Amazon Redshift cluster. Write the logs to the EBS volume.

Answer: A

Explanation:
Redshift audit logging integrates natively with S3, making it secure, scalable, and compliant.

- B, C, D: Not supported destinations for Redshift logs.

Question: 83
A company wants to migrate a data warehouse from Teradata to Amazon Redshift.
Which solution will meet this requirement with the LEAST operational effort?

Options
A. Use AWS Database Migration Service (AWS DMS) Schema Conversion to migrate the schema. Use AWS DMS to migrate the data.
B. Use the AWS Schema Conversion Tool (AWS SCT) to migrate the schema. Use AWS Database Migration Service (AWS DMS) to migrate the data.
C. Use AWS Database Migration Service (AWS DMS) to migrate the data. Use automatic schema conversion.
D. Manually export the schema definition from Teradata. Apply the schema to the Amazon Redshift database. Use AWS Database Migration Service (AWS DMS) to migrate the data.

Answer: B

Explanation:
AWS SCT is the standard tool for converting Teradata schemas to Redshift-compatible formats, and DMS handles the data migration.

- A: "DMS Schema Conversion" is not a feature.
- C: DMS does not do schema conversion.
- D: Manual schema export requires high effort.

Question: 84
A company uses a variety of AWS and third-party data stores. The company wants to consolidate all the data into a central data warehouse to perform analytics. Users need fast response times for analytics queries.
The company uses Amazon QuickSight in direct query mode to visualize the data. Users normally run queries during a few hours each day with unpredictable spikes.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A. Use Amazon Redshift Serverless to load all the data into Amazon Redshift managed storage (RMS).
B. Use Amazon Athena to load all the data into Amazon S3 in Apache Parquet format.
C. Use Amazon Redshift provisioned clusters to load all the data into Amazon Redshift managed storage (RMS).
D. Use Amazon Aurora PostgreSQL to load all the data into Aurora.

Answer: A

Explanation:
Redshift Serverless auto-scales with unpredictable workloads, requires no cluster management, and provides fast queries.

- B: Athena is cost-efficient but not ideal for fast response times in direct query mode.
- C: Provisioned Redshift requires capacity management.
- D: Aurora is not a data warehouse solution.

Question: 85
A data engineer uses Amazon Kinesis Data Streams to ingest and process records that contain user behavior data from an application every day.
The data engineer notices that the data stream is experiencing throttling because hot shards receive much more data than other shards in the data stream.
How should the data engineer resolve the throttling issue?

Options
A. Use a random partition key to distribute the ingested records.
B. Increase the number of shards in the data stream. Distribute the records across the shards.
C. Limit the number of records that are sent each second by the producer to match the capacity of the stream.
D. Decrease the size of the records that the producer sends to match the capacity of the stream.

Answer: A

Explanation:
Using random partition keys avoids hot shard problems by distributing data evenly across shards.

- B: Adding shards without fixing partitioning logic won't solve hot shard skew.
- C, D: Throttling isn't due to total record size but uneven distribution.

Question: 86
A company has a data processing pipeline that includes several dozen steps. The data processing pipeline needs to send alerts in real time when a step fails or succeeds. The data processing pipeline uses a combination of Amazon S3 buckets, AWS Lambda functions, and AWS Step Functions state machines.
A data engineer needs to create a solution to monitor the entire pipeline.
Which solution will meet these requirements?

Options
A. Configure the Step Functions state machines to store notifications in an Amazon S3 bucket when the state machines finish running. Enable S3 event notifications on the S3 bucket.
B. Configure the AWS Lambda functions to store notifications in an Amazon S3 bucket when the state machines finish running. Enable S3 event notifications on the S3 bucket.
C. Use AWS CloudTrail to send a message to an Amazon Simple Notification Service (Amazon SNS) topic that sends notifications when a state machine fails to run or succeeds to run.
D. Configure an Amazon EventBridge rule to react when the execution status of a state machine changes. Configure the rule to send a message to an Amazon SNS topic that sends notifications.

Answer: D

Explanation:
EventBridge integrates natively with Step Functions to react to execution status changes and send SNS notifications in real time.

- A, B: Storing notifications in S3 adds latency and overhead.
- C: CloudTrail logs API activity, not real-time status changes.

Question: 87
A company has an application that uses an Amazon API Gateway REST API and an AWS Lambda function to retrieve data from an Amazon DynamoDB instance. Users recently reported intermittent high latency in the application's response times. A data engineer finds that the Lambda function experiences frequent throttling when the company's other Lambda functions experience increased invocations.
The company wants to ensure the API's Lambda function operate without being affected by other Lambda functions.
Which solution will meet this requirement MOST cost-effectively?

Options
A. Increase the number of read capacity unit (RCU) in DynamoDB.
B. Configure provisioned concurrency for the Lambda function.
C. Configure reserved concurrency for the Lambda function.
D. Increase the Lambda function timeout and allocated memory.

Answer: C

Explanation:
Reserved concurrency ensures that a specific function always has guaranteed capacity, isolating it from throttling caused by other functions.

- A: DynamoDB RCU doesn’t address Lambda throttling.
- B: Provisioned concurrency ensures warm starts but is costlier.
- D: Timeout/memory doesn’t prevent throttling.

Question: 88
A company has as JSON file that contains personally identifiable information (PII) data and non-PII data. The company needs to make the data available for querying and analysis.
The non-PII data must be available to everyone in the company. The PII data must be available only to a limited group of employees.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A. Store the JSON file in an Amazon S3 bucket. Configure AWS Glue to split the file into one file that contains the PII data and one file that contains the non-PII data. Store the output files in separate S3 buckets. Grant the required access to the buckets based on the type of user.
B. Store the JSON file in an Amazon S3 bucket. Use Amazon Macie to identify PII data and to grant access based on the type of user.
C. Store the JSON file in an Amazon S3 bucket. Catalog the file schema in AWS Lake Formation. Use Lake Formation permissions to provide access to the required data based on the type of user.
D. Create two Amazon RDS PostgreSQL databases. Load the PII data and the non-PII data into the separate databases. Grant access to the databases based on the type of user.

Answer: C

Explanation:
Lake Formation provides fine-grained access control at column/row level directly on S3 data, minimizing effort.

- A: Splitting files and managing separate buckets adds overhead.
- B: Macie detects PII but doesn’t enforce ongoing access control.
- D: Moving to RDS adds cost and unnecessary complexity.

Question: 89
A company uses AWS Key Management Service (AWS KMS) to encrypt an Amazon Redshift cluster. The company wants to configure a cross-Region snapshot of the Redshift cluster as part of disaster recovery (DR) strategy.
A data engineer needs to use the AWS CLI to create the cross-Region snapshot.
Which combination of steps will meet these requirements? (Choose two.)

Options
A. Create a KMS key and configure a snapshot copy grant in the source AWS Region.
B. In the source AWS Region, enable snapshot copying. Specify the name of the snapshot copy grant that is created in the destination AWS Region.
C. In the source AWS Region, enable snapshot copying. Specify the name of the snapshot copy grant that is created in the source AWS Region.
D. Create a KMS key and configure a snapshot copy grant in the destination AWS Region.
E. Convert the cluster to a Multi-AZ deployment.

Answer: AC

Explanation:
You must create a snapshot copy grant in the **source Region** (A) and configure snapshot copying with that grant in the same Region (C).

- B: Grants must be created in the source, not destination.
- D: Destination KMS key is not required for setup.
- E: Multi-AZ doesn't apply to Redshift.

Question: 90
A company is using Amazon S3 to build a data lake. The company needs to replicate records from multiple source databases into Apache Parquet format. Most of the source databases are hosted on Amazon RDS. However, one source database is an on-premises Microsoft SQL Server Enterprise instance. The company needs to implement a solution to replicate existing data from all source databases and all future changes to the target S3 data lake.
Which solution will meet these requirements MOST cost-effectively?

Options
A. Use one AWS Glue job to replicate existing data. Use a second AWS Glue job to replicate future changes.
B. Use AWS Database Migration Service (AWS DMS) to replicate existing data. Use AWS Glue jobs to replicate future changes.
C. Use AWS Database Migration Service (AWS DMS) to replicate existing data and future changes.
D. Use AWS Glue jobs to replicate existing data. Use Amazon Kinesis Data Streams to replicate future changes.

Answer: C

Explanation:
AWS DMS supports both full load and change data capture (CDC) for ongoing replication to S3 in Parquet, handling both existing and future changes with minimal cost/overhead.

- A, B, D: Splitting load/future changes across Glue or Kinesis adds unnecessary complexity.

Question: 91
A data engineer uses Amazon Managed Workflows for Apache Airflow (Amazon MWAA) to run data pipelines in an AWS account.
A workflow recently failed to run. The data engineer needs to use Apache Airflow logs to diagnose the failure of the workflow.
Which log type should the data engineer use to diagnose the cause of the failure?

Options
A.YourEnvironmentName-WebServer
B.YourEnvironmentName-Scheduler
C.YourEnvironmentName-DAGProcessing
D.YourEnvironmentName-Task

Answer: D

Explanation:
Task logs contain detailed information about task execution and failures. They are the most useful to troubleshoot workflow errors.

- A: WebServer logs capture UI/API requests, not workflow execution.
- B: Scheduler logs show scheduling events but not detailed task errors.
- C: DAGProcessing logs help in DAG parsing issues but not runtime errors.

Question: 92
A finance company uses Amazon Redshift as a data warehouse. The company stores the data in a shared Amazon S3 bucket. The company uses Amazon Redshift Spectrum to access the data that is stored in the S3 bucket. The data comes from certified third-party data providers. Each third-party data provider has unique connection details.
To comply with regulations, the company must ensure that none of the data is accessible from outside the company's AWS environment.
Which combination of steps should the company take to meet these requirements? (Choose two.)

Options
A.Replace the existing Redshift cluster with a new Redshift cluster that is in a private subnet. Use an interface VPC endpoint to connect to the Redshift cluster. Use a NAT gateway to give Redshift access to the S3 bucket.
B.Create an AWS CloudHSM hardware security module (HSM) for each data provider. Encrypt each data provider's data by using the corresponding HSM for each data provider.
C.Turn on enhanced VPC routing for the Amazon Redshift cluster. Set up an AWS Direct Connect connection and configure a connection between each data provider and the finance company’s VPC.
D.Define table constraints for the primary keys and the foreign keys.
E.Use federated queries to access the data from each data provider. Do not upload the data to the S3 bucket. Perform the federated queries through a gateway VPC endpoint.

Answer: AC

Explanation:
Enhanced VPC routing (C) ensures all Redshift Spectrum traffic goes through the VPC, and a private subnet with endpoints (A) ensures data cannot be accessed publicly.

- B: HSM is for encryption, not access control.
- D: Constraints enforce schema, not security.
- E: Federated queries don’t address security for existing S3 data.

Question: 93
Files from multiple data sources arrive in an Amazon S3 bucket on a regular basis. A data engineer wants to ingest new files into Amazon Redshift in near real time when the new files arrive in the S3 bucket.
Which solution will meet these requirements?

Options
A.Use the query editor v2 to schedule a COPY command to load new files into Amazon Redshift.
B.Use the zero-ETL integration between Amazon Aurora and Amazon Redshift to load new files into Amazon Redshift.
C.Use AWS Glue job bookmarks to extract, transform, and load (ETL) load new files into Amazon Redshift.
D.Use S3 Event Notifications to invoke an AWS Lambda function that loads new files into Amazon Redshift.

Answer: D

Explanation:
S3 Event Notifications + Lambda allow near real-time ingestion when new files land. Lambda can trigger Redshift COPY efficiently.

- A: Query editor scheduling is batch, not near real-time.
- B: Aurora-Redshift zero-ETL is unrelated (data is from S3, not Aurora).
- C: Glue bookmarks handle deduplication but are not real-time.

Question: 94
A technology company currently uses Amazon Kinesis Data Streams to collect log data in real time. The company wants to use Amazon Redshift for downstream real-time queries and to enrich the log data.
Which solution will ingest data into Amazon Redshift with the LEAST operational overhead?

Options
A.Set up an Amazon Kinesis Data Firehose delivery stream to send data to a Redshift provisioned cluster table.
B.Set up an Amazon Kinesis Data Firehose delivery stream to send data to Amazon S3. Configure a Redshift provisioned cluster to load data every minute.
C.Configure Amazon Managed Service for Apache Flink (previously known as Amazon Kinesis Data Analytics) to send data directly to a Redshift provisioned cluster table.
D.Use Amazon Redshift streaming ingestion from Kinesis Data Streams and to present data as a materialized view.

Answer: D

Explanation:
Redshift streaming ingestion (D) allows direct integration with Kinesis Data Streams, reducing latency and removing the need for intermediate services.

- A: Firehose → Redshift adds buffering delay and management overhead.
- B: Firehose to S3 then COPY into Redshift introduces batch latency.
- C: Flink direct output requires more custom development and ops.

Question: 95
A company maintains a data warehouse in an on-premises Oracle database. The company wants to build a data lake on AWS. The company wants to load data warehouse tables into Amazon S3 and synchronize the tables with incremental data that arrives from the data warehouse every day.
Each table has a column that contains monotonically increasing values. The size of each table is less than 50 GB. The data warehouse tables are refreshed every night between 1 AM and 2 AM. A business intelligence team queries the tables between 10 AM and 8 PM every day.
Which solution will meet these requirements in the MOST operationally efficient way?

Options
A.Use an AWS Database Migration Service (AWS DMS) full load plus CDC job to load tables that contain monotonically increasing data columns from the on-premises data warehouse to Amazon S3. Use custom logic in AWS Glue to append the daily incremental data to a full-load copy that is in Amazon S3.
B.Use an AWS Glue Java Database Connectivity (JDBC) connection. Configure a job bookmark for a column that contains monotonically increasing values. Write custom logic to append the daily incremental data to a full-load copy that is in Amazon S3.
C.Use an AWS Database Migration Service (AWS DMS) full load migration to load the data warehouse tables into Amazon S3 every day. Overwrite the previous day's full-load copy every day.
D.Use AWS Glue to load a full copy of the data warehouse tables into Amazon S3 every day. Overwrite the previous day's full-load copy every day.

Answer: A

Explanation:
DMS full load + CDC captures ongoing changes efficiently and appends incremental updates to S3. This ensures operational efficiency and low latency.

- B: JDBC + bookmarks require custom coding.
- C: Full reload daily is inefficient.
- D: Glue full reload daily increases processing cost.

Question: 96
A company is building a data lake for a new analytics team. The company is using Amazon S3 for storage and Amazon Athena for query analysis. All data that is in Amazon S3 is in Apache Parquet format.
The company is running a new Oracle database as a source system in the company's data center. The company has 70 tables in the Oracle database. All the tables have primary keys. Data can occasionally change in the source system. The company wants to ingest the tables every day into the data lake.
Which solution will meet this requirement with the LEAST effort?

Options
A.Create an Apache Sqoop job in Amazon EMR to read the data from the Oracle database. Configure the Sqoop job to write the data to Amazon S3 in Parquet format.
B.Create an AWS Glue connection to the Oracle database. Create an AWS Glue bookmark job to ingest the data incrementally and to write the data to Amazon S3 in Parquet format.
C.Create an AWS Database Migration Service (AWS DMS) task for ongoing replication. Set the Oracle database as the source. Set Amazon S3 as the target. Configure the task to write the data in Parquet format.
D.Create an Oracle database in Amazon RDS. Use AWS Database Migration Service (AWS DMS) to migrate the on-premises Oracle database to Amazon RDS. Configure triggers on the tables to invoke AWS Lambda functions to write changed records to Amazon S3 in Parquet format.

Answer: B

Explanation:
Glue bookmarks track incremental changes, making daily ingestion straightforward and automated. Glue also supports writing directly in Parquet.

- A: EMR + Sqoop adds operational burden.
- C: DMS replication is possible but more complex than Glue bookmark for daily ingestion.
- D: Migrating to RDS + Lambda triggers adds unnecessary overhead.

Question: 97
A transportation company wants to track vehicle movements by capturing geolocation records. The records are 10 bytes in size. The company receives up to 10,000 records every second. Data transmission delays of a few minutes are acceptable because of unreliable network conditions. The transportation company wants to use Amazon Kinesis Data Streams to ingest the geolocation data. The company needs a reliable mechanism to send data to Kinesis Data Streams. The company needs to maximize the throughput efficiency of the Kinesis shards.
Which solution will meet these requirements in the MOST operationally efficient way?

Options
A.Kinesis Agent
B.Kinesis Producer Library (KPL)
C.Amazon Kinesis Data Firehose
D.Kinesis SDK

Answer: B

Explanation:
KPL batches and aggregates records efficiently before sending them to Kinesis Data Streams, maximizing shard throughput and efficiency.

- A: Kinesis Agent is for file-based streaming, not geolocation data.
- C: Firehose is for delivery into sinks, not direct stream ingestion.
- D: Kinesis SDK lacks built-in batching/aggregation, less efficient.

Question: 98
An investment company needs to manage and extract insights from a volume of semi-structured data that grows continuously.
A data engineer needs to deduplicate the semi-structured data, remove records that are duplicates, and remove common misspellings of duplicates.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Use the FindMatches feature of AWS Glue to remove duplicate records.
B.Use non-Windows functions in Amazon Athena to remove duplicate records.
C.Use Amazon Neptune ML and an Apache Gremlin script to remove duplicate records.
D.Use the global tables feature of Amazon DynamoDB to prevent duplicate data.

Answer: A

Explanation:
Glue FindMatches is built for deduplication and fuzzy matching (e.g., misspellings) with minimal setup.

- B: Athena SQL can remove exact duplicates but not fuzzy matches.
- C: Neptune ML requires graph modeling, too complex.
- D: DynamoDB global tables prevent duplicates across regions, not within semi-structured data.

Question: 99
A company is building an inventory management system and an inventory reordering system to automatically reorder products. Both systems use Amazon Kinesis Data Streams. The inventory management system uses the Amazon Kinesis Producer Library (KPL) to publish data to a stream. The inventory reordering system uses the Amazon Kinesis Client Library (KCL) to consume data from the stream. The company configures the stream to scale up and down as needed.
Before the company deploys the systems to production, the company discovers that the inventory reordering system received duplicated data.
Which factors could have caused the reordering system to receive duplicated data? (Choose two.)

Options
A.The producer experienced network-related timeouts.
B.The stream's value for the IteratorAgeMilliseconds metric was too high.
C.There was a change in the number of shards, record processors, or both.
D.The AggregationEnabled configuration property was set to true.
E.The max_records configuration property was set to a number that was too high.

Answer: AC

Explanation:
Network retries (A) can cause duplicate writes. Scaling shards or processors (C) can cause consumers to reprocess records during resharding or failover.

- B: IteratorAgeMilliseconds indicates delay, not duplication.
- D: AggregationEnabled improves efficiency, doesn't create duplicates.
- E: max_records affects batch size, not duplication.

Question: 100
An ecommerce company operates a complex order fulfilment process that spans several operational systems hosted in AWS. Each of the operational systems has a Java Database Connectivity (JDBC)-compliant relational database where the latest processing state is captured.
The company needs to give an operations team the ability to track orders on an hourly basis across the entire fulfillment process.
Which solution will meet these requirements with the LEAST development overhead?

Options
A.Use AWS Glue to build ingestion pipelines from the operational systems into Amazon Redshift. Build dashboards in Amazon QuickSight that track the orders.
B.Use AWS Glue to build ingestion pipelines from the operational systems into Amazon DynamoDB. Build dashboards in Amazon QuickSight that track the orders.
C.Use AWS Database Migration Service (AWS DMS) to capture changed records in the operational systems. Publish the changes to an Amazon DynamoDB table in a different AWS region from the source database. Build Grafana dashboards that track the orders.
D.Use AWS Database Migration Service (AWS DMS) to capture changed records in the operational systems. Publish the changes to an Amazon DynamoDB table in a different AWS region from the source database. Build Amazon QuickSight dashboards that track the orders.

Answer: A

Explanation:
Redshift is best for combining relational data from multiple systems. Glue can ingest JDBC sources into Redshift easily, and QuickSight integrates natively for dashboards.

- B: DynamoDB is not optimized for cross-system relational analytics.
- C: DynamoDB + Grafana adds more development work and is not suited for relational analytics.
- D: DynamoDB still not the right fit for this relational aggregation.

Question: 101
A data engineer must orchestrate a data pipeline that consists of one AWS Lambda function and one AWS Glue job. The solution must integrate with AWS services. Which solution will meet these requirements with the LEAST management overhead?

Options
A.Use an AWS Step Functions workflow that includes a state machine. Configure the state machine to run the Lambda function and then the AWS Glue job.
B.Use an Apache Airflow workflow that is deployed on an Amazon EC2 instance. Define a directed acyclic graph (DAG) in which the first task is to call the Lambda function and the second task is to call the AWS Glue job.
C.Use an AWS Glue workflow to run the Lambda function and then the AWS Glue job.
D.Use an Apache Airflow workflow that is deployed on Amazon Elastic Kubernetes Service (Amazon EKS). Define a directed acyclic graph (DAG) in which the first task is to call the Lambda function and the second task is to call the AWS Glue job.

Answer: A

Explanation:
Step Functions is a fully managed AWS-native orchestration service. It can directly integrate with Lambda and Glue, allowing the pipeline to be managed with minimal overhead.

- B is incorrect because managing Airflow on EC2 introduces more operational burden.
- C is incorrect because Glue workflows cannot orchestrate Lambda.
- D is incorrect because running Airflow on EKS increases complexity and management.

Question: 102
A company needs to set up a data catalog and metadata management for data sources that run in the AWS Cloud. The company will use the data catalog to maintain the metadata of all the objects that are in a set of data stores. The data stores include structured sources such as Amazon RDS and Amazon Redshift. The data stores also include semistructured sources such as JSON files and .xml files that are stored in Amazon S3.
The company needs a solution that will update the data catalog on a regular basis. The solution also must detect changes to the source metadata.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Use Amazon Aurora as the data catalog. Create AWS Lambda functions that will connect to the data catalog. Configure the Lambda functions to gather the metadata information from multiple sources and to update the Aurora data catalog. Schedule the Lambda functions to run periodically.
B.Use the AWS Glue Data Catalog as the central metadata repository. Use AWS Glue crawlers to connect to multiple data stores and to update the Data Catalog with metadata changes. Schedule the crawlers to run periodically to update the metadata catalog.
C.Use Amazon DynamoDB as the data catalog. Create AWS Lambda functions that will connect to the data catalog. Configure the Lambda functions to gather the metadata information from multiple sources and to update the DynamoDB data catalog. Schedule the Lambda functions to run periodically.
D.Use the AWS Glue Data Catalog as the central metadata repository. Extract the schema for Amazon RDS and Amazon Redshift sources, and build the Data Catalog. Use AWS Glue crawlers for data that is in Amazon S3 to infer the schema and to automatically update the Data Catalog.

Answer: B

Explanation:
The Glue Data Catalog is AWS's managed metadata repository. Glue crawlers can automatically detect schema changes across RDS, Redshift, and S3 and update the catalog with minimal management.

- A is incorrect because Aurora is a database, not a metadata catalog.
- C is incorrect because DynamoDB is not designed for metadata management.
- D is incorrect because relying only on manual schema extraction adds operational overhead.

Question: 103
A company stores data from an application in an Amazon DynamoDB table that operates in provisioned capacity mode. The workloads of the application have predictable throughput load on a regular schedule. Every Monday, there is an immediate increase in activity early in the morning. The application has very low usage during weekends.
The company must ensure that the application performs consistently during peak usage times.
Which solution will meet these requirements in the MOST cost-effective way?

Options
A.Increase the provisioned capacity to the maximum capacity that is currently present during peak load times.
B.Divide the table into two tables. Provision each table with half of the provisioned capacity of the original table. Spread queries evenly across both tables.
C.Use AWS Application Auto Scaling to schedule higher provisioned capacity for peak usage times. Schedule lower capacity during off-peak times.
D.Change the capacity mode from provisioned to on-demand. Configure the table to scale up and scale down based on the load on the table.

Answer: C

Explanation:
Application Auto Scaling allows DynamoDB provisioned mode to scale automatically on a schedule. This ensures enough throughput during predictable peak loads and reduces cost during off-peak times.

- A is incorrect because overprovisioning for all times wastes cost.
- B is incorrect because splitting tables adds complexity without solving capacity issues.
- D is incorrect because on-demand is more expensive if usage patterns are predictable.

Question: 104
A company is planning to migrate on-premises Apache Hadoop clusters to Amazon EMR. The company also needs to migrate a data catalog into a persistent storage solution.
The company currently stores the data catalog in an on-premises Apache Hive metastore on the Hadoop clusters. The company requires a serverless solution to migrate the data catalog.
Which solution will meet these requirements MOST cost-effectively?

Options
A.Use AWS Database Migration Service (AWS DMS) to migrate the Hive metastore into Amazon S3. Configure AWS Glue Data Catalog to scan Amazon S3 to produce the data catalog.
B.Configure a Hive metastore in Amazon EMR. Migrate the existing on-premises Hive metastore into Amazon EMR. Use AWS Glue Data Catalog to store the company's data catalog as an external data catalog.
C.Configure an external Hive metastore in Amazon EMR. Migrate the existing on-premises Hive metastore into Amazon EMR. Use Amazon Aurora MySQL to store the company's data catalog.
D.Configure a new Hive metastore in Amazon EMR. Migrate the existing on-premises Hive metastore into Amazon EMR. Use the new metastore as the company's data catalog.

Answer: B

Explanation:
Glue Data Catalog is the serverless and persistent solution for metadata management.
Migrating the Hive metastore into EMR and connecting it with Glue as the external catalog is cost-effective and serverless.

- A is incorrect because S3 is object storage, not a metastore.
- C is incorrect because Aurora adds unnecessary cost and management.
- D is incorrect because EMR metastore is tied to clusters, not serverless.

Question: 105
A company uses an Amazon Redshift provisioned cluster as its database. The Redshift cluster has five reserved ra3.4xlarge nodes and uses key distribution.
A data engineer notices that one of the nodes frequently has a CPU load over 90%. SQL Queries that run on the node are queued. The other four nodes usually have a CPU load under 15% during daily operations.
The data engineer wants to maintain the current number of compute nodes. The data engineer also wants to balance the load more evenly across all five compute nodes.
Which solution will meet these requirements?

Options
A.Change the sort key to be the data column that is most often used in a WHERE clause of the SQL SELECT statement.
B.Change the distribution key to the table column that has the largest dimension.
C.Upgrade the reserved node from ra3.4xlarge to ra3.16xlarge.
D.Change the primary key to be the data column that is most often used in a WHERE clause of the SQL SELECT statement.

Answer: B

Explanation:
Key distribution can cause skew if the distribution key isn't chosen correctly. Using the column with the largest dimension spreads data more evenly across nodes, balancing load.

- A is incorrect because sort keys improve query efficiency, not distribution.
- C is incorrect because upgrading nodes does not solve skew; it only adds cost.
- D is incorrect because primary key is logical, not used for physical distribution.

Question: 106
A security company stores IoT data that is in JSON format in an Amazon S3 bucket. The data structure can change when the company upgrades the IoT devices. The company wants to create a data catalog that includes the IoT data. The company's analytics department will use the data catalog to index the data.
Which solution will meet these requirements MOST cost-effectively?

Options
A.Create an AWS Glue Data Catalog. Configure an AWS Glue Schema Registry. Create a new AWS Glue workload to orchestrate the ingestion of the data that the analytics department will use into Amazon Redshift Serverless.
B.Create an Amazon Redshift provisioned cluster. Create an Amazon Redshift Spectrum database for the analytics department to explore the data that is in Amazon S3. Create Redshift stored procedures to load the data into Amazon Redshift.
C.Create an Amazon Athena workgroup. Explore the data that is in Amazon S3 by using Apache Spark through Athena. Provide the Athena workgroup schema and tables to the analytics department.
D.Create an AWS Glue Data Catalog. Configure an AWS Glue Schema Registry. Create AWS Lambda user defined functions (UDFs) by using the Amazon Redshift Data API. Create an AWS Step Functions job to orchestrate the ingestion of the data that the analytics department will use into Amazon Redshift Serverless.

Answer: A

Explanation:
Glue Data Catalog with Schema Registry automatically handles schema evolution for semi-structured JSON data in S3. This provides cost-effective indexing for the analytics team.

- B is incorrect because Redshift provisioning adds cost and isn't serverless.
- C is incorrect because Athena with Spark adds complexity and cost.
- D is incorrect because orchestrating ingestion into Redshift Serverless is unnecessary when analysis can be done directly.

Question: 107
A company stores details about transactions in an Amazon S3 bucket. The company wants to log all writes to the S3 bucket into another S3 bucket that is in the same AWS Region.
Which solution will meet this requirement with the LEAST operational effort?

Options
A.Configure an S3 Event Notifications rule for all activities on the transactions S3 bucket to invoke an AWS Lambda function. Program the Lambda function to write the event to Amazon Kinesis Data Firehose. Configure Kinesis Data Firehose to write the event to the logs S3 bucket.
B.Create a trail of management events in AWS CloudTraiL. Configure the trail to receive data from the transactions S3 bucket. Specify an empty prefix and write-only events. Specify the logs S3 bucket as the destination bucket.
C.Configure an S3 Event Notifications rule for all activities on the transactions S3 bucket to invoke an AWS Lambda function. Program the Lambda function to write the events to the logs S3 bucket.
D.Create a trail of data events in AWS CloudTraiL. Configure the trail to receive data from the transactions S3 bucket. Specify an empty prefix and write-only events. Specify the logs S3 bucket as the destination bucket.

Answer: D

Explanation:
CloudTrail data events provide object-level logging for S3. Configuring data events for write-only activity ensures all writes are captured automatically in the log bucket.

- A is incorrect because setting up Lambda + Firehose adds unnecessary components.
- B is incorrect because management events don't include S3 object-level data writes.
- C is incorrect because using Lambda adds overhead when CloudTrail data events already cover this.

Question: 108
A data engineer needs to maintain a central metadata repository that users access through Amazon EMR and Amazon Athena queries. The repository needs to provide the schema and properties of many tables. Some of the metadata is stored in Apache Hive. The data engineer needs to import the metadata from Hive into the central metadata repository.
Which solution will meet these requirements with the LEAST development effort?

Options
A.Use Amazon EMR and Apache Ranger.
B.Use a Hive metastore on an EMR cluster.
C.Use the AWS Glue Data Catalog.
D.Use a metastore on an Amazon RDS for MySQL DB instance.

Answer: C

Explanation:
AWS Glue Data Catalog is serverless, integrates with Athena and EMR, and can directly import Hive metadata with minimal development.

- A is incorrect because Ranger is for fine-grained access control, not metadata management.
- B is incorrect because an EMR-based Hive metastore ties metadata to clusters.
- D is incorrect because RDS requires custom setup and management.

Question: 109
A company needs to build a data lake in AWS. The company must provide row-level data access and column-level data access to specific teams. The teams will access the data by using Amazon Athena, Amazon Redshift Spectrum, and Apache Hive from Amazon EMR.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Use Amazon S3 for data lake storage. Use S3 access policies to restrict data access by rows and columns. Provide data access through Amazon S3.
B.Use Amazon S3 for data lake storage. Use Apache Ranger through Amazon EMR to restrict data access by rows and columns. Provide data access by using Apache Pig.
C.Use Amazon Redshift for data lake storage. Use Redshift security policies to restrict data access by rows and columns. Provide data access by using Apache Spark and Amazon Athena federated queries.
D.Use Amazon S3 for data lake storage. Use AWS Lake Formation to restrict data access by rows and columns. Provide data access through AWS Lake Formation.

Answer: D

Explanation:
Lake Formation integrates with Athena, Redshift Spectrum, and EMR. It provides native row-level and column-level access controls for S3-based data lakes with minimal overhead.

- A is incorrect because S3 IAM policies can't enforce row/column access.
- B is incorrect because Apache Ranger requires heavy operational management.
- C is incorrect because Redshift isn't meant to be the primary data lake storage.

Question: 110
An airline company is collecting metrics about flight activities for analytics. The company is conducting a proof of concept (POC) test to show how analytics can provide insights that the company can use to increase on-time departures.
The POC test uses objects in Amazon S3 that contain the metrics in .csv format. The POC test uses Amazon Athena to query the data. The data is partitioned in the S3 bucket by date.
As the amount of data increases, the company wants to optimize the storage solution to improve query performance.
Which combination of solutions will meet these requirements? (Choose two.)

Options
A.Add a randomized string to the beginning of the keys in Amazon S3 to get more throughput across partitions.
B.Use an S3 bucket that is in the same account that uses Athena to query the data.
C.Use an S3 bucket that is in the same AWS Region where the company runs Athena queries.
D.Preprocess the .csv data to JSON format by fetching only the document keys that the query requires.
E.Preprocess the .csv data to Apache Parquet format by fetching only the data blocks that are needed for predicates.

Answer: CE

Explanation:
Storing data in the same Region as Athena minimizes latency. Converting CSV to Parquet enables column pruning and efficient scanning, improving Athena query performance.

- A is incorrect because adding random prefixes reduces partition pruning efficiency.
- B is incorrect because cross-account S3 access does not significantly impact query performance compared to same-Region optimization.
- D is incorrect because JSON is less efficient than Parquet for analytics queries

Question: 111
During a security review, a company identified a vulnerability in an AWS Glue job. The company discovered that credentials to access an Amazon Redshift cluster were hard coded in the job script.
A data engineer must remediate the security vulnerability in the AWS Glue job. The solution must securely store the credentials.
Which combination of steps should the data engineer take to meet these requirements? (Choose two.)

Options
A.Store the credentials in the AWS Glue job parameters.
B.Store the credentials in a configuration file that is in an Amazon S3 bucket.
C.Access the credentials from a configuration file that is in an Amazon S3 bucket by using the AWS Glue job.
D.Store the credentials in AWS Secrets Manager.
E.Grant the AWS Glue job IAM role access to the stored credentials.

Answer: DE

Explanation:
Storing secrets in AWS Secrets Manager provides secure and encrypted management of credentials. Granting the Glue job IAM role permission ensures the job retrieves credentials securely at runtime.

- A is incorrect because job parameters are not secure for storing credentials.
- B is incorrect because storing credentials in S3 is insecure.
- C is incorrect because accessing credentials from S3 doesn’t protect them properly.

Question: 112
A data engineer uses Amazon Redshift to run resource-intensive analytics processes once every month. Every month, the data engineer creates a new Redshift provisioned cluster. The data engineer deletes the Redshift provisioned cluster after the analytics processes are complete every month. Before the data engineer deletes the cluster each month, the data engineer unloads backup data from the cluster to an Amazon S3 bucket.
The data engineer needs a solution to run the monthly analytics processes that does not require the data engineer to manage the infrastructure manually.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Use Amazon Step Functions to pause the Redshift cluster when the analytics processes are complete and to resume the cluster to run new processes every month.
B.Use Amazon Redshift Serverless to automatically process the analytics workload.
C.Use the AWS CLI to automatically process the analytics workload.
D.Use AWS CloudFormation templates to automatically process the analytics workload.

Answer: B

Explanation:
Redshift Serverless eliminates infrastructure management. It automatically provisions and scales compute to run workloads only when needed, ideal for infrequent monthly analytics.

- A is incorrect because pausing/resuming provisioned clusters still requires management.
- C is incorrect because CLI automation doesn't remove infrastructure overhead.
- D is incorrect because CloudFormation provisions clusters but doesn't avoid managing them.

Question: 113
A company receives a daily file that contains customer data in .xls format. The company stores the file in Amazon S3. The daily file is approximately 2 GB in size.
A data engineer concatenates the column in the file that contains customer first names and the column that contains customer last names. The data engineer needs to determine the number of distinct customers in the file.
Which solution will meet this requirement with the LEAST operational effort?

Options
A.Create and run an Apache Spark job in an AWS Glue notebook. Configure the job to read the S3 file and calculate the number of distinct customers.
B.Create an AWS Glue crawler to create an AWS Glue Data Catalog of the S3 file. Run SQL queries from Amazon Athena to calculate the number of distinct customers.
C.Create and run an Apache Spark job in Amazon EMR Serverless to calculate the number of distinct customers.
D.Use AWS Glue DataBrew to create a recipe that uses the COUNT_DISTINCT aggregate function to calculate the number of distinct customers.

Answer: D

Explanation:
AWS Glue DataBrew provides a no-code option with built-in aggregate functions like COUNT_DISTINCT, reducing coding effort.

- A is incorrect because Spark jobs in Glue notebooks require custom coding.
- B is incorrect because Athena doesn't natively support .xls format.
- C is incorrect because EMR Serverless with Spark introduces higher setup complexity.

Question: 114
A healthcare company uses Amazon Kinesis Data Streams to stream real-time health data from wearable devices, hospital equipment, and patient records.
A data engineer needs to find a solution to process the streaming data. The data engineer needs to store the data in an Amazon Redshift Serverless warehouse. The solution must support near real-time analytics of the streaming data and the previous day's data.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Load data into Amazon Kinesis Data Firehose. Load the data into Amazon Redshift.
B.Use the streaming ingestion feature of Amazon Redshift.
C.Load the data into Amazon S3. Use the COPY command to load the data into Amazon Redshift.
D.Use the Amazon Aurora zero-ETL integration with Amazon Redshift.

Answer: B

Explanation:
Redshift streaming ingestion allows direct ingestion from Kinesis Data Streams into Redshift Serverless with near real-time performance and minimal management.

- A is incorrect because Firehose adds an extra component with more latency.
- C is incorrect because COPY from S3 introduces batch loading, not real-time.
- D is incorrect because Aurora zero-ETL applies to Aurora sources, not Kinesis.

Question: 115
A data engineer needs to use an Amazon QuickSight dashboard that is based on Amazon Athena queries on data that is stored in an Amazon S3 bucket. When the data engineer connects to the QuickSight dashboard, the data engineer receives an error message that indicates insufficient permissions.
Which factors could cause to the permissions-related errors? (Choose two.)

Options
A.There is no connection between QuickSight and Athena.
B.The Athena tables are not cataloged.
C.QuickSight does not have access to the S3 bucket.
D.QuickSight does not have access to decrypt S3 data.
E.There is no IAM role assigned to QuickSight.

Answer: CD

Explanation:
QuickSight requires explicit permissions to access S3 buckets where Athena data resides, and if data is encrypted, it needs permission to use the KMS key for decryption.

- A is incorrect because QuickSight connects to Athena natively if permissions are set.
- B is incorrect because uncataloged tables would cause query errors, not permissions errors.
- E is incorrect because QuickSight uses service-linked roles, not manual IAM roles.

Question: 116
A company stores datasets in JSON format and .csv format in an Amazon S3 bucket. The company has Amazon RDS for Microsoft SQL Server databases, Amazon DynamoDB tables that are in provisioned capacity mode, and an Amazon Redshift cluster. A data engineering team must develop a solution that will give data scientists the ability to query all data sources by using syntax similar to SQL.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Use AWS Glue to crawl the data sources. Store metadata in the AWS Glue Data Catalog. Use Amazon Athena to query the data. Use SQL for structured data sources. Use PartiQL for data that is stored in JSON format.
B.Use AWS Glue to crawl the data sources. Store metadata in the AWS Glue Data Catalog. Use Redshift Spectrum to query the data. Use SQL for structured data sources. Use PartiQL for data that is stored in JSON format.
C.Use AWS Glue to crawl the data sources. Store metadata in the AWS Glue Data Catalog. Use AWS Glue jobs to transform data that is in JSON format to Apache Parquet or .csv format. Store the transformed data in an S3 bucket. Use Amazon Athena to query the original and transformed data from the S3 bucket.
D.Use AWS Lake Formation to create a data lake. Use Lake Formation jobs to transform the data from all data sources to Apache Parquet format. Store the transformed data in an S3 bucket. Use Amazon Athena or Redshift Spectrum to query the data.

Answer: A

Explanation:
Athena with Glue Data Catalog allows SQL-like queries across multiple data sources. PartiQL support enables querying semi-structured JSON with familiar syntax.

- B is incorrect because Redshift Spectrum is less flexible for multi-source queries.
- C is incorrect because transformation is unnecessary; Athena + PartiQL can query JSON directly.
- D is incorrect because Lake Formation adds extra setup not required.

Question: 117
A data engineer is configuring Amazon SageMaker Studio to use AWS Glue interactive sessions to prepare data for machine learning (ML) models.
The data engineer receives an access denied error when the data engineer tries to prepare the data by using SageMaker Studio.
Which change should the engineer make to gain access to SageMaker Studio?

Options
A.Add the AWSGlueServiceRole managed policy to the data engineer's IAM user.
B.Add a policy to the data engineer's IAM user that includes the sts:AssumeRole action for the AWS Glue and SageMaker service principals in the trust policy.
C.Add the AmazonSageMakerFullAccess managed policy to the data engineer's IAM user.
D.Add a policy to the data engineer's IAM user that allows the sts:AddAssociation action for the AWS Glue and SageMaker service principals in the trust policy.

Answer: B

Explanation:
The SageMaker Studio user must assume the role with permissions to use Glue interactive sessions. This requires configuring the trust policy to allow sts:AssumeRole for Glue and SageMaker service principals.

- A is incorrect because AWSGlueServiceRole is meant for Glue jobs, not Studio users.
- C is incorrect because SageMakerFullAccess does not cover Glue interactive session permissions.
- D is incorrect because sts:AddAssociation is not the correct action.

Question: 118
A company extracts approximately 1 TB of data every day from data sources such as SAP HANA, Microsoft SQL Server, MongoDB, Apache Kafka, and Amazon DynamoDB. Some of the data sources have undefined data schemas or data schemas that change.
A data engineer must implement a solution that can detect the schema for these data sources. The solution must extract, transform, and load the data to an Amazon S3 bucket. The company has a service level agreement (SLA) to load the data into the S3 bucket within 15 minutes of data creation.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Use Amazon EMR to detect the schema and to extract, transform, and load the data into the S3 bucket. Create a pipeline in Apache Spark.
B.Use AWS Glue to detect the schema and to extract, transform, and load the data into the S3 bucket. Create a pipeline in Apache Spark.
C.Create a PySpark program in AWS Lambda to extract, transform, and load the data into the S3 bucket.
D.Create a stored procedure in Amazon Redshift to detect the schema and to extract, transform, and load the data into a Redshift Spectrum table. Access the table from Amazon S3.

Answer: B

Explanation:
AWS Glue automatically detects schema changes and supports Spark ETL pipelines. It can ingest data from heterogeneous sources with minimal management and within SLA requirements.

- A is incorrect because EMR requires cluster setup and more management.
- C is incorrect because Lambda is not suitable for large-scale ETL (1 TB daily).
- D is incorrect because Redshift procedures are not designed for schema detection or multi-source ETL.

Question: 119
A company has multiple applications that use datasets that are stored in an Amazon S3 bucket. The company has an ecommerce application that generates a dataset that contains personally identifiable information (PII). The company has an internal analytics application that does not require access to the PII.
To comply with regulations, the company must not share PII unnecessarily. A data engineer needs to implement a solution that will redact PII dynamically, based on the needs of each application that accesses the dataset.
Which solution will meet the requirements with the LEAST operational overhead?

Options
A.Create an S3 bucket policy to limit the access each application has. Create multiple copies of the dataset. Give each dataset copy the appropriate level of redaction for the needs of the application that accesses the copy.
B.Create an S3 Object Lambda endpoint. Use the S3 Object Lambda endpoint to read data from the S3 bucket. Implement redaction logic within an S3 Object Lambda function to dynamically redact PII based on the needs of each application that accesses the data.
C.Use AWS Glue to transform the data for each application. Create multiple copies of the dataset. Give each dataset copy the appropriate level of redaction for the needs of the application that accesses the copy.
D.Create an API Gateway endpoint that has custom authorizers. Use the API Gateway endpoint to read data from the S3 bucket. Initiate a REST API call to dynamically redact PII based on the needs of each application that accesses the data.

Answer: B

Explanation:
S3 Object Lambda allows dynamic transformation/redaction of objects as they are accessed, eliminating the need for multiple dataset copies.

- A is incorrect because it requires creating and maintaining multiple datasets.
- C is incorrect because Glue transformations create redundant copies and add overhead.
- D is incorrect because API Gateway adds unnecessary infrastructure and complexity.

Question: 120
A data engineer needs to build an extract, transform, and load (ETL) job. The ETL job will process daily incoming .csv files that users upload to an Amazon S3 bucket. The size of each S3 object is less than 100 MB.
Which solution will meet these requirements MOST cost-effectively?

Options
A.Write a custom Python application. Host the application on an Amazon Elastic Kubernetes Service (Amazon EKS) cluster.
B.Write a PySpark ETL script. Host the script on an Amazon EMR cluster.
C.Write an AWS Glue PySpark job. Use Apache Spark to transform the data.
D.Write an AWS Glue Python shell job. Use pandas to transform the data.

Answer: C

Explanation:
AWS Glue PySpark jobs are cost-effective for small-to-medium ETL workloads. They provide managed Spark without cluster management and handle schema evolution automatically.

- A is incorrect because EKS requires significant setup and management.
- B is incorrect because EMR clusters are overkill for <100 MB file sizes.
- D is incorrect because Python shell jobs with pandas are less scalable for ongoing ETL pipelines.

Question: 121
A marketing company uses Amazon S3 to store clickstream data. The company queries the data at the end of each day by using a SQL JOIN clause on S3 objects that are stored in separate buckets. The company creates key performance indicators (KPIs) based on the objects. The company needs a serverless solution that will give users the ability to query data by partitioning the data. The solution must maintain the atomicity, consistency, isolation, and durability (ACID) properties of the data. Which solution will meet these requirements MOST cost-effectively?

Options
A.Amazon S3 Select
B.Amazon Redshift Spectrum
C.Amazon Athena
D.Amazon EMR

Answer: C

Explanation:
Athena is serverless, supports partitioning in the AWS Glue Data Catalog, and integrates with ACID transaction tables when paired with AWS Lake Formation and Apache Iceberg. It is cost-effective because you pay only per query.

- A: S3 Select works at the object level and does not support JOINs across datasets or ACID transactions.
- B: Redshift Spectrum queries S3 but requires provisioning and managing a Redshift cluster, increasing cost.
- D: EMR can handle this but requires cluster management and incurs higher costs than serverless Athena.

Question: 122
A company wants to migrate data from an Amazon RDS for PostgreSQL DB instance in the eu-east-1 Region of an AWS account named Account_A. The company will migrate the data to an Amazon Redshift cluster in the eu-west-1 Region of an AWS account named Account_B. Which solution will give AWS Database Migration Service (AWS DMS) the ability to replicate data between two data stores?

Options
A.Set up an AWS DMS replication instance in Account_B in eu-west-1.
B.Set up an AWS DMS replication instance in Account_B in eu-east-1.
C.Set up an AWS DMS replication instance in a new AWS account in eu-west-1.
D.Set up an AWS DMS replication instance in Account_A in eu-east-1.

Answer: A

Explanation:
DMS replication instances must be in the same Region as the target (Amazon Redshift in eu-west-1). Placing it in Account_B ensures network access to both source and target.

- B: If in eu-east-1, it can reach the source but not the target in eu-west-1 efficiently.
- C: Creating a new AWS account adds unnecessary complexity.
- D: Same as B, in source Region, not target Region.

Question: 123
A company uses Amazon S3 as a data lake. The company sets up a data warehouse by using a multi-node Amazon Redshift cluster. The company organizes the data files in the data lake based on the data source of each data file. The company loads all the data files into one table in the Redshift cluster by using a separate COPY command for each data file location. This approach takes a long time to load all the data files into the table. The company must increase the speed of the data ingestion. The company does not want to increase the cost of the process. Which solution will meet these requirements?

Options
A.Use a provisioned Amazon EMR cluster to copy all the data files into one folder. Use a COPY command to load the data into Amazon Redshift.
B.Load all the data files in parallel into Amazon Aurora. Run an AWS Glue job to load the data into Amazon Redshift.
C.Use an AWS Glue job to copy all the data files into one folder. Use a COPY command to load the data into Amazon Redshift.
D.Create a manifest file that contains the data file locations. Use a COPY command to load the data into Amazon Redshift.

Answer: D

Explanation:
A manifest file lists multiple S3 file paths so Redshift can load them in parallel with one COPY command. This improves speed without additional cost.

- A: EMR adds cost and operational overhead.
- B: Loading into Aurora first is unnecessary and expensive.
- C: Glue jobs introduce extra cost; not needed if manifest can handle parallel ingestion.

Question: 124
A company plans to use Amazon Kinesis Data Firehose to store data in Amazon S3. The source data consists of 2 MB .csv files. The company must convert the .csv files to JSON format. The company must store the files in Apache Parquet format. Which solution will meet these requirements with the LEAST development effort?

Options
A.Use Kinesis Data Firehose to convert the .csv files to JSON. Use an AWS Lambda function to store the files in Parquet format.
B.Use Kinesis Data Firehose to convert the .csv files to JSON and to store the files in Parquet format.
C.Use Kinesis Data Firehose to invoke an AWS Lambda function that transforms the .csv files to JSON and stores the files in Parquet format.
D.Use Kinesis Data Firehose to invoke an AWS Lambda function that transforms the .csv files to JSON. Use Kinesis Data Firehose to store the files in Parquet format.

Answer: D

Explanation:
Firehose supports Parquet output natively. The Lambda function can handle CSV → JSON, and then Firehose itself handles Parquet conversion, minimizing development work.

- A: Requires custom Lambda to handle Parquet, increasing development effort.
- B: Firehose cannot directly convert CSV → JSON before Parquet.
- C: Offloads both JSON and Parquet conversion to Lambda, requiring more complex code.

Question: 125
A company is using an AWS Transfer Family server to migrate data from an on-premises environment to AWS. Company policy mandates the use of TLS 1.2 or above to encrypt the data in transit. Which solution will meet these requirements?

Options
A.Generate new SSH keys for the Transfer Family server. Make the old keys and the new keys available for use.
B.Update the security group rules for the on-premises network to allow only connections that use TLS 1.2 or above.
C.Update the security policy of the Transfer Family server to specify a minimum protocol version of TLS 1.2
D.Install an SSL certificate on the Transfer Family server to encrypt data transfers by using TLS 1.2.

Answer: C

Explanation:
Transfer Family supports security policies that specify minimum TLS versions. Setting the minimum to TLS 1.2 enforces encryption requirements.

- A: SSH keys do not enforce TLS versions.
- B: Security groups control IP/port access, not TLS protocol versions.
- D: SSL certificates secure communication but do not enforce protocol version.

Question: 126
A company wants to migrate an application and an on-premises Apache Kafka server to AWS. The application processes incremental updates that an on-premises Oracle database sends to the Kafka server. The company wants to use the replatform migration strategy instead of the refactor strategy. Which solution will meet these requirements with the LEAST management overhead?

Options
A.Amazon Kinesis Data Streams
B.Amazon Managed Streaming for Apache Kafka (Amazon MSK) provisioned cluster
C.Amazon Kinesis Data Firehose
D.Amazon Managed Streaming for Apache Kafka (Amazon MSK) Serverless

Answer: D

Explanation:
MSK Serverless is a fully managed Kafka service, eliminating the need to manage brokers or clusters. It matches the replatforming approach (lift-and-shift with managed infra).

- A: Kinesis Data Streams is a refactor, as it requires changing from Kafka APIs.
- B: MSK provisioned cluster requires managing capacity and scaling.
- C: Firehose is for delivery, not for full Kafka migration compatibility.

Question: 127
A data engineer is building an automated extract, transform, and load (ETL) ingestion pipeline by using AWS Glue. The pipeline ingests compressed files that are in an Amazon S3 bucket. The ingestion pipeline must support incremental data processing. Which AWS Glue feature should the data engineer use to meet this requirement?

Options
A.Workflows
B.Triggers
C.Job bookmarks
D.Classifiers

Answer: C

Explanation:
Job bookmarks track processed files and prevent reprocessing, enabling incremental data ingestion in Glue.

- A: Workflows orchestrate multiple jobs but don’t manage incremental state.
- B: Triggers schedule or chain jobs, not incremental processing.
- D: Classifiers help detect schema, not track incremental progress.

Question: 128
A banking company uses an application to collect large volumes of transactional data. The company uses Amazon Kinesis Data Streams for real-time analytics. The company's application uses the PutRecord action to send data to Kinesis Data Streams. A data engineer has observed network outages during certain times of day. The data engineer wants to configure exactly-once delivery for the entire processing pipeline. Which solution will meet this requirement?

Options
A.Design the application so it can remove duplicates during processing by embedding a unique ID in each record at the source.
B.Update the checkpoint configuration of the Amazon Managed Service for Apache Flink (previously known as Amazon Kinesis Data Analytics) data collection application to avoid duplicate processing of events.
C.Design the data source so events are not ingested into Kinesis Data Streams multiple times.
D.Stop using Kinesis Data Streams. Use Amazon EMR instead. Use Apache Flink and Apache Spark Streaming in Amazon EMR.

Answer: A

Explanation:
Exactly-once delivery in Kinesis relies on de-duplication at the consumer side. Embedding unique IDs ensures that duplicates caused by retries during outages can be removed.

- B: Checkpointing ensures at-least-once processing, not exactly-once without deduplication.
- C: Preventing duplicate ingestion is unrealistic under retries; unique IDs are required.
- D: Switching to EMR introduces unnecessary complexity and doesn't guarantee exactly-once.

Question: 129
A company stores logs in an Amazon S3 bucket. When a data engineer attempts to access several log files, the data engineer discovers that some files have been unintentionally deleted. The data engineer needs a solution that will prevent unintentional file deletion in the future. Which solution will meet this requirement with the LEAST operational overhead?

Options
A.Manually back up the S3 bucket on a regular basis.
B.Enable S3 Versioning for the S3 bucket.
C.Configure replication for the S3 bucket.
D.Use an Amazon S3 Glacier storage class to archive the data that is in the S3 bucket.

Answer: B

Explanation:
S3 Versioning keeps older versions of files, so even if a file is deleted, it can be restored. This requires no manual work.

- A: Manual backups add operational burden.
- C: Replication creates a copy but accidental deletion can still replicate.
- D: Glacier archives data but doesn't prevent deletion of active logs.

Question: 130
A telecommunications company collects network usage data throughout each day at a rate of several thousand data points each second. The company runs an application to process the usage data in real time. The company aggregates and stores the data in an Amazon Aurora DB instance. Sudden drops in network usage usually indicate a network outage. The company must be able to identify sudden drops in network usage so the company can take immediate remedial actions. Which solution will meet this requirement with the LEAST latency?

Options
A.Create an AWS Lambda function to query Aurora for drops in network usage. Use Amazon EventBridge to automatically invoke the Lambda function every minute.
B.Modify the processing application to publish the data to an Amazon Kinesis data stream. Create an Amazon Managed Service for Apache Flink (previously known as Amazon Kinesis Data Analytics) application to detect drops in network usage.
C.Replace the Aurora database with an Amazon DynamoDB table. Create an AWS Lambda function to query the DynamoDB table for drops in network usage every minute. Use DynamoDB Accelerator (DAX) between the processing application and DynamoDB table.
D.Create an AWS Lambda function within the Database Activity Streams feature of Aurora to detect drops in network usage.

Answer: B

Explanation:
Publishing data into Kinesis and using Apache Flink enables continuous streaming analytics with millisecond-level latency, ideal for real-time anomaly detection.

- A: Invoking Lambda every minute introduces latency and misses near-real-time detection.
- C: DynamoDB + Lambda polling still introduces latency and is unnecessary.
- D: Aurora Database Activity Streams is for auditing/logging, not real-time anomaly detection.

Question: 131
A data engineer must ingest a source of structured data that is in .csv format into an Amazon S3 data lake. The .csv files contain 15 columns. Data analysts need to run Amazon Athena queries on one or two columns of the dataset. The data analysts rarely query the entire file.
Which solution will meet these requirements MOST cost-effectively?

Options
A.Use an AWS Glue PySpark job to ingest the source data into the data lake in .csv format.
B.Create an AWS Glue extract, transform, and load (ETL) job to read from the .csv structured data source.
Configure the job to ingest the data into the data lake in JSON format.
C.Use an AWS Glue PySpark job to ingest the source data into the data lake in Apache Avro format.
D.Create an AWS Glue extract, transform, and load (ETL) job to read from the .csv structured data source.
Configure the job to write the data into the data lake in Apache Parquet format.

Answer: D

Explanation:
Parquet is a columnar format that enables column pruning and efficient compression, so Athena reads far less data when analysts query one or two columns, reducing both latency and cost.

- A is incorrect because CSV is row-based; Athena must scan entire rows even if only a few columns are needed.
- B is incorrect because JSON is verbose and row-oriented, increasing scan size and cost.
- C is incorrect because Avro is row-based; it doesn't provide the same column-pruning benefits as Parquet.

Question: 132
A company has five offices in different AWS Regions. Each office has its own human resources (HR) department that uses a unique IAM role. The company stores employee records in a data lake that is based on Amazon S3 storage.
A data engineering team needs to limit access to the records. Each HR department should be able to access records for only employees who are within the HR department's Region.
Which combination of steps should the data engineering team take to meet this requirement with the LEAST operational overhead? (Choose two.)

Options
A.Use data filters for each Region to register the S3 paths as data locations.
B.Register the S3 path as an AWS Lake Formation location.
C.Modify the IAM roles of the HR departments to add a data filter for each department's Region.
D.Enable fine-grained access control in AWS Lake Formation. Add a data filter for each Region.
E.Create a separate S3 bucket for each Region. Configure an IAM policy to allow S3 access. Restrict access
based on Region.

Answer: BD

Explanation:
Registering the S3 path in Lake Formation and enabling Lake Formation fine-grained access control with regional data filters lets you enforce row/partition-level access centrally with minimal ops.

- A is incorrect because data filters apply within Lake Formation governance, not during location registration alone.
- C is incorrect because IAM alone doesn't provide Lake Formation's fine-grained table/partition-level filtering.
- E is incorrect because creating separate buckets increases operational overhead and doesn't provide fine-grained controls.

Question: 133
A company uses AWS Step Functions to orchestrate a data pipeline. The pipeline consists of Amazon EMR jobs that ingest data from data sources and store the data in an Amazon S3 bucket. The pipeline also includes EMR jobs that load the data to Amazon Redshift.
The company's cloud infrastructure team manually built a Step Functions state machine. The cloud infrastructure team launched an EMR cluster into a VPC to support the EMR jobs. However, the deployed Step Functions state machine is not able to run the EMR jobs.
Which combination of steps should the company take to identify the reason the Step Functions state machine is not able to run the EMR jobs? (Choose two.)

Options
A.Use AWS CloudFormation to automate the Step Functions state machine deployment. Create a step to pause
the state machine during the EMR jobs that fail. Configure the step to wait for a human user to send approval
through an email message. Include details of the EMR task in the email message for further analysis.
B.Verify that the Step Functions state machine code has all IAM permissions that are necessary to create and
run the EMR jobs. Verify that the Step Functions state machine code also includes IAM permissions to access
the Amazon S3 buckets that the EMR jobs use. Use Access Analyzer for S3 to check the S3 access properties.
C.Check for entries in Amazon CloudWatch for the newly created EMR cluster. Change the AWS Step Functions
state machine code to use Amazon EMR on EKS. Change the IAM access policies and the security group
configuration for the Step Functions state machine code to reflect inclusion of Amazon Elastic Kubernetes
Service (Amazon EKS).
D.Query the flow logs for the VPC. Determine whether the traffic that originates from the EMR cluster can
successfully reach the data providers. Determine whether any security group that might be attached to the
Amazon EMR cluster allows connections to the data source servers on the informed ports.
E.Check the retry scenarios that the company configured for the EMR jobs. Increase the number of seconds in
the interval between each EMR task. Validate that each fallback state has the appropriate catch for each
decision state. Configure an Amazon Simple Notification Service (Amazon SNS) topic to store the error
messages.

Answer: BD

Explanation:
First validate IAM permissions for Step Functions to start/monitor EMR and access S3, then check VPC flow logs and security groups to verify network reachability from EMR to data sources.

- A is incorrect because adding manual pauses/emails doesn't diagnose IAM/VPC causes.
- C is incorrect because switching to EMR on EKS is unrelated and adds complexity.
- E is incorrect because tuning retries doesn't fix permission or networking failures.

Question: 134
A company is developing an application that runs on Amazon EC2 instances. Currently, the data that the application generates is temporary. However, the company needs to persist the data, even if the EC2 instances are terminated.
A data engineer must launch new EC2 instances from an Amazon Machine Image (AMI) and configure the instances to preserve the data.
Which solution will meet this requirement?

Options
A.Launch new EC2 instances by using an AMI that is backed by an EC2 instance store volume that contains the
application data. Apply the default settings to the EC2 instances.
B.Launch new EC2 instances by using an AMI that is backed by a root Amazon Elastic Block Store (Amazon
EBS) volume that contains the application data. Apply the default settings to the EC2 instances.
C.Launch new EC2 instances by using an AMI that is backed by an EC2 instance store volume. Attach an
Amazon Elastic Block Store (Amazon EBS) volume to contain the application data. Apply the default settings to
the EC2 instances.
D.Launch new EC2 instances by using an AMI that is backed by an Amazon Elastic Block Store (Amazon EBS)
volume. Attach an additional EC2 instance store volume to contain the application data. Apply the default
settings to the EC2 instances.

Answer: C

Explanation:
EBS volumes are persistent across instance stops/terminations (when preserved). Using an additional EBS volume for application data ensures durability, while the AMI can be instance-store–backed if desired.

- A is incorrect because instance store is ephemeral and data is lost on termination.
- B is incorrect because coupling data to the root EBS volume is risky; separate data volume is standard.
- D is incorrect because instance store is ephemeral; adding it won't preserve data.

Question: 135
A company uses Amazon Athena to run SQL queries for extract, transform, and load (ETL) tasks by using Create Table As Select (CTAS). The company must use Apache Spark instead of SQL to generate analytics.
Which solution will give the company the ability to use Spark to access Athena?

Options
A.Athena query settings
B.Athena workgroup
C.Athena data source
D.Athena query editor

Answer: B

Explanation:
Athena for Spark is enabled and managed via Athena workgroups, allowing Spark-based analytics that read from the same data catalog and S3 data.

- A is incorrect because query settings alone don't enable Spark sessions.
- C is incorrect because data sources map connectors, not Spark enablement.
- D is incorrect because the editor is a UI and doesn't control Spark capability.

Question: 136
A company needs to partition the Amazon S3 storage that the company uses for a data lake. The partitioning will use a path of the S3 object keys in the following format: s3://bucket/prefix/year=2023/month=01/day=01.
A data engineer must ensure that the AWS Glue Data Catalog synchronizes with the S3 storage when the company adds new partitions to the bucket.
Which solution will meet these requirements with the LEAST latency?

Options
A.Schedule an AWS Glue crawler to run every morning.
B.Manually run the AWS Glue CreatePartition API twice each day.
C.Use code that writes data to Amazon S3 to invoke the Boto3 AWS Glue create_partition API call.
D.Run the MSCK REPAIR TABLE command from the AWS Glue console.

Answer: C

Explanation:
Publishing partitions as part of the write path with create_partition updates the catalog immediately, minimizing latency versus periodic crawls or manual repairs.

- A is incorrect because scheduled crawlers introduce delay.
- B is incorrect due to manual effort and latency.
- D is incorrect because MSCK REPAIR is a batch metadata sync, not low-latency.

Question: 137
A media company uses software as a service (SaaS) applications to gather data by using third-party tools. The company needs to store the data in an Amazon S3 bucket. The company will use Amazon Redshift to perform analytics based on the data.
Which AWS service or feature will meet these requirements with the LEAST operational overhead?

Options
A.Amazon Managed Streaming for Apache Kafka (Amazon MSK)
B.Amazon AppFlow
C.AWS Glue Data Catalog
D.Amazon Kinesis

Answer: B

Explanation:
AppFlow provides managed, low-code integrations from many SaaS apps directly into S3, with scheduling and transformations, minimizing ops.

- A is incorrect because MSK requires cluster management and custom connectors.
- C is incorrect because the Data Catalog is metadata, not ingestion.
- D is incorrect because Kinesis would require custom producers from each SaaS.

Question: 138
A data engineer is using Amazon Athena to analyze sales data that is in Amazon S3. The data engineer writes a query to retrieve sales amounts for 2023 for several products from a table named sales_data. However, the query does not return results for all of the products that are in the sales_data table. The data engineer needs to troubleshoot the query to resolve the issue.
The data engineer's original query is as follows:
SELECT product_name, sum(sales_amount)
FROM sales_data -
WHERE year = 2023 -
GROUP BY product_name -
How should the data engineer modify the Athena query to meet these requirements?

Options
A.Replace sum(sales_amount) with count(*) for the aggregation.
B.Change WHERE year = 2023 to WHERE extract(year FROM sales_data) = 2023.
C.Add HAVING sum(sales_amount) > 0 after the GROUP BY clause.
D.Remove the GROUP BY clause.

Answer: B

Explanation:
Using extract(year FROM ...) against the correct timestamp column (rather than a partition or mismatched column) ensures filtering by the actual year value; the provided option indicates using extract(year FROM sales_data) = 2023.

- A is incorrect because changing the aggregation doesn’t fix the filtering issue.
- C is incorrect because adding HAVING filters results, not the underlying cause.
- D is incorrect because removing GROUP BY changes aggregation semantics.

Question: 139
A data engineer has a one-time task to read data from objects that are in Apache Parquet format in an Amazon S3 bucket. The data engineer needs to query only one column of the data. Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Configure an AWS Lambda function to load data from the S3 bucket into a pandas dataframe. Write a SQL
SELECT statement on the dataframe to query the required column.
B.Use S3 Select to write a SQL SELECT statement to retrieve the required column from the S3 objects.
C.Prepare an AWS Glue DataBrew project to consume the S3 objects and to query the required column.
D.Run an AWS Glue crawler on the S3 objects. Use a SQL SELECT statement in Amazon Athena to query the
required column.

Answer: B

Explanation:
S3 Select allows querying a subset (columns) from supported formats directly from S3 with minimal setup—ideal for a one-off column retrieval.

- A is incorrect because building and running Lambda/pandas adds unnecessary code and limits data size.
- C is incorrect because DataBrew setup is heavier for a one-time query.
- D is incorrect because crawling and Athena setup is overkill for a simple one-off read.

Question: 140
A company uses Amazon Redshift for its data warehouse. The company must automate refresh schedules for Amazon Redshift materialized views.
Which solution will meet this requirement with the LEAST effort?

Options
A.Use Apache Airflow to refresh the materialized views.
B.Use an AWS Lambda user-defined function (UDF) within Amazon Redshift to refresh the materialized views.
C.Use the query editor v2 in Amazon Redshift to refresh the materialized views.
D.Use an AWS Glue workflow to refresh the materialized views.

Answer: C

Explanation:
Query editor v2 supports creating scheduled queries (including REFRESH MATERIALIZED VIEW), providing a simple, native way to automate refreshes with minimal effort.

- A is incorrect because Airflow adds orchestration infrastructure overhead.
- B is incorrect because Lambda UDFs don’t schedule tasks; scheduling still needed.
- D is incorrect because Glue workflows are unnecessary for simple SQL refresh scheduling.

Question: 141
A data engineer creates an AWS Glue Data Catalog table by using an AWS Glue crawler that is named Orders. The
data engineer wants to add the following new partitions:
s3://transactions/orders/order_date=2023-01-01
s3://transactions/orders/order_date=2023-01-02
The data engineer must edit the metadata to include the new partitions in the table without scanning all the
folders and files in the location of the table.
Which data definition language (DDL) statement should the data engineer use in Amazon Athena?

Options
A.ALTER TABLE Orders ADD PARTITION(order_date='2023-01-01') LOCATION 's3://transactions/orders/order_date=2023-01-01'; ALTER TABLE Orders ADD PARTITION(order_date='2023-01-02') LOCATION 's3://transactions/orders/order_date=2023-01-02';
B.MSCK REPAIR TABLE Orders;
C.REPAIR TABLE Orders;
D.ALTER TABLE Orders MODIFY PARTITION(order_date='2023-01-01') LOCATION 's3://transactions/orders/2023-01-01';
ALTER TABLE Orders MODIFY PARTITION(order_date='2023-01-02') LOCATION 's3://transactions/orders/2023-01-02';

Answer: A

Explanation:
The ALTER TABLE ... ADD PARTITION command directly adds new partitions to the Glue Data Catalog metadata without rescanning the entire dataset. This is efficient when exact partition locations are already known.

- B is incorrect because MSCK REPAIR TABLE rescans all paths, which is what the engineer wants to avoid.
- C is invalid since REPAIR TABLE is not supported in Athena.
- D is incorrect because MODIFY PARTITION updates locations for existing partitions, not new ones.

Question: 142
A company stores 10 to 15 TB of uncompressed .csv files in Amazon S3. The company is evaluating Amazon Athena
as a one-time query engine.
The company wants to transform the data to optimize query runtime and storage costs.
Which file format and compression solution will meet these requirements for Athena queries?

Options
A..csv format compressed with zip
B.JSON format compressed with bzip2
C.Apache Parquet format compressed with Snappy
D.Apache Avro format compressed with LZO

Answer: C

Explanation:
Parquet is a columnar storage format optimized for analytical queries. Combined with Snappy compression, it reduces storage size and speeds up queries significantly.

- A is incorrect because compressed CSV does not improve columnar query performance.
- B is incorrect because JSON with bzip2 is inefficient for large-scale analytics.
- D is incorrect because Avro is row-based, less efficient for Athena queries compared to Parquet.

Question: 143
A company uses Apache Airflow to orchestrate the company's current on-premises data pipelines. The company
runs SQL data quality check tasks as part of the pipelines. The company wants to migrate the pipelines to AWS
and to use AWS managed services.
Which solution will meet these requirements with the LEAST amount of refactoring?

Options
A.Setup AWS Outposts in the AWS Region that is nearest to the location where the company uses Airflow.
Migrate the servers into Outposts hosted Amazon EC2 instances. Update the pipelines to interact with the
Outposts hosted EC2 instances instead of the on-premises pipelines.
B.Create a custom Amazon Machine Image (AMI) that contains the Airflow application and the code that the
company needs to migrate. Use the custom AMI to deploy Amazon EC2 instances. Update the network
connections to interact with the newly deployed EC2 instances.
C.Migrate the existing Airflow orchestration configuration into Amazon Managed Workflows for Apache Airflow
(Amazon MWAA). Create the data quality checks during the ingestion to validate the data quality by using SQL
tasks in Airflow.
D.Convert the pipelines to AWS Step Functions workflows. Recreate the data quality checks in SQL as Python
based AWS Lambda functions.

Answer: C

Explanation:
Amazon MWAA is a managed service for Airflow, allowing direct migration of existing Airflow DAGs with minimal refactoring.

- A is incorrect because Outposts is expensive and unnecessary for managed Airflow migration.
- B is incorrect because deploying EC2 with a custom AMI still leaves infrastructure management overhead.
- D is incorrect because rewriting into Step Functions + Lambda requires significant refactoring.

Question: 144
A company uses Amazon EMR as an extract, transform, and load (ETL) pipeline to transform data that comes from
multiple sources. A data engineer must orchestrate the pipeline to maximize performance. Which AWS service will meet this requirement MOST cost effectively?

Options
A.Amazon EventBridge
B.Amazon Managed Workflows for Apache Airflow (Amazon MWAA)
C.AWS Step Functions
D.AWS Glue Workflows

Answer: C

Explanation:
AWS Step Functions provides serverless orchestration with pay-per-use pricing, ideal for orchestrating EMR jobs efficiently without additional infrastructure.

- A is incorrect because EventBridge is an event bus, not a full orchestrator.
- B is incorrect because MWAA adds management cost compared to Step Functions.
- D is incorrect because Glue Workflows are designed for Glue jobs, not EMR.

Question: 145
An online retail company stores Application Load Balancer (ALB) access logs in an Amazon S3 bucket. The
company wants to use Amazon Athena to query the logs to analyze traffic patterns.
A data engineer creates an unpartitioned table in Athena. As the amount of the data gradually increases, the
response time for queries also increases. The data engineer wants to improve the query performance in Athena.
Which solution will meet these requirements with the LEAST operational effort?

Options
A.Create an AWS Glue job that determines the schema of all ALB access logs and writes the partition metadata
to AWS Glue Data Catalog.
B.Create an AWS Glue crawler that includes a classifier that determines the schema of all ALB access logs and
writes the partition metadata to AWS Glue Data Catalog.
C.Create an AWS Lambda function to transform all ALB access logs. Save the results to Amazon S3 in Apache
Parquet format. Partition the metadata. Use Athena to query the transformed data.
D.Use Apache Hive to create bucketed tables. Use an AWS Lambda function to transform all ALB access logs.

Answer: B

Explanation:
Using an AWS Glue crawler automates schema detection and partition management in the Data Catalog, which improves Athena query performance without manual work.

- A is incorrect because Glue jobs require more setup than a crawler.
- C is correct technically but requires building a transformation pipeline, which is more effort.
- D is incorrect because Hive and custom Lambda transformation increase complexity.

Question: 146
A company has a business intelligence platform on AWS. The company uses an AWS Storage Gateway Amazon S3
File Gateway to transfer files from the company's on-premises environment to an Amazon S3 bucket.
A data engineer needs to setup a process that will automatically launch an AWS Glue workflow to run a series of
AWS Glue jobs when each file transfer finishes successfully.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Determine when the file transfers usually finish based on previous successful file transfers. Set up an
Amazon EventBridge scheduled event to initiate the AWS Glue jobs at that time of day.
B.Set up an Amazon EventBridge event that initiates the AWS Glue workflow after every successful S3 File
Gateway file transfer event.
C.Set up an on-demand AWS Glue workflow so that the data engineer can start the AWS Glue workflow when
each file transfer is complete.
D.Set up an AWS Lambda function that will invoke the AWS Glue Workflow. Set up an event for the creation of
an S3 object as a trigger for the Lambda function.

Answer: B

Explanation:
EventBridge can directly capture S3 File Gateway transfer events and trigger Glue workflows with minimal setup and no extra components.

- A is incorrect because scheduling jobs by time is unreliable for event-driven needs.
- C is incorrect because manual starts increase operational effort.
- D is incorrect because adding Lambda introduces an unnecessary middle layer.

Question: 147
A retail company uses Amazon Aurora PostgreSQL to process and store live transactional data. The company uses
an Amazon Redshift cluster for a data warehouse.
An extract, transform, and load (ETL) job runs every morning to update the Redshift cluster with new data from the
PostgreSQL database. The company has grown rapidly and needs to cost optimize the Redshift cluster.
A data engineer needs to create a solution to archive historical data. The data engineer must be able to run
analytics queries that effectively combine data from live transactional data in PostgreSQL, current data in
Redshift, and archived historical data. The solution must keep only the most recent 15 months of data in Amazon
Redshift to reduce costs.
Which combination of steps will meet these requirements? (Choose two.)

Options
A.Configure the Amazon Redshift Federated Query feature to query live transactional data that is in the
PostgreSQL database.
B.Configure Amazon Redshift Spectrum to query live transactional data that is in the PostgreSQL database.
C.Schedule a monthly job to copy data that is older than 15 months to Amazon S3 by using the UNLOAD
command. Delete the old data from the Redshift cluster. Configure Amazon Redshift Spectrum to access
historical data in Amazon S3.
D.Schedule a monthly job to copy data that is older than 15 months to Amazon S3 Glacier Flexible Retrieval by
using the UNLOAD command. Delete the old data from the Redshift cluster. Configure Redshift Spectrum to
access historical data from S3 Glacier Flexible Retrieval.
E.Create a materialized view in Amazon Redshift that combines live, current, and historical data from different
sources.

Answer: AC

Explanation:
Redshift Federated Query allows querying live Aurora PostgreSQL data. Spectrum enables access to archived historical data in S3, keeping Redshift optimized with only 15 months of current data.

- B is incorrect because Spectrum cannot query PostgreSQL directly.
- D is incorrect because Glacier retrieval is not suitable for analytics.
- E is incorrect because materialized views don’t combine live external sources like Aurora.

Question: 148
A manufacturing company has many IoT devices in facilities around the world. The company uses Amazon Kinesis
Data Streams to collect data from the devices. The data includes device ID, capture date, measurement type,
measurement value, and facility ID. The company uses facility ID as the partition key.
The company's operations team recently observed many WriteThroughputExceeded exceptions. The operations
team found that some shards were heavily used but other shards were generally idle.
How should the company resolve the issues that the operations team observed?

Options
A.Change the partition key from facility ID to a randomly generated key.
B.Increase the number of shards.
C.Archive the data on the producer's side.
D.Change the partition key from facility ID to capture date.

Answer: A

Explanation:
Using facility ID as partition key causes hot shards when few facilities send large amounts of data. A random key distributes records evenly across shards, preventing throttling.

- B is incorrect because simply adding shards won't fix uneven distribution.
- C is incorrect because archiving doesn't address throughput issues.
- D is incorrect because capture date won't evenly distribute data.

Question: 149
A data engineer wants to improve the performance of SQL queries in Amazon Athena that run against a sales data
table.
The data engineer wants to understand the execution plan of a specific SQL statement. The data engineer also
wants to see the computational cost of each operation in a SQL query.
Which statement does the data engineer need to run to meet these requirements?

Options
A.EXPLAIN SELECT * FROM sales;
B.EXPLAIN ANALYZE FROM sales;
C.EXPLAIN ANALYZE SELECT * FROM sales;
D.EXPLAIN FROM sales;

Answer: C

Explanation:
EXPLAIN ANALYZE provides both the query execution plan and actual runtime statistics, including computational cost, in Athena.

- A is incorrect because EXPLAIN shows only the logical execution plan, not runtime costs.
- B is incorrect because syntax is invalid.
- D is incorrect because syntax is invalid.

Question: 150
A company plans to provision a log delivery stream within a VPC. The company configured the VPC flow logs to
publish to Amazon CloudWatch Logs. The company needs to send the flow logs to Splunk in near real time for
further analysis.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Configure an Amazon Kinesis Data Streams data stream to use Splunk as the destination. Create a
CloudWatch Logs subscription filter to send log events to the data stream.
B.Create an Amazon Kinesis Data Firehose delivery stream to use Splunk as the destination. Create a
CloudWatch Logs subscription filter to send log events to the delivery stream.
C.Create an Amazon Kinesis Data Firehose delivery stream to use Splunk as the destination. Create an AWS
Lambda function to send the flow logs from CloudWatch Logs to the delivery stream.
D.Configure an Amazon Kinesis Data Streams data stream to use Splunk as the destination. Create an AWS
Lambda function to send the flow logs from CloudWatch Logs to the data stream.

Answer: B

Explanation:
Kinesis Data Firehose has native Splunk integration and can directly receive CloudWatch Logs through subscription filters. It requires minimal management and scales automatically.

- A is incorrect because Data Streams requires custom consumer management.
- C is incorrect because Lambda adds unnecessary overhead when Firehose can integrate directly.
- D is incorrect because Lambda + Data Streams is complex and higher overhead.

Question: 151
A data engineer is configuring an AWS Glue job to read data from an Amazon S3 bucket. The data engineer has set up the necessary AWS Glue connection details and an associated IAM role. However, when the data engineer attempts to run the AWS Glue job, the data engineer receives an error message that indicates that there are problems with the Amazon S3 VPC gateway endpoint. The data engineer must resolve the error and connect the AWS Glue job to the S3 bucket. Which solution will meet this requirement?

Options
A.Update the AWS Glue security group to allow inbound traffic from the Amazon S3 VPC gateway endpoint.
B.Configure an S3 bucket policy to explicitly grant the AWS Glue job permissions to access the S3 bucket.
C.Review the AWS Glue job code to ensure that the AWS Glue connection details include a fully qualified domain name.
D.Verify that the VPC's route table includes inbound and outbound routes for the Amazon S3 VPC gateway endpoint.

Answer: D

Explanation:
The issue is with the VPC S3 gateway endpoint configuration. For Glue to connect to S3 through a gateway endpoint, the correct routes must exist in the VPC route tables. Without them, requests to S3 won't reach the endpoint. That's why verifying and updating the VPC's route table is the right solution.

- A is incorrect because security groups are not applied to S3 VPC gateway endpoints. Security groups apply to ENIs, not gateway endpoints.
- B is incorrect because IAM/bucket policies control permissions but don't resolve networking issues.
- C is incorrect because Glue does not require FQDNs for S3 connections. The issue is networking, not DNS.

Question: 152
A retail company has a customer data hub in an Amazon S3 bucket. Employees from many countries use the data hub to support company-wide analytics. A governance team must ensure that the company's data analysts can access data only for customers who are within the same country as the analysts. Which solution will meet these requirements with the LEAST operational effort?

Options
A.Create a separate table for each country's customer data. Provide access to each analyst based on the country that the analyst serves.
B.Register the S3 bucket as a data lake location in AWS Lake Formation. Use the Lake Formation row-level security features to enforce the company's access policies.
C.Move the data to AWS Regions that are close to the countries where the customers are. Provide access to each analyst based on the country that the analyst serves.
D.Load the data into Amazon Redshift. Create a view for each country. Create separate IAM roles for each country to provide access to data from each country. Assign the appropriate roles to the analysts.

Answer: B

Explanation:
Lake Formation supports fine-grained security, including row-level security, making it the easiest and most scalable way to enforce per-country access with minimal operational burden.

- A is incorrect because manually splitting tables by country is complex, hard to scale, and requires ongoing maintenance.
- C is incorrect because moving data between Regions doesn’t solve governance or row-level restrictions.
- D is incorrect because Redshift with views and roles adds operational overhead compared to Lake Formation’s built-in row-level access controls.

Question: 153
A media company wants to improve a system that recommends media content to customer based on user behavior and preferences. To improve the recommendation system, the company needs to incorporate insights from third-party datasets into the company's existing analytics platform. The company wants to minimize the effort and time required to incorporate third-party datasets. Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Use API calls to access and integrate third-party datasets from AWS Data Exchange.
B.Use API calls to access and integrate third-party datasets from AWS DataSync.
C.Use Amazon Kinesis Data Streams to access and integrate third-party datasets from AWS CodeCommit repositories.
D.Use Amazon Kinesis Data Streams to access and integrate third-party datasets from Amazon Elastic Container Registry (Amazon ECR).

Answer: A

Explanation:
AWS Data Exchange provides ready-to-use third-party datasets through APIs, minimizing the integration effort and operational overhead.

- B is incorrect because DataSync is for file transfer between storage systems, not consuming curated datasets.
- C is incorrect because CodeCommit is for source code, not data marketplaces.
- D is incorrect because ECR stores container images, not data sets.

Question: 154
A financial company wants to implement a data mesh. The data mesh must support centralized data governance, data analysis, and data access control. The company has decided to use AWS Glue for data catalogs and extract, transform, and load (ETL) operations. Which combination of AWS services will implement a data mesh? (Choose two.)

Options
A.Use Amazon Aurora for data storage. Use an Amazon Redshift provisioned cluster for data analysis.
B.Use Amazon S3 for data storage. Use Amazon Athena for data analysis.
C.Use AWS Glue DataBrew for centralized data governance and access control.
D.Use Amazon RDS for data storage. Use Amazon EMR for data analysis.
E.Use AWS Lake Formation for centralized data governance and access control.

Answer: BE

Explanation:
Amazon S3 + Athena provide a scalable, serverless storage and query layer ideal for a data mesh. Lake Formation adds centralized governance, security, and access control across domains, which is essential.

- A is incorrect because Aurora/Redshift are not decentralized or lightweight enough for a mesh pattern.
- C is incorrect because Glue DataBrew is for data preparation, not centralized governance.
- D is incorrect because RDS + EMR are not optimal for federated governance in a data mesh.

Question: 155
A data engineer maintains custom Python scripts that perform a data formatting process that many AWS Lambda functions use. When the data engineer needs to modify the Python scripts, the data engineer must manually update all the Lambda functions. The data engineer requires a less manual way to update the Lambda functions. Which solution will meet this requirement?

Options
A.Store a pointer to the custom Python scripts in the execution context object in a shared Amazon S3 bucket.
B.Package the custom Python scripts into Lambda layers. Apply the Lambda layers to the Lambda functions.
C.Store a pointer to the custom Python scripts in environment variables in a shared Amazon S3 bucket.
D.Assign the same alias to each Lambda function. Call reach Lambda function by specifying the function's alias.

Answer: B

Explanation:
Lambda layers let you package libraries and custom scripts once, and reuse them across multiple Lambda functions. Updating the layer automatically updates all functions referencing it.

- A is incorrect because storing scripts in S3 still requires manual fetch and integration into each Lambda.
- C is incorrect because environment variables cannot store actual script code.
- D is incorrect because aliases manage versions of Lambda, not shared code.

Question: 156
A company created an extract, transform, and load (ETL) data pipeline in AWS Glue. A data engineer must crawl a table that is in Microsoft SQL Server. The data engineer needs to extract, transform, and load the output of the crawl to an Amazon S3 bucket. The data engineer also must orchestrate the data pipeline. Which AWS service or feature will meet these requirements MOST cost-effectively?

Options
A.AWS Step Functions
B.AWS Glue workflows
C.AWS Glue Studio
D.Amazon Managed Workflows for Apache Airflow (Amazon MWAA)

Answer: B

Explanation:
AWS Glue workflows can orchestrate Glue crawlers and Glue ETL jobs natively at no extra cost beyond Glue usage, making it the most cost-effective choice.

- A is incorrect because Step Functions incurs additional costs and is unnecessary if Glue workflows suffice.
- C is incorrect because Glue Studio is a UI for job authoring, not orchestration.
- D is incorrect because MWAA adds extra cost and complexity compared to built-in Glue workflows.

Question: 157
A financial services company stores financial data in Amazon Redshift. A data engineer wants to run real-time queries on the financial data to support a web-based trading application. The data engineer wants to run the queries from within the trading application. Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Establish WebSocket connections to Amazon Redshift.
B.Use the Amazon Redshift Data API.
C.Set up Java Database Connectivity (JDBC) connections to Amazon Redshift.
D.Store frequently accessed data in Amazon S3. Use Amazon S3 Select to run the queries.

Answer: B

Explanation:
The Redshift Data API allows applications to run queries directly over HTTPS without managing persistent connections, making it low-overhead and ideal for serverless apps.

- A is incorrect because Redshift does not support WebSocket connections.
- C is incorrect because JDBC requires managing connections, drivers, and scaling.
- D is incorrect because S3 Select works only on S3 objects, not Redshift data.

Question: 158
A company uses Amazon Athena for one-time queries against data that is in Amazon S3. The company has several use cases. The company must implement permission controls to separate query processes and access to query history among users, teams, and applications that are in the same AWS account. Which solution will meet these requirements?

Options
A.Create an S3 bucket for each use case. Create an S3 bucket policy that grants permissions to appropriate individual IAM users. Apply the S3 bucket policy to the S3 bucket.
B.Create an Athena workgroup for each use case. Apply tags to the workgroup. Create an IAM policy that uses the tags to apply appropriate permissions to the workgroup.
C.Create an IAM role for each use case. Assign appropriate permissions to the role for each use case. Associate the role with Athena.
D.Create an AWS Glue Data Catalog resource policy that grants permissions to appropriate individual IAM users for each use case. Apply the resource policy to the specific tables that Athena uses.

Answer: B

Explanation:
Athena workgroups allow separation of queries, query history, and resource usage. Using IAM with tags on workgroups provides fine-grained access control.

- A is incorrect because splitting S3 buckets is unnecessary and does not manage query history separation.
- C is incorrect because IAM roles don't inherently separate query history within Athena.
- D is incorrect because Glue Data Catalog policies control metadata, not Athena query isolation.

Question: 159
A data engineer needs to schedule a workflow that runs a set of AWS Glue jobs every day. The data engineer does not require the Glue jobs to run or finish at a specific time. Which solution will run the Glue jobs in the MOST cost-effective way?

Options
A.Choose the FLEX execution class in the Glue job properties.
B.Use the Spot Instance type in Glue job properties.
C.Choose the STANDARD execution class in the Glue job properties.
D.Choose the latest version in the GlueVersion field in the Glue job properties.

Answer: A

Explanation:
The FLEX execution class allows Glue jobs to run on spare compute capacity, reducing costs if timing is not critical.

- B is incorrect because Glue does not provide Spot instances; that applies to EC2, not Glue.
- C is incorrect because STANDARD is faster but more costly than FLEX.
- D is incorrect because GlueVersion only sets job runtime version, not cost optimization.

Question: 160
A data engineer needs to create an AWS Lambda function that converts the format of data from .csv to Apache Parquet. The Lambda function must run only if a user uploads a .csv file to an Amazon S3 bucket. Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Create an S3 event notification that has an event type of s3:ObjectCreated:*. Use a filter rule to generate notifications only when the suffix includes .csv. Set the Amazon Resource Name (ARN) of the Lambda function as the destination for the event notification.*
B.Create an S3 event notification that has an event type of s3:ObjectTagging: for objects that have a tag set to .csv. Set the Amazon Resource Name (ARN) of the Lambda function as the destination for the event notification.
C.Create an S3 event notification that has an event type of s3:*. Use a filter rule to generate notifications only when the suffix includes .csv. Set the Amazon Resource Name (ARN) of the Lambda function as the destination for the event notification.*
D.Create an S3 event notification that has an event type of s3:ObjectCreated:. Use a filter rule to generate notifications only when the suffix includes .csv. Set an Amazon Simple Notification Service (Amazon SNS) topic as the destination for the event notification. Subscribe the Lambda function to the SNS topic.

Answer: A

Explanation:
The simplest approach is to directly trigger the Lambda with an S3 ObjectCreated:* event, filtered for .csv suffix. This ensures only .csv files cause execution, with minimal overhead.

- B is incorrect because S3 tagging events do not match file creation and require extra tagging.
- C is incorrect because s3:* is too broad and inefficient.
- D is incorrect because adding SNS adds unnecessary complexity when direct S3-to-Lambda integration is possible.

Question: 161
A company has a data lake on AWS. The data lake ingests sources of data from business units. The company uses
Amazon Athena for queries. The storage layer is Amazon S3 with an AWS Glue Data Catalog as a metadata
repository.
The company wants to make the data available to data scientists and business analysts. However, the company
first needs to manage fine-grained, column-level data access for Athena based on the user roles and
responsibilities.
Which solution will meet these requirements?

Options
A.Set up AWS Lake Formation. Define security policy-based rules for the users and applications by IAM role in Lake Formation.
B.Define an IAM resource-based policy for AWS Glue tables. Attach the same policy to IAM user groups.
C.Define an IAM identity-based policy for AWS Glue tables. Attach the same policy to IAM roles. Associate the IAM roles with IAM groups that contain the users.
D.Create a resource share in AWS Resource Access Manager (AWS RAM) to grant access to IAM users.

Answer: A

Explanation:
Lake Formation provides fine-grained access control, including row-level and column-level permissions for Athena queries. It integrates with Glue Data Catalog and IAM seamlessly.

- B is incorrect because Glue resource policies don't provide column-level controls.
- C is incorrect because IAM policies can't enforce column-level access.
- D is incorrect because AWS RAM is for sharing resources across accounts, not for fine-grained data access.

Question: 162
A company has developed several AWS Glue extract, transform, and load (ETL) jobs to validate and transform data
from Amazon S3. The ETL jobs load the data into Amazon RDS for MySQL in batches once every day. The ETL jobs
use a DynamicFrame to read the S3 data. The ETL jobs currently process all the data that is in the S3 bucket. However, the company wants the jobs to
process only the daily incremental data.
Which solution will meet this requirement with the LEAST coding effort?

Options
A.Create an ETL job that reads the S3 file status and logs the status in Amazon DynamoDB.
B.Enable job bookmarks for the ETL jobs to update the state after a run to keep track of previously processed data.
C.Enable job metrics for the ETL jobs to help keep track of processed objects in Amazon CloudWatch.
D.Configure the ETL jobs to delete processed objects from Amazon S3 after each run.

Answer: B

Explanation:
Glue job bookmarks allow ETL jobs to process only new or changed data, avoiding reprocessing and requiring minimal code changes.

- A is incorrect because manually logging status in DynamoDB requires extra development.
- C is incorrect because job metrics are for monitoring, not incremental data tracking.
- D is incorrect because deleting source data is risky and not best practice.

Question: 163
An online retail company has an application that runs on Amazon EC2 instances that are in a VPC. The company
wants to collect flow logs for the VPC and analyze network traffic.
Which solution will meet these requirements MOST cost-effectively?

Options
A.Publish flow logs to Amazon CloudWatch Logs. Use Amazon Athena for analytics.
B.Publish flow logs to Amazon CloudWatch Logs. Use an Amazon OpenSearch Service cluster for analytics.
C.Publish flow logs to Amazon S3 in text format. Use Amazon Athena for analytics.
D.Publish flow logs to Amazon S3 in Apache Parquet format. Use Amazon Athena for analytics.

Answer: D

Explanation:
Storing VPC flow logs in S3 in Parquet format optimizes cost and performance. Parquet reduces storage size and Athena query costs while providing efficient analytics.

- A is incorrect because Athena cannot query CloudWatch Logs directly.
- B is incorrect because OpenSearch is more expensive and unnecessary.
- C is incorrect because querying text data in Athena is slower and more costly than Parquet.

Question: 164
A retail company stores transactions, store locations, and customer information tables in four reserved ra3.4xlarge
Amazon Redshift cluster nodes. All three tables use even table distribution.
The company updates the store location table only once or twice every few years.
A data engineer notices that Redshift queues are slowing down because the whole store location table is
constantly being broadcast to all four compute nodes for most queries. The data engineer wants to speed up the
query performance by minimizing the broadcasting of the store location table.
Which solution will meet these requirements in the MOST cost-effective way?

Options
A.Change the distribution style of the store location table from EVEN distribution to ALL distribution.
B.Change the distribution style of the store location table to KEY distribution based on the column that has the highest dimension.
C.Add a join column named store_id into the sort key for all the tables.
D.Upgrade the Redshift reserved node to a larger instance size in the same instance family.

Answer: A

Explanation:
Using ALL distribution replicates small, rarely updated tables across all nodes, avoiding repeated broadcasts and improving performance.

- B is incorrect because KEY distribution is not suitable for a small, static table.
- C is incorrect because sort keys affect query scan efficiency, not broadcast behavior.
- D is incorrect because upgrading nodes increases cost without solving the distribution issue.

Question: 165
A company has a data warehouse that contains a table that is named Sales. The company stores the table in
Amazon Redshift. The table includes a column that is named city_name. The company wants to query the table to
find all rows that have a city_name that starts with "San" or "El".
Which SQL query will meet this requirement?

Options
A.Select * from Sales where city_name ~ '$(San|El)*';
*B.Select * from Sales where city_name ~ '^(San|El)*'*;
C.Select * from Sales where city_name ~'$(San&El)*';
*D.Select * from Sales where city_name ~ '^(San&El)*'*;

Answer: B

Explanation:
~ '^(San|El)*' correctly uses regex to match values starting with "San" or "El". ^ anchors the start of the string, and (San|El) matches either prefix.

- A is incorrect because $ anchors the end of the string, not the start.
- C and D are incorrect because & is not valid in regex for "OR".

Question: 166
A company needs to send customer call data from its on-premises PostgreSQL database to AWS to generate near
real-time insights. The solution must capture and load updates from operational data stores that run in the
PostgreSQL database. The data changes continuously.
A data engineer configures an AWS Database Migration Service (AWS DMS) ongoing replication task. The task
reads changes in near real time from the PostgreSQL source database transaction logs for each table. The task
then sends the data to an Amazon Redshift cluster for processing.
The data engineer discovers latency issues during the change data capture (CDC) of the task. The data engineer
thinks that the PostgreSQL source database is causing the high latency.
Which solution will confirm that the PostgreSQL database is the source of the high latency?

Options
A.Use Amazon CloudWatch to monitor the DMS task. Examine the CDCIncomingChanges metric to identify delays in the CDC from the source database.
B.Verify that logical replication of the source database is configured in the postgresql.conf configuration file.
C.Enable Amazon CloudWatch Logs for the DMS endpoint of the source database. Check for error messages.
D.Use Amazon CloudWatch to monitor the DMS task. Examine the CDCLatencySource metric to identify delays in the CDC from the source database.

Answer: D

Explanation:
The CDCLatencySource metric shows the latency between the source database and DMS, confirming whether PostgreSQL is the bottleneck.

- A is incorrect because CDCIncomingChanges shows volume of changes, not latency.
- B is incorrect because checking configuration doesn’t confirm runtime latency.
- C is incorrect because logs may not clearly indicate latency issues.

Question: 167
A lab uses IoT sensors to monitor humidity, temperature, and pressure for a project. The sensors send 100 KB of
data every 10 seconds. A downstream process will read the data from an Amazon S3 bucket every 30 seconds.
Which solution will deliver the data to the S3 bucket with the LEAST latency?

Options
A.Use Amazon Kinesis Data Streams and Amazon Kinesis Data Firehose to deliver the data to the S3 bucket. Use the default buffer interval for Kinesis Data Firehose.
B.Use Amazon Kinesis Data Streams to deliver the data to the S3 bucket. Configure the stream to use 5 provisioned shards.
C.Use Amazon Kinesis Data Streams and call the Kinesis Client Library to deliver the data to the S3 bucket. Use a 5 second buffer interval from an application.
D.Use Amazon Managed Service for Apache Flink (previously known as Amazon Kinesis Data Analytics) and Amazon Kinesis Data Firehose to deliver the data to the S3 bucket. Use a 5 second buffer interval for Kinesis Data Firehose.

Answer: C

Explanation:
Using Kinesis Data Streams with Kinesis Client Library and a short buffer interval minimizes latency while still providing reliable delivery to S3.

- A is incorrect because Firehose default buffer interval is 300 seconds, too high latency.
- B is incorrect because Data Streams alone doesn't write to S3; it requires consumers.
- D is incorrect because Managed Flink + Firehose adds extra services and latency.

Question: 168
A company wants to use machine learning (ML) to perform analytics on data that is in an Amazon S3 data lake. The
company has two data transformation requirements that will give consumers within the company the ability to
create reports.
The company must perform daily transformations on 300 GB of data that is in a variety format that must arrive in
Amazon S3 at a scheduled time. The company must perform one-time transformations of terabytes of archived
data that is in the S3 data lake. The company uses Amazon Managed Workflows for Apache Airflow (Amazon
MWAA) Directed Acyclic Graphs (DAGs) to orchestrate processing.
Which combination of tasks should the company schedule in the Amazon MWAA DAGs to meet these requirements
MOST cost-effectively? (Choose two.)

Options
A.For daily incoming data, use AWS Glue crawlers to scan and identify the schema.
B.For daily incoming data, use Amazon Athena to scan and identify the schema.
C.For daily incoming data, use Amazon Redshift to perform transformations.
D.For daily and archived data, use Amazon EMR to perform data transformations.
E.For archived data, use Amazon SageMaker to perform data transformations.

Answer: AD

Explanation:
AWS Glue crawlers efficiently identify schema for daily datasets, and Amazon EMR provides scalable transformation for both daily and large archived data.

- B is incorrect because Athena is for queries, not schema detection.
- C is incorrect because Redshift is not cost-efficient for raw ETL.
- E is incorrect because SageMaker is for ML, not bulk ETL.

Question: 169
A retail company uses AWS Glue for extract, transform, and load (ETL) operations on a dataset that contains
information about customer orders. The company wants to implement specific validation rules to ensure data
accuracy and consistency.
Which solution will meet these requirements?

Options
A.Use AWS Glue job bookmarks to track the data for accuracy and consistency.
B.Create custom AWS Glue Data Quality rulesets to define specific data quality checks.
C.Use the built-in AWS Glue Data Quality transforms for standard data quality validations.
D.Use AWS Glue Data Catalog to maintain a centralized data schema and metadata repository.

Answer: B

Explanation:
Glue Data Quality rulesets allow custom validation logic to enforce business-specific accuracy and consistency checks.

- A is incorrect because bookmarks track incremental data, not quality.
- C is incorrect because built-in transforms only handle general validations, not custom rules.
- D is incorrect because Data Catalog stores metadata, not validation logic.

Question: 170
An insurance company stores transaction data that the company compressed with gzip.
The company needs to query the transaction data for occasional audits.
Which solution will meet this requirement in the MOST cost-effective way?

Options
A.Store the data in Amazon Glacier Flexible Retrieval. Use Amazon S3 Glacier Select to query the data.
B.Store the data in Amazon S3. Use Amazon S3 Select to query the data.
C.Store the data in Amazon S3. Use Amazon Athena to query the data.
D.Store the data in Amazon Glacier Instant Retrieval. Use Amazon Athena to query the data.

Answer: A

Explanation:
Glacier Flexible Retrieval with Glacier Select is the lowest-cost storage and provides query-in-place for occasional audits.

- B is incorrect because keeping all data in S3 costs more over time.
- C is incorrect because Athena queries all data, increasing cost unnecessarily for rare access.
- D is incorrect because Glacier Instant Retrieval is more expensive than Flexible Retrieval.

Question: 171
A company uses Amazon RDS for MySQL as the database for a critical application. The database workload is mostly writes, with a small number of reads.
A data engineer notices that the CPU utilization of the DB instance is very high. The high CPU utilization is slowing down the application. The data engineer must reduce the CPU utilization of the DB Instance.
Which actions should the data engineer take to meet this requirement? (Choose two.)

Options
A.Use the Performance Insights feature of Amazon RDS to identify queries that have high CPU utilization. Optimize the problematic queries.
B.Modify the database schema to include additional tables and indexes.
C.Reboot the RDS DB instance once each week.
D.Upgrade to a larger instance size.
E.Implement caching to reduce the database query load.

Answer: AD

Explanation:
Using Performance Insights helps identify and optimize queries causing CPU spikes. Upgrading to a larger instance size increases available CPU capacity to handle workloads.

- B is incorrect because schema changes may not directly reduce CPU utilization without targeted optimization.
- C is incorrect because rebooting doesn't solve high CPU utilization issues.
- E is incorrect because caching helps reduce reads, but this workload is mostly writes.

Question: 172
A company has used an Amazon Redshift table that is named Orders for 6 months. The company performs weekly updates and deletes on the table. The table has an interleaved sort key on a column that contains AWS Regions. The company wants to reclaim disk space so that the company will not run out of storage space. The company also wants to analyze the sort key column.
Which Amazon Redshift command will meet these requirements?

Options
A.VACUUM FULL Orders
B.VACUUM DELETE ONLY Orders
C.VACUUM REINDEX Orders
D.VACUUM SORT ONLY Orders

Answer: C

Explanation:
VACUUM REINDEX is used for tables with interleaved sort keys to rebuild them and reclaim space while optimizing query performance.

- A is incorrect because VACUUM FULL reclaims space but is not ideal for interleaved keys.
- B is incorrect because VACUUM DELETE ONLY just reclaims deleted rows, not reindexes.
- D is incorrect because VACUUM SORT ONLY is used to resort rows, not reindex interleaved keys.

Question: 173
A manufacturing company wants to collect data from sensors. A data engineer needs to implement a solution that ingests sensor data in near real time.
The solution must store the data to a persistent data store. The solution must store the data in nested JSON format. The company must have the ability to query from the data store with a latency of less than 10 milliseconds.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Use a self-hosted Apache Kafka cluster to capture the sensor data. Store the data in Amazon S3 for querying.
B.Use AWS Lambda to process the sensor data. Store the data in Amazon S3 for querying.
C.Use Amazon Kinesis Data Streams to capture the sensor data. Store the data in Amazon DynamoDB for querying.
D.Use Amazon Simple Queue Service (Amazon SQS) to buffer incoming sensor data. Use AWS Glue to store the data in Amazon RDS for querying.

Answer: C

Explanation:
Kinesis Data Streams provides near real-time ingestion, and DynamoDB can persist JSON with sub-10 ms query latency. This combination meets requirements with minimal management.

- A is incorrect because self-hosting Kafka adds heavy operational overhead.
- B is incorrect because S3 is not suitable for sub-10 ms queries.
- D is incorrect because RDS queries have higher latency and SQS isn't designed for real-time streams.

Question: 174
A company stores data in a data lake that is in Amazon S3. Some data that the company stores in the data lake contains personally identifiable information (PII). Multiple user groups need to access the raw data. The company must ensure that user groups can access only the PII that they require.
Which solution will meet these requirements with the LEAST effort?

Options
A.Use Amazon Athena to query the data. Set up AWS Lake Formation and create data filters to establish levels of access for the company's IAM roles. Assign each user to the IAM role that matches the user's PII access requirements.
B.Use Amazon QuickSight to access the data. Use column-level security features in QuickSight to limit the PII that users can retrieve from Amazon S3 by using Amazon Athena. Define QuickSight access levels based on the PII access requirements of the users.
C.Build a custom query builder UI that will run Athena queries in the background to access the data. Create user groups in Amazon Cognito. Assign access levels to the user groups based on the PII access requirements of the users.
D.Create IAM roles that have different levels of granular access. Assign the IAM roles to IAM user groups. Use an identity-based policy to assign access levels to user groups at the column level.

Answer: A

Explanation:
Lake Formation provides fine-grained access control for S3-based data lakes, allowing filtering by row or column. It integrates directly with Athena, enabling secure, minimal-effort governance.

- B is incorrect because QuickSight column-level security applies only in QuickSight dashboards, not in the data lake itself.
- C is incorrect because building a custom UI adds unnecessary development overhead.
- D is incorrect because IAM policies cannot enforce column-level access.

Question: 175
A data engineer must build an extract, transform, and load (ETL) pipeline to process and load data from 10 source systems into 10 tables that are in an Amazon Redshift database. All the source systems generate .csv, JSON, or Apache Parquet files every 15 minutes. The source systems all deliver files into one Amazon S3 bucket. The file sizes range from 10 MB to 20 GB. The ETL pipeline must function correctly despite changes to the data schema.
Which data pipeline solutions will meet these requirements? (Choose two.)

Options
A.Use an Amazon EventBridge rule to run an AWS Glue job every 15 minutes. Configure the AWS Glue job to process and load the data into the Amazon Redshift tables.
B.Use an Amazon EventBridge rule to invoke an AWS Glue workflow job every 15 minutes. Configure the AWS Glue workflow to have an on-demand trigger that runs an AWS Glue crawler and then runs an AWS Glue job when the crawler finishes running successfully. Configure the AWS Glue job to process and load the data into the Amazon Redshift tables.
C.Configure an AWS Lambda function to invoke an AWS Glue crawler when a file is loaded into the S3 bucket. Configure an AWS Glue job to process and load the data into the Amazon Redshift tables. Create a second Lambda function to run the AWS Glue job. Create an Amazon EventBridge rule to invoke the second Lambda function when the AWS Glue crawler finishes running successfully.
D.Configure an AWS Lambda function to invoke an AWS Glue workflow when a file is loaded into the S3 bucket. Configure the AWS Glue workflow to have an on-demand trigger that runs an AWS Glue crawler and then runs an AWS Glue job when the crawler finishes running successfully. Configure the AWS Glue job to process and load the data into the Amazon Redshift tables.
E.Configure an AWS Lambda function to invoke an AWS Glue job when a file is loaded into the S3 bucket. Configure the AWS Glue job to read the files from the S3 bucket into an Apache Spark DataFrame. Configure the AWS Glue job to also put smaller partitions of the DataFrame into an Amazon Kinesis Data Firehose delivery stream. Configure the delivery stream to load data into the Amazon Redshift tables.

Answer: BD

Explanation:
Glue workflows orchestrated with crawlers handle schema changes dynamically and reliably. EventBridge triggers provide scheduling and automation with minimal management.

- A is incorrect because directly scheduling Glue jobs without schema crawlers may fail on schema changes.
- C is incorrect because chaining multiple Lambdas increases complexity unnecessarily.
- E is incorrect because introducing Firehose for Redshift adds complexity not required.

Question: 176
A financial company wants to use Amazon Athena to run on-demand SQL queries on a petabyte-scale dataset to support a business intelligence (BI) application. An AWS Glue job that runs during non-business hours updates the dataset once every day. The BI application has a standard data refresh frequency of 1 hour to comply with company policies.
A data engineer wants to cost optimize the company's use of Amazon Athena without adding any additional infrastructure costs.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Configure an Amazon S3 Lifecycle policy to move data to the S3 Glacier Deep Archive storage class after 1 day.
B.Use the query result reuse feature of Amazon Athena for the SQL queries.
C.Add an Amazon ElastiCache cluster between the BI application and Athena.
D.Change the format of the files that are in the dataset to Apache Parquet.

Answer: B

Explanation:
Athena’s query result reuse feature caches query results. For repetitive queries within the refresh window, Athena can reuse results, reducing cost and improving performance.

- A is incorrect because Glacier Deep Archive prevents fast query access.
- C is incorrect because ElastiCache does not integrate directly with Athena.
- D is incorrect because Parquet optimizes scans but doesn’t help with hourly query refreshes.

Question: 177
A company's data engineer needs to optimize the performance of table SQL queries. The company stores data in an Amazon Redshift cluster. The data engineer cannot increase the size of the cluster because of budget constraints.
The company stores the data in multiple tables and loads the data by using the EVEN distribution style. Some tables are hundreds of gigabytes in size. Other tables are less than 10 MB in size.
Which solution will meet these requirements?

Options
A.Keep using the EVEN distribution style for all tables. Specify primary and foreign keys for all tables.
B.Use the ALL distribution style for large tables. Specify primary and foreign keys for all tables.
C.Use the ALL distribution style for rarely updated small tables. Specify primary and foreign keys for all tables.
D.Specify a combination of distribution, sort, and partition keys for all tables.

Answer: C

Explanation:
The ALL distribution style replicates small tables across all nodes, eliminating joins overhead. This is efficient for small dimension tables under 10 MB. Large tables should remain evenly distributed.

- A is incorrect because EVEN distribution for all tables doesn't optimize joins with small tables.
- B is incorrect because replicating large tables with ALL wastes storage.
- D is incorrect because sort/partition keys help, but distribution style choice is the key performance factor here.

Question: 178
A company receives .csv files that contain physical address data. The data is in columns that have the following names: Door_No, Street_Name, City, and Zip_Code. The company wants to create a single column to store these values in the following format:
Which solution will meet this requirement with the LEAST coding effort?

Options
A.Use AWS Glue DataBrew to read the files. Use the NEST_TO_ARRAY transformation to create the new column.
B.Use AWS Glue DataBrew to read the files. Use the NEST_TO_MAP transformation to create the new column.
C.Use AWS Glue DataBrew to read the files. Use the PIVOT transformation to create the new column.
D.Write a Lambda function in Python to read the files. Use the Python data dictionary type to create the new column.

Answer: B

Explanation:
DataBrew's NEST_TO_MAP transformation combines multiple columns into a single column with minimal coding.

- A is incorrect because NEST_TO_ARRAY generates an array, not a formatted string column.
- C is incorrect because PIVOT is used for restructuring rows/columns, not concatenating values.
- D is incorrect because writing a Lambda function requires custom code and more effort.

Question: 179
A company receives call logs as Amazon S3 objects that contain sensitive customer information. The company must protect the S3 objects by using encryption. The company must also use encryption keys that only specific employees can access.
Which solution will meet these requirements with the LEAST effort?

Options
A.Use an AWS CloudHSM cluster to store the encryption keys. Configure the process that writes to Amazon S3 to make calls to CloudHSM to encrypt and decrypt the objects. Deploy an IAM policy that restricts access to the CloudHSM cluster.
B.Use server-side encryption with customer-provided keys (SSE-C) to encrypt the objects that contain customer information. Restrict access to the keys that encrypt the objects.
C.Use server-side encryption with AWS KMS keys (SSE-KMS) to encrypt the objects that contain customer information. Configure an IAM policy that restricts access to the KMS keys that encrypt the objects.
D.Use server-side encryption with Amazon S3 managed keys (SSE-S3) to encrypt the objects that contain customer information. Configure an IAM policy that restricts access to the Amazon S3 managed keys that encrypt the objects.

Answer: C

Explanation:
SSE-KMS provides key-level access control using IAM policies, ensuring only specific employees can use the keys. It requires minimal setup compared to HSM or custom key management.

- A is incorrect because CloudHSM adds unnecessary complexity and overhead.
- B is incorrect because SSE-C requires manual key management for every request.
- D is incorrect because SSE-S3 uses AWS-managed keys without fine-grained employee access control.

Question: 180
A company stores petabytes of data in thousands of Amazon S3 buckets in the S3 Standard storage class. The data supports analytics workloads that have unpredictable and variable data access patterns.
The company does not access some data for months. However, the company must be able to retrieve all data within milliseconds. The company needs to optimize S3 storage costs.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Use S3 Storage Lens standard metrics to determine when to move objects to more cost-optimized storage classes. Create S3 Lifecycle policies for the S3 buckets to move objects to cost-optimized storage classes. Continue to refine the S3 Lifecycle policies in the future to optimize storage costs.
B.Use S3 Storage Lens activity metrics to identify S3 buckets that the company accesses infrequently. Configure S3 Lifecycle rules to move objects from S3 Standard to the S3 Standard-Infrequent Access (S3 Standard-IA) and S3 Glacier storage classes based on the age of the data.
C.Use S3 Intelligent-Tiering. Activate the Deep Archive Access tier.
D.Use S3 Intelligent-Tiering. Use the default access tier.

Answer: D

Explanation:
S3 Intelligent-Tiering automatically optimizes storage costs for unpredictable access patterns without lifecycle management. The default tier meets millisecond access requirements while reducing cost.

- A is incorrect because manually refining lifecycle rules increases operational overhead.
- B is incorrect because Glacier does not provide millisecond retrieval.
- C is incorrect because Deep Archive retrieval takes hours, not milliseconds.

Question: 181
A data engineer must use AWS services to ingest a dataset into an Amazon S3 data lake. The data engineer profiles the dataset and discovers that the dataset contains personally identifiable information (PII). The data engineer must implement a solution to profile the dataset and obfuscate the PII. Which solution will meet this requirement with the LEAST operational effort?

Options
A.Use an Amazon Kinesis Data Firehose delivery stream to process the dataset. Create an AWS Lambda transform function to identify the PII. Use an AWS SDK to obfuscate the PII. Set the S3 data lake as the target for the delivery stream.
B.Use the Detect PII transform in AWS Glue Studio to identify the PII. Obfuscate the PII. Use an AWS Step Functions state machine to orchestrate a data pipeline to ingest the data into the S3 data lake.
C.Use the Detect PII transform in AWS Glue Studio to identify the PII. Create a rule in AWS Glue Data Quality to obfuscate the PII. Use an AWS Step Functions state machine to orchestrate a data pipeline to ingest the data into the S3 data lake.
D.Ingest the dataset into Amazon DynamoDB. Create an AWS Lambda function to identify and obfuscate the PII in the DynamoDB table and to transform the data. Use the same Lambda function to ingest the data into the S3 data lake.

Answer: B

Explanation:
AWS Glue Studio has a built-in Detect PII transform, which can easily detect sensitive data and obfuscate it with minimal coding effort. Combining this with orchestration in Step Functions makes the pipeline simple and serverless.

- A is incorrect because Kinesis + Lambda requires custom logic for PII detection and adds operational effort.
- C is incorrect because AWS Glue Data Quality is for rules-based validation, not obfuscation.
- D is incorrect because ingesting into DynamoDB first adds unnecessary complexity.

Question: 182
A company maintains multiple extract, transform, and load (ETL) workflows that ingest data from the company's operational databases into an Amazon S3 based data lake. The ETL workflows use AWS Glue and Amazon EMR to process data. The company wants to improve the existing architecture to provide automated orchestration and to require minimal manual effort. Which solution will meet these requirements with the LEAST operational overhead?

Options
A.AWS Glue workflows
B.AWS Step Functions tasks
C.AWS Lambda functions
D.Amazon Managed Workflows for Apache Airflow (Amazon MWAA) workflows

Answer: B

Explanation:
AWS Step Functions provides a serverless way to orchestrate multiple workflows with minimal manual effort. It integrates easily with Glue and EMR jobs and provides retries, parallelism, and monitoring.

- A is incorrect because Glue workflows only handle Glue jobs, not EMR.
- C is incorrect because Lambda alone cannot orchestrate multiple workflows reliably.
- D is incorrect because MWAA adds operational overhead compared to Step Functions.

Question: 183
A company currently stores all of its data in Amazon S3 by using the S3 Standard storage class. A data engineer examined data access patterns to identify trends. During the first 6 months, most data files are accessed several times each day. Between 6 months and 2 years, most data files are accessed once or twice each month. After 2 years, data files are accessed only once or twice each year. The data engineer needs to use an S3 Lifecycle policy to develop new data storage rules. The new storage solution must continue to provide high availability. Which solution will meet these requirements in the MOST cost-effective way?

Options
A.Transition objects to S3 One Zone-Infrequent Access (S3 One Zone-IA) after 6 months. Transfer objects to S3 Glacier Flexible Retrieval after 2 years.
B.Transition objects to S3 Standard-Infrequent Access (S3 Standard-IA) after 6 months. Transfer objects to S3 Glacier Flexible Retrieval after 2 years.
C.Transition objects to S3 Standard-Infrequent Access (S3 Standard-IA) after 6 months. Transfer objects to S3 Glacier Deep Archive after 2 years.
D.Transition objects to S3 One Zone-Infrequent Access (S3 One Zone-IA) after 6 months. Transfer objects to S3 Glacier Deep Archive after 2 years.

Answer: C

Explanation:
S3 Standard-IA is designed for infrequently accessed data with high availability, which fits the 6 months–2 years range. After 2 years, Glacier Deep Archive is the most cost-effective for long-term retention with rare access.

- A is incorrect because S3 One Zone-IA does not provide high availability across multiple AZs.
- B is incorrect because Glacier Flexible Retrieval is more expensive than Deep Archive for data rarely accessed after 2 years.

- D is incorrect because One Zone-IA sacrifices durability/availability, not meeting the requirement.

Question: 184
A company maintains an Amazon Redshift provisioned cluster that the company uses for extract, transform, and load (ETL) operations to support critical analysis tasks. A sales team within the company maintains a Redshift cluster that the sales team uses for business intelligence (BI) tasks. The sales team recently requested access to the data that is in the ETL Redshift cluster so the team can perform weekly summary analysis tasks. The sales team needs to join data from the ETL cluster with data that is in the sales team's BI cluster. The company needs a solution that will share the ETL cluster data with the sales team without interrupting the critical analysis tasks. The solution must minimize usage of the computing resources of the ETL cluster. Which solution will meet these requirements?

Options
A.Set up the sales team BI cluster as a consumer of the ETL cluster by using Redshift data sharing.
B.Create materialized views based on the sales team's requirements. Grant the sales team direct access to the ETL cluster.
C.Create database views based on the sales team's requirements. Grant the sales team direct access to the ETL cluster.
D.Unload a copy of the data from the ETL cluster to an Amazon S3 bucket every week. Create an Amazon Redshift Spectrum table based on the content of the ETL cluster.

Answer: A

Explanation:
Redshift Data Sharing allows one cluster to share data in near real time with another cluster without duplicating data or impacting compute resources of the producer cluster.

- B is incorrect because materialized views still consume compute on the ETL cluster.
- C is incorrect because direct queries also increase workload on the ETL cluster.
- D is incorrect because unloading to S3 is batch, adds latency, and adds overhead.

Question: 185
A data engineer needs to join data from multiple sources to perform a one-time analysis job. The data is stored in Amazon DynamoDB, Amazon RDS, Amazon Redshift, and Amazon S3. Which solution will meet this requirement MOST cost-effectively?

Options
A.Use an Amazon EMR provisioned cluster to read from all sources. Use Apache Spark to join the data and perform the analysis.
B.Copy the data from DynamoDB, Amazon RDS, and Amazon Redshift into Amazon S3. Run Amazon Athena queries directly on the S3 files.
C.Use Amazon Athena Federated Query to join the data from all data sources.
D.Use Redshift Spectrum to query data from DynamoDB, Amazon RDS, and Amazon S3 directly from Redshift.

Answer: C

Explanation:
Athena Federated Query can query across multiple sources (S3, RDS, DynamoDB, Redshift) without moving data, making it cost-effective and serverless for a one-time analysis.

- A is incorrect because EMR adds cluster provisioning and costs for a one-time analysis.
- B is incorrect because copying large data into S3 adds extra cost and effort.
- D is incorrect because Redshift Spectrum does not natively query DynamoDB or RDS.

Question: 186
A company is planning to use a provisioned Amazon EMR cluster that runs Apache Spark jobs to perform big data analysis. The company requires high reliability. A big data team must follow best practices for running cost-optimized and long-running workloads on Amazon EMR. The team must find a solution that will maintain the company's current level of performance. Which combination of resources will meet these requirements MOST cost-effectively? (Choose two.)

Options
A.Use Hadoop Distributed File System (HDFS) as a persistent data store.
B.Use Amazon S3 as a persistent data store.
C.Use x86-based instances for core nodes and task nodes.
D.Use Graviton instances for core nodes and task nodes.
E.Use Spot Instances for all primary nodes.

Answer: BD

Explanation:
Best practice is to use Amazon S3 (with EMRFS) for persistent storage, as HDFS is tied to cluster lifecycle. Using Graviton instances reduces cost while maintaining performance.

- A is incorrect because HDFS increases cluster dependency and is less reliable.
- C is incorrect because x86 instances are generally more expensive than Graviton.
- E is incorrect because Spot instances should not be used for primary nodes (risk of termination).

Question: 187
A company wants to implement real-time analytics capabilities. The company wants to use Amazon Kinesis Data Streams and Amazon Redshift to ingest and process streaming data at the rate of several gigabytes per second. The company wants to derive near real-time insights by using existing business intelligence (BI) and analytics tools. Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Use Kinesis Data Streams to stage data in Amazon S3. Use the COPY command to load data from Amazon S3 directly into Amazon Redshift to make the data immediately available for real-time analysis.
B.Access the data from Kinesis Data Streams by using SQL queries. Create materialized views directly on top of the stream. Refresh the materialized views regularly to query the most recent stream data.
C.Create an external schema in Amazon Redshift to map the data from Kinesis Data Streams to an Amazon Redshift object. Create a materialized view to read data from the stream. Set the materialized view to auto refresh.
D.Connect Kinesis Data Streams to Amazon Kinesis Data Firehose. Use Kinesis Data Firehose to stage the data in Amazon S3. Use the COPY command to load the data from Amazon S3 to a table in Amazon Redshift.

Answer: C

Explanation:
Redshift's native streaming ingestion feature (external schema for Kinesis) allows near real-time queries with materialized views that auto-refresh, minimizing latency and operational overhead.

- A is incorrect because staging in S3 introduces delay.
- B is incorrect because SQL queries cannot directly run on Kinesis without connectors.
- D is incorrect because Firehose to S3 adds buffer-based delays, not real time.

Question: 188
A company uses an Amazon QuickSight dashboard to monitor usage of one of the company's applications. The company uses AWS Glue jobs to process data for the dashboard. The company stores the data in a single Amazon S3 bucket. The company adds new data every day. A data engineer discovers that dashboard queries are becoming slower over time. The data engineer determines that the root cause of the slowing queries is long-running AWS Glue jobs. Which actions should the data engineer take to improve the performance of the AWS Glue jobs? (Choose two.)

Options
A.Partition the data that is in the S3 bucket. Organize the data by year, month, and day.
B.Increase the AWS Glue instance size by scaling up the worker type.
C.Convert the AWS Glue schema to the DynamicFrame schema class.
D.Adjust AWS Glue job scheduling frequency so the jobs run half as many times each day.
E.Modify the IAM role that grants access to AWS glue to grant access to all S3 features.

Answer: AB

Explanation:
Partitioning the data improves query performance by scanning less data. Increasing Glue worker size improves job execution speed.

- C is incorrect because DynamicFrame vs DataFrame doesn't solve long-running jobs.
- D is incorrect because reducing job frequency does not solve performance issues.
- E is incorrect because IAM permissions don't impact Glue job runtime performance.

Question: 189
A data engineer needs to use AWS Step Functions to design an orchestration workflow. The workflow must parallel process a large collection of data files and apply a specific transformation to each file. Which Step Functions state should the data engineer use to meet these requirements?

Options
A.Parallel state
B.Choice state
C.Map state
D.Wait state

Answer: C

Explanation:
The Map state in Step Functions is designed to apply the same operation in parallel to multiple items (e.g., files). It enables parallelism while maintaining orchestration simplicity.

- A is incorrect because Parallel state runs multiple branches, not parallel iterations over a dataset.
- B is incorrect because Choice state handles branching logic, not parallel processing.
- D is incorrect because Wait state is for delays, not processing.

Question: 190
A company is migrating a legacy application to an Amazon S3 based data lake. A data engineer reviewed data that is associated with the legacy application. The data engineer found that the legacy data contained some duplicate information. The data engineer must identify and remove duplicate information from the legacy application data. Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Write a custom extract, transform, and load (ETL) job in Python. Use the DataFrame.drop_duplicates() function by importing the Pandas library to perform data deduplication.
B.Write an AWS Glue extract, transform, and load (ETL) job. Use the FindMatches machine learning (ML) transform to transform the data to perform data deduplication.
C.Write a custom extract, transform, and load (ETL) job in Python. Import the Python dedupe library. Use the dedupe library to perform data deduplication.
D.Write an AWS Glue extract, transform, and load (ETL) job. Import the Python dedupe library. Use the dedupe library to perform data deduplication.

Answer: B

Explanation:
AWS Glue FindMatches ML transform is a managed solution that automatically detects and removes duplicates with minimal coding and operational effort.

- A is incorrect because using Pandas requires managing infrastructure and scaling issues for large datasets.
- C is incorrect because dedupe library requires custom coding and infrastructure management.
- D is incorrect because using dedupe inside Glue adds unnecessary coding effort when Glue has a built-in FindMatches transform.

Question: 191
A data engineer is launching an Amazon EMR cluster. The data that the data engineer needs to load into the new cluster is currently in an Amazon S3 bucket. The data engineer needs to ensure that data is encrypted both at rest and in transit.
The data that is in the S3 bucket is encrypted by an AWS Key Management Service (AWS KMS) key. The data engineer has an Amazon S3 path that has a Privacy Enhanced Mail (PEM) file.
Which solution will meet these requirements?

Options
A. Create an Amazon EMR security configuration. Specify the appropriate AWS KMS key for at-rest encryption for the S3 bucket. Create a second security configuration. Specify the Amazon S3 path of the PEM file for in-transit encryption. Create the EMR cluster, and attach both security configurations to the cluster.
B. Create an Amazon EMR security configuration. Specify the appropriate AWS KMS key for local disk encryption for the S3 bucket. Specify the Amazon S3 path of the PEM file for in-transit encryption. Use the security configuration during EMR cluster creation.
C. Create an Amazon EMR security configuration. Specify the appropriate AWS KMS key for at-rest encryption for the S3 bucket. Specify the Amazon S3 path of the PEM file for in-transit encryption. Use the security configuration during EMR cluster creation.
D. Create an Amazon EMR security configuration. Specify the appropriate AWS KMS key for at-rest encryption for the S3 bucket. Specify the Amazon S3 path of the PEM file for in-transit encryption. Create the EMR cluster, and attach the security configuration to the cluster.

Answer: C

Explanation:
Creating a single EMR security configuration allows you to define both at-rest encryption (via KMS) and in-transit encryption (via PEM file). This meets requirements in a simple, managed way.

- A: EMR only allows one active security configuration, not multiple.
- B: Refers incorrectly to local disk encryption instead of S3 at-rest encryption.
- D: Attaching after creation is not possible; must be applied at creation.

Question: 192
A retail company is using an Amazon Redshift cluster to support real-time inventory management. The company has deployed an ML model on a real-time endpoint in Amazon SageMaker.
The company wants to make real-time inventory recommendations. The company also wants to make predictions about future inventory needs.
Which solutions will meet these requirements? (Choose two.)

Options
A. Use Amazon Redshift ML to generate inventory recommendations.
B. Use SQL to invoke a remote SageMaker endpoint for prediction.
C. Use Amazon Redshift ML to schedule regular data exports for offline model training.
D. Use SageMaker Autopilot to create inventory management dashboards in Amazon Redshift.
E. Use Amazon Redshift as a file storage system to archive old inventory management reports.

Answer: AB

Explanation:
Redshift ML (A) allows direct integration of ML inside Redshift for recommendations. Redshift's ability to invoke remote SageMaker endpoints via SQL (B) enables real-time predictions without exporting data.

- C: Regular exports are unnecessary; ML can run directly in Redshift.
- D: Autopilot creates models, not dashboards in Redshift.
- E: Redshift is not a storage system for archiving reports.

Question: 193
A company stores CSV files in an Amazon S3 bucket. A data engineer needs to process the data in the CSV files and store the processed data in a new S3 bucket.
The process needs to rename a column, remove specific columns, ignore the second row of each file, create a new column based on the values of the first row of the data, and filter the results by a numeric value of a column.
Which solution will meet these requirements with the LEAST development effort?

Options
A. Use AWS Glue Python jobs to read and transform the CSV files.
B. Use an AWS Glue custom crawler to read and transform the CSV files.
C. Use an AWS Glue workflow to build a set of jobs to crawl and transform the CSV files.
D. Use AWS Glue DataBrew recipes to read and transform the CSV files.

Answer: D

Explanation:
AWS Glue DataBrew offers no-code transformations like renaming/removing columns, ignoring rows, adding derived columns, and filtering — ideal for this scenario with minimal development effort.

- A: Glue Python jobs require custom coding.
- B: Crawlers only catalog data; they don't perform transformations.
- C: Workflows still require custom jobs to implement transformations.

Question: 194
A company uses Amazon Redshift as its data warehouse. Data encoding is applied to the existing tables of the data warehouse. A data engineer discovers that the compression encoding applied to some of the tables is not the best fit for the data.
The data engineer needs to improve the data encoding for the tables that have sub-optimal encoding.
Which solution will meet this requirement?

Options
A. Run the ANALYZE command against the identified tables. Manually update the compression encoding of columns based on the output of the command.
B. Run the ANALYZE COMPRESSION command against the identified tables. Manually update the compression encoding of columns based on the output of the command.
C. Run the VACUUM REINDEX command against the identified tables.
D. Run the VACUUM RECLUSTER command against the identified tables.

Answer: B

Explanation:
ANALYZE COMPRESSION evaluates sample data and recommends optimal encoding for each column. The engineer can then update encodings to improve storage and query performance.

- A: ANALYZE provides statistics but not compression recommendations.
- C: REINDEX is unrelated to encoding optimization.
- D: RECLUSTER reorganizes data distribution, not column encoding.

Question: 195
The company stores a large volume of customer records in Amazon S3. To comply with regulations, the company must be able to access new customer records immediately for the first 30 days after the records are created. The company accesses records that are older than 30 days infrequently.
The company needs to cost-optimize its Amazon S3 storage.
Which solution will meet these requirements MOST cost-effectively?

Options
A. Apply a lifecycle policy to transition records to S3 Standard Infrequent-Access (S3 Standard-IA) storage after 30 days.
B. Use S3 Intelligent-Tiering storage.
C. Transition records to S3 Glacier Deep Archive storage after 30 days.
D. Use S3 Standard-Infrequent Access (S3 Standard-IA) storage for all customer records.

Answer: A

Explanation:
Lifecycle policies can move objects to Standard-IA after 30 days, reducing cost while still providing millisecond access as required.

- B: Intelligent-Tiering is costlier if access patterns are already predictable.
- C: Glacier Deep Archive has hours of retrieval latency, not acceptable here.
- D: Using Standard-IA for all data is unnecessary for frequently accessed data (first 30 days).

Question: 196
A data engineer is using Amazon QuickSight to build a dashboard to report a company's revenue in multiple AWS Regions. The data engineer wants the dashboard to display the total revenue for a Region, regardless of the drilldown levels shown in the visual.
Which solution will meet these requirements?

Options
A. Create a table calculation.
B. Create a simple calculated field.
C. Create a level-aware calculation - aggregate (LAC-A) function.
D. Create a level-aware calculation - window (LAC-W) function.

Answer: C

Explanation:
Level-aware aggregation (LAC-A) ensures that metrics like revenue are always calculated at a fixed level (e.g., Region), independent of drilldown levels in visuals.

- A: Table calculations depend on visible fields and drilldowns.
- B: Simple calculated fields change with drilldowns.
- D: Window functions calculate across partitions, not fixed drill levels.

Question: 197
A retail company stores customer data in an Amazon S3 bucket. Some of the customer data contains personally identifiable information (PII) about customers. The company must not share PII data with business partners.
A data engineer must determine whether a dataset contains PII before making objects in the dataset available to business partners.
Which solution will meet this requirement with the LEAST manual intervention?

Options
A. Configure the S3 bucket and S3 objects to allow access to Amazon Macie. Use automated sensitive data discovery in Macie.
B. Configure AWS CloudTrail to monitor S3 PUT operations. Inspect the CloudTrail trails to identify operations that save PII.
C. Create an AWS Lambda function to identify PII in S3 objects. Schedule the function to run periodically.
D. Create a table in AWS Glue Data Catalog. Write custom SQL queries to identify PII in the table. Use Amazon Athena to run the queries.

Answer: A

Explanation:
Amazon Macie provides automated PII discovery and classification in S3 with minimal manual work.

- B: CloudTrail monitors operations, not content.
- C: Custom Lambda requires ongoing maintenance.
- D: Glue + Athena queries require manual setup and ongoing checks.

Question: 198
A data engineer needs to create an empty copy of an existing table in Amazon Athena to perform data processing tasks. The existing table in Athena contains 1,000 rows.
Which query will meet this requirement?

Options
A. CREATE TABLE new_table LIKE old_table;
B. CREATE TABLE new_table AS SELECT * FROM old_table WITH NO DATA;
C. CREATE TABLE new_table AS SELECT * FROM old_table;
D. CREATE TABLE new_table as SELECT * FROM old_table WHERE 1=1;

Answer: B

Explanation:
CREATE TABLE ... AS SELECT ... WITH NO DATA creates a new table with the schema only, no rows, which matches the requirement.

- A: Syntax not valid for Athena.
- C: Would copy both schema and data.
- D: WHERE 1=1 would still copy all rows.

Question: 199
A company has a data lake in Amazon S3. The company collects AWS CloudTrail logs for multiple applications. The company stores the logs in the data lake, catalogs the logs in AWS Glue, and partitions the logs based on the year. The company uses Amazon Athena to analyze the logs.
Recently, customers reported that a query on one of the Athena tables did not return any data. A data engineer must resolve the issue.
Which combination of troubleshooting steps should the data engineer take? (Choose two.)

Options
A. Confirm that Athena is pointing to the correct Amazon S3 location.
B. Increase the query timeout duration.
C. Use the MSCK REPAIR TABLE command.
D. Restart Athena.
E. Delete and recreate the problematic Athena table.

Answer: AC

Explanation:
Athena queries may fail if the table points to the wrong S3 location (A). Also, when new partitions are added in S3, they must be loaded into Athena with MSCK REPAIR TABLE (C).

- B: Timeout won't fix missing data.
- D: Athena is serverless; restarting doesn't apply.
- E: Recreating the table is unnecessary and high effort.

Question: 200
A data engineer wants to orchestrate a set of extract, transform, and load (ETL) jobs that run on AWS. The ETL jobs contain tasks that must run Apache Spark jobs on Amazon EMR, make API calls to Salesforce, and load data into Amazon Redshift.
The ETL jobs need to handle failures and retries automatically. The data engineer needs to use Python to orchestrate the jobs.
Which service will meet these requirements?

Options
A. Amazon Managed Workflows for Apache Airflow (Amazon MWAA)
B. AWS Step Functions
C. AWS Glue
D. Amazon EventBridge

Answer: A

Explanation:
Amazon MWAA supports Python DAGs, integrates with EMR, Redshift, and external APIs, and provides built-in retry/failure handling. Ideal for complex ETL orchestration.

- B: Step Functions orchestration is JSON/YAML, not Python-based.
- C: Glue focuses on ETL execution, not orchestration across diverse services.
- D: EventBridge schedules/triggers events, not full workflow orchestration.

Question: 201
A company runs multiple applications on AWS. The company configured each application to output logs. The company wants to query and visualize the application logs in near real time.
Which solution will meet these requirements?

Options
A. Configure the applications to output logs to Amazon CloudWatch Logs log groups. Create an Amazon S3 bucket. Create an AWS Lambda function that runs on a schedule to export the required log groups to the S3 bucket. Use Amazon Athena to query the log data in the S3 bucket.
B. Create an Amazon OpenSearch Service domain. Configure the applications to output logs to Amazon CloudWatch Logs log groups. Create an OpenSearch Service subscription filter for each log group to stream the data to OpenSearch. Create the required queries and dashboards in OpenSearch Service to analyze and visualize the data.
C. Configure the applications to output logs to Amazon CloudWatch Logs log groups. Use CloudWatch log anomaly detection to query and visualize the log data.
D. Update the application code to send the log data to Amazon QuickSight by using Super-fast, Parallel, In-memory Calculation Engine (SPICE). Create the required analyses and dashboards in QuickSight.

Answer: B

Explanation:
OpenSearch integrates natively with CloudWatch Logs via subscription filters, supports real-time indexing, querying, and dashboards.

- A: Athena queries on exported S3 data are batch-based, not near real-time.
- C: CloudWatch anomaly detection is for metrics, not querying raw logs.
- D: QuickSight is not designed for direct log ingestion.

Question: 202
An ecommerce company processes millions of orders each day. The company uses AWS Glue ETL to collect data from multiple sources, clean the data, and store the data in an Amazon S3 bucket in CSV format by using the S3 Standard storage class. The company uses the stored data to conduct daily analysis.
The company wants to optimize costs for data storage and retrieval.
Which solution will meet this requirement?

Options
A. Transition the data to Amazon S3 Glacier Flexible Retrieval.
B. Transition the data from Amazon S3 to an Amazon Aurora cluster.
C. Configure AWS Glue ETL to transform the incoming data to Apache Parquet format.
D. Configure AWS Glue ETL to use Amazon EMR to process incoming data in parallel.

Answer: C

Explanation:
Parquet is columnar and highly compressed, reducing storage and query scan costs.

- A: Glacier is for archival, not daily analysis.
- B: Aurora is transactional DB, not for analytics.
- D: EMR doesn't reduce storage/query costs.

Question: 203
A data engineer is optimizing query performance in Amazon Athena notebooks that use Apache Spark to analyze large datasets that are stored in Amazon S3. The data is partitioned.
An AWS Glue crawler updates the partitions.
The data engineer wants to minimize the amount of data that is scanned to improve efficiency of Athena queries.
Which solution will meet these requirements?

Options
A. Apply partition filters in the queries.
B. Increase the frequency of AWS Glue crawler invocations to update the data catalog more often.
C. Organize the data that is in Amazon S3 by using a nested directory structure.
D. Configure Spark to use in-memory caching for frequently accessed data.

Answer: A

Explanation:
Partition pruning ensures Athena scans only relevant partitions, minimizing scanned data.

- B: More frequent crawls don't reduce scan size.
- C: Already partitioned; nested directories help but pruning in queries is key.
- D: Caching helps repeated queries but doesn't reduce scan size.

Question: 204
A company manages an Amazon Redshift data warehouse. The data warehouse is in a public subnet inside a custom VPC. A security group allows only traffic from within itself. An ACL is open to all traffic.
The company wants to generate several visualizations in Amazon QuickSight for an upcoming sales event. The company will run QuickSight Enterprise edition in a second AWS account inside a public subnet within a second custom VPC. The new public subnet has a security group that allows outbound traffic to the existing Redshift cluster.
A data engineer needs to establish connections between Amazon Redshift and QuickSight. QuickSight must refresh dashboards by querying the Redshift cluster.
Which solution will meet these requirements?

Options
A. Configure the Redshift security group to allow inbound traffic on the Redshift port from the QuickSight security group.
B. Assign Elastic IP addresses to the QuickSight visualizations. Configure the QuickSight security group to allow inbound traffic on the Redshift port from the Elastic IP addresses.
C. Confirm that the CIDR ranges of the Redshift VPC and the QuickSight VPC are the same. If CIDR ranges are different, reconfigure one CIDR range to match the other. Establish network peering between the VPCs.
D. Create a QuickSight gateway endpoint in the Redshift VPC. Attach an endpoint policy to the gateway endpoint to ensure only specific QuickSight accounts can use the endpoint.

Answer: A

Explanation:
Allowing inbound access from QuickSight's security group to Redshift's port enables cross-account access.

- B: Elastic IPs don't apply to QuickSight managed service.
- C: VPC peering helps only if SG rules allow; not required.
- D: No QuickSight gateway endpoint exists.

Question: 205
A data engineer is building a data pipeline. A large data file is uploaded to an Amazon S3 bucket once each day at unpredictable times. An AWS Glue workflow uses hundreds of workers to process the file and load the data into Amazon Redshift. The company wants to process the file as quickly as possible.
Which solution will meet these requirements?

Options
A. Create an on-demand AWS Glue trigger to start the workflow. Create an AWS Lambda function that runs every 15 minutes to check the S3 bucket for the daily file. Configure the function to start the AWS Glue workflow if the file is present.
B. Create an event-based AWS Glue trigger to start the workflow. Configure Amazon S3 to log events to AWS CloudTrail. Create a rule in Amazon EventBridge to forward PutObject events to the AWS Glue trigger.
C. Create a scheduled AWS Glue trigger to start the workflow. Create a cron job that runs the AWS Glue job every 15 minutes. Set up the AWS Glue job to check the S3 bucket for the daily file. Configure the job to stop if the file is not present.
D. Create an on-demand AWS Glue trigger to start the workflow. Create an AWS Database Migration Service (AWS DMS) migration task. Set the DMS source as the S3 bucket. Set the target endpoint as the AWS Glue workflow.

Answer: B

Explanation:
S3 event + EventBridge + Glue trigger ensures immediate, event-driven processing.

- A, C: Polling increases latency and overhead.
- D: DMS is for database replication, not ETL orchestration.

Question: 206
A data engineer needs to run a data transformation job whenever a user adds a file to an Amazon S3 bucket. The job will run for less than 1 minute. The job must send the output through an email message to the data engineer. The data engineer expects users to add one file every hour of the day.
Which solution will meet these requirements in the MOST operationally efficient way?

Options
A. Create a small Amazon EC2 instance that polls the S3 bucket for new files. Run transformation code on a schedule to generate the output. Use operating system commands to send email messages.
B. Run an Amazon Elastic Container Service (Amazon ECS) task to poll the S3 bucket for new files. Run transformation code on a schedule to generate the output. Use operating system commands to send email messages.
C. Create an AWS Lambda function to transform the data. Use Amazon S3 Event Notifications to invoke the Lambda function when a new object is created. Publish the output to an Amazon Simple Notification Service (Amazon SNS) topic. Subscribe the data engineer's email account to the topic.
D. Deploy an Amazon EMR cluster. Use EMR File System (EMRFS) to access the files in the S3 bucket. Run transformation code on a schedule to generate the output to a second S3 bucket. Create an Amazon Simple Notification Service (Amazon SNS) topic. Configure Amazon S3 Event Notifications to notify the topic when a new object is created.

Answer: C

Explanation:
Lambda + S3 Event + SNS email is serverless, fast, and low-cost.

- A, B: Polling adds unnecessary cost and management.
- D: EMR is overkill for <1 minute job.

Question: 207
A company uses Amazon S3 and AWS Glue Data Catalog to manage a data lake that contains contact information for customers. The company uses PySpark and AWS Glue jobs with a DynamicFrame to run a workflow that processes data within the data lake.
A data engineer notices that the workflow is generating errors as a result of how customer postal codes are stored in the data lake. Some postal codes include unnecessary numbers or invalid characters.
The data engineer needs a solution to address the errors and correct the postal codes in the data lake.

Options
A. Create a schema definition for PySpark that matches the format the processing workflow requires for postal codes. Pass the schema to the DynamicFrame during processing.
B. Use AWS Glue workflow properties to allow job state sharing. Configure the AWS Glue jobs to read values from the postal code column by using the properties from a previously successful run of the jobs.
C. Configure the column.push_down_predicate setting and the catalogPartitionPredicate settings for the postal code column in the DynamicFrame.
D. Set the DynamicFrame additional_options parameter 'useS3ListImplementation' to True.

Answer: A

Explanation:
Explicit schema enforcement ensures postal codes follow the correct format, fixing invalid values.

- B: Workflow properties don't fix data errors.
- C: Predicate pushdown is for filtering, not data correction.
- D: S3ListImplementation is irrelevant here.

Question: 208
A data engineer is troubleshooting an AWS Glue workflow that occasionally fails. The engineer determines that the failures are a result of data quality issues. A business reporting team needs to receive an email notification any time the workflow fails in the future.
Which solution will meet this requirement?

Options
A. Create an Amazon Simple Notification Service (Amazon SNS) FIFO topic. Subscribe the team's email account to the SNS topic. Create an AWS Lambda function that initiates when the AWS Glue job state changes to FAILED. Set the SNS topic as the target.
B. Create an Amazon Simple Notification Service (Amazon SNS) standard topic. Subscribe the team's email account to the SNS topic. Create an Amazon EventBridge rule that triggers when the AWS Glue job state changes to FAILED. Set the SNS topic as the target.
C. Create an Amazon Simple Queue Service (Amazon SQS) FIFO queue. Subscribe the team's email account to the SQS queue. Create an AWS Config rule that triggers when the AWS Glue job state changes to FAILED. Set the SQS queue as the target.
D. Create an Amazon Simple Queue Service (Amazon SQS) standard queue. Subscribe the team's email account to the SQS queue. Create an Amazon EventBridge rule that triggers when the AWS Glue job state changes to FAILED. Set the SQS queue as the target.

Answer: B

Explanation:
EventBridge directly captures Glue job state change events and routes to SNS for email.

- A: FIFO SNS not needed.
- C, D: SQS not required for simple email alerts.

Question: 209
A company uses AWS Glue jobs to implement several data pipelines. The pipelines are critical to the company. The company needs to implement a monitoring mechanism that will alert stakeholders if the pipelines fail.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A. Create an Amazon EventBridge rule to match AWS Glue job failure events. Configure the rule to target an AWS Lambda function to process events. Configure the function to send notifications to an Amazon Simple Notification Service (Amazon SNS) topic.
B. Configure an Amazon CloudWatch Logs log group for the AWS Glue jobs. Create an Amazon EventBridge rule to match new log creation events in the log group. Configure the rule to target an AWS Lambda function that reads the logs and sends notifications to an Amazon Simple Notification Service (Amazon SNS) topic if AWS Glue job failure logs are present.
C. Create an Amazon EventBridge rule to match AWS Glue job failure events. Define an Amazon CloudWatch metric based on the EventBridge rule. Set up a CloudWatch alarm based on the metric to send notifications to an Amazon Simple Notification Service (Amazon SNS) topic.
D. Configure an Amazon CloudWatch Logs log group for the AWS Glue jobs. Create an Amazon EventBridge rule to match new log creation events in the log group. Configure the rule to send notifications to an Amazon Simple Notification Service (Amazon SNS) topic.

Answer: C

Explanation:
EventBridge → CloudWatch metric → CloudWatch alarm → SNS ensures automated monitoring with minimal overhead.

- A: Requires extra Lambda unnecessarily.
- B, D: Log parsing is heavier and unnecessary.

Question: 210
A company uses AWS Glue Apache Spark jobs to handle extract, transform, and load (ETL) workloads. The company has enabled logging and monitoring for all AWS Glue jobs.
One of the AWS Glue jobs begins to fail. A data engineer investigates the error and wants to examine metrics for all individual stages within the job.
How can the data engineer access the stage metrics?

Options
A. Examine the AWS Glue job and stage details in the Spark UI.
B. Examine the AWS Glue job and stage metrics in Amazon CloudWatch.
C. Examine the AWS Glue job and stage logs in AWS CloudTrail logs.
D. Examine the AWS Glue job and stage details by using the run insights feature on the job.

Answer: A

Explanation:
Spark UI provides detailed stage-level metrics (tasks, shuffle, memory usage) for Glue Spark jobs.

- B: CloudWatch shows overall job metrics, not stage-level detail.
- C: CloudTrail tracks API calls, not Spark execution details.
- D: Run insights provides aggregated Glue job info, not stage metrics.

Question: 211
A data engineer is processing and analyzing multiple terabytes of raw data that is in Amazon S3. The data engineer needs to clean and prepare the data. Then the data engineer needs to load the data into Amazon Redshift for analytics.
The data engineer needs a solution that will give data analysts the ability to perform complex queries. The solution must eliminate the need to perform complex extract, transform, and load (ETL) processes or to manage infrastructure.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Use Amazon EMR to prepare the data. Use AWS Step Functions to load the data into Amazon Redshift. Use Amazon QuickSight to run queries.
B.Use AWS Glue DataBrew to prepare the data. Use AWS Glue to load the data into Amazon Redshift. Use Amazon Redshift to run queries.
C.Use AWS Lambda to prepare the data. Use Amazon Kinesis Data Firehose to load the data into Amazon Redshift. Use Amazon Athena to run queries.
D.Use AWS Glue to prepare the data. Use AWS Database Migration Service (AVVS DMS) to load the data into Amazon Redshift. Use Amazon Redshift Spectrum to run queries.

Answer: B

Explanation:
Glue DataBrew is serverless and provides a no-code way to clean and prepare raw data. Glue then loads data into Redshift, where analysts can run complex queries. This eliminates heavy ETL coding and infrastructure management.

- A: EMR requires cluster management and is overkill compared to Glue DataBrew.
- C: Lambda and Firehose are not suited for bulk ETL at terabyte scale. Athena is not the target since Redshift is required.
- D: DMS is for database replication, not for cleaning and preparing large S3 datasets.

Question: 212
A company uses an AWS Lambda function to transfer files from a legacy SFTP environment to Amazon S3 buckets. The Lambda function is VPC enabled to ensure that all communications between the Lambda function and other AVS services that are in the same VPC environment will occur over a secure network.
The Lambda function is able to connect to the SFTP environment successfully. However, when the Lambda function attempts to upload files to the S3 buckets, the Lambda function returns timeout errors. A data engineer must resolve the timeout issues in a secure way.
Which solution will meet these requirements in the MOST cost-effective way?

Options
A.Create a NAT gateway in the public subnet of the VPC. Route network traffic to the NAT gateway.
B.Create a VPC gateway endpoint for Amazon S3. Route network traffic to the VPC gateway endpoint.
C.Create a VPC interface endpoint for Amazon S3. Route network traffic to the VPC interface endpoint.
D.Use a VPC internet gateway to connect to the internet. Route network traffic to the VPC internet gateway.

Answer: B

Explanation:
A VPC gateway endpoint allows private, secure communication from VPC resources to S3 without internet or NAT. It is the most cost-effective and secure way.

- A: NAT gateway allows internet access, but it adds costs and is unnecessary for S3.
- C: Interface endpoints for S3 exist but are less cost-effective than gateway endpoints.
- D: Internet gateway would force traffic over the internet, which is less secure and not required.

Question: 213
A company reads data from customer databases that run on Amazon RDS. The databases contain many inconsistent fields. For example, a customer record field that is named place_id in one database is named location_id in another database. The company needs to link customer records across different databases, even when customer record fields do not match. Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Create a provisioned Amazon EMR cluster to process and analyze data in the databases. Connect to the Apache Zeppelin notebook. Use the FindMatches transform to find duplicate records in the data.
B.Create an AWS Glue crawler to craw the databases. Use the FindMatches transform to find duplicate records in the data. Evaluate and tune the transform by evaluating the performance and results.
C.Create an AWS Glue crawler to craw the databases. Use Amazon SageMaker to construct Apache Spark ML pipelines to find duplicate records in the data.
D.Create a provisioned Amazon EMR cluster to process and analyze data in the databases. Connect to the Apache Zeppelin notebook. Use an Apache Spark ML model to find duplicate records in the data. Evaluate and tune the model by evaluating the performance and results.

Answer: B

Explanation:
Glue crawlers catalog the databases, and Glue FindMatches is purpose-built for deduplication and fuzzy matching with minimal setup. This avoids custom ML.

- A: EMR requires cluster management and more coding effort.
- C: SageMaker ML pipelines require building custom models, adding overhead.
- D: Spark ML models require heavy customization and cluster management.

Question: 214
A finance company receives data from third-party data providers and stores the data as objects in an Amazon S3 bucket.
The company ran an AWS Glue crawler on the objects to create a data catalog. The AWS Glue crawler created multiple tables. However, the company expected that the crawler would create only one table.
The company needs a solution that will ensure the AVS Glue crawler creates only one table.
Which combination of solutions will meet this requirement? (Choose two.)

Options
A.Ensure that the object format, compression type, and schema are the same for each object.
B.Ensure that the object format and schema are the same for each object. Do not enforce consistency for the compression type of each object.
C.Ensure that the schema is the same for each object. Do not enforce consistency for the file format and compression type of each object.
D.Ensure that the structure of the prefix for each S3 object name is consistent.
E.Ensure that all S3 object names follow a similar pattern.

Answer: AD

Explanation:
Glue crawler creates separate tables when file formats, compression, or schemas differ, or when prefixes look like separate datasets. Keeping consistent schema, format, and compression (A) plus a uniform prefix structure (D) ensures one table.

- B: Ignoring compression consistency may still lead to multiple tables.
- C: File format inconsistencies would break unification.
- E: Naming pattern alone doesn't guarantee schema and format consistency.

Question: 215
An application consumes messages from an Amazon Simple Queue Service (Amazon SQS) queue. The application experiences occasional downtime. As a result of the downtime, messages within the queue expire and are deleted after 1 day. The message deletions cause data loss for the application.
Which solutions will minimize data loss for the application? (Choose two.)

Options
A.Increase the message retention period
B.Increase the visibility timeout.
C.Attach a dead-letter queue (DLQ) to the SQS queue.
D.Use a delay queue to delay message delivery
E.Reduce message processing time.

Answer: AC

Explanation:
Increasing message retention (A) ensures messages remain in the queue longer during downtime. Attaching a DLQ (C) captures failed messages for later processing, preventing data loss.

- B: Visibility timeout affects in-progress processing, not expired messages.
- D: Delay queues postpone delivery but don't help during downtime.
- E: Reducing processing time doesn't prevent loss when the app is down.

Question: 216
A company is creating near real-time dashboards to visualize time series data. The company ingests data into Amazon Managed Streaming for Apache Kafka (Amazon MSK). A customized data pipeline consumes the data. The pipeline then writes data to Amazon Keyspaces (for Apache Cassandra), Amazon OpenSearch Service, and Apache Avro objects in Amazon S3.
Which solution will make the data available for the data visualizations with the LEAST latency?

Options
A.Create OpenSearch Dashboards by using the data from OpenSearch Service.
B.Use Amazon Athena with an Apache Hive metastore to query the Avro objects in Amazon S3. Use Amazon Managed Grafana to connect to Athena and to create the dashboards.
C.Use Amazon Athena to query the data from the Avro objects in Amazon S3. Configure Amazon Keyspaces as the data catalog. Connect Amazon QuickSight to Athena to create the dashboards.
D.Use AWS Glue to catalog the data. Use S3 Select to query the Avro objects in Amazon S3. Connect Amazon QuickSight to the S3 bucket to create the dashboards.

Answer: A

Explanation:
OpenSearch indexes provide sub-second search and analytics. Using OpenSearch Dashboards directly offers the lowest latency for visualizations.

- B: Athena queries S3 data in batch mode, which has higher latency.
- C: Athena with Keyspaces as catalog is not supported and still batch.
- D: S3 Select is not designed for real-time visualization at scale.

Question: 217
A company stores petabytes of data in thousands of Amazon S3 buckets in the S3 Standard storage class. The data supports analytics workloads that have unpredictable and variable data access patterns.
The company does not access some data for months. However, the company must be able to retrieve all data within milliseconds. The company needs to optimize S3 storage costs.
Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Use S3 Storage Lens standard metrics to determine when to move objects to more cost-optimized storage classes. Create S3 Lifecycle policies for the S3 buckets to move objects to cost-optimized storage classes. Continue to refine the S3 Lifecycle policies in the future to optimize storage costs.
B.Use S3 Storage Lens activity metrics to identify S3 buckets that the company accesses infrequently. Configure S3 Lifecycle rules to move objects from S3 Standard to the S3 Standard-Infrequent Access (S3 Standard-IA) and S3 Glacier storage classes based on the age of the data.
C.Use S3 Intelligent-Tiering. Activate the Deep Archive Access tier.
D.Use S3 Intelligent-Tiering. Use the default access tier.

Answer: D

Explanation:
S3 Intelligent-Tiering automatically moves data between frequent and infrequent access tiers with no operational overhead. Retrieval remains milliseconds across tiers.

- A: Manual lifecycle policies require ongoing management.
- B: Glacier storage does not provide millisecond retrieval.
- C: Deep Archive has retrieval delays, not suitable here.

Question: 218
A media company wants to use Amazon OpenSearch Service to analyze rea-time data about popular musical artists and songs. The company expects to ingest millions of new data events every day. The new data events will arrive through an Amazon Kinesis data stream. The company must transform the data and then ingest the data into the OpenSearch Service domain.
Which method should the company use to ingest the data with the LEAST operational overhead?

Options
A.Use Amazon Kinesis Data Firehose and an AWS Lambda function to transform the data and deliver the transformed data to OpenSearch Service.
B.Use a Logstash pipeline that has prebuilt filters to transform the data and deliver the transformed data to OpenSearch Service.
C.Use an AWS Lambda function to call the Amazon Kinesis Agent to transform the data and deliver the transformed data OpenSearch Service.
D.Use the Kinesis Client Library (KCL) to transform the data and deliver the transformed data to OpenSearch Service.

Answer: A

Explanation:
Kinesis Data Firehose natively integrates with OpenSearch. Adding a Lambda transform in the delivery stream is managed and serverless, minimizing operational overhead.

- B: Logstash requires cluster setup and management.
- C: Kinesis Agent is for file-based ingestion, not for streaming.
- D: KCL requires building and running custom consumers.

Question: 219
A company stores customer data tables that include customer addresses in an AWS Lake Formation data lake. To comply with new regulations, the company must ensure that users cannot access data for customers who are in Canada.
The company needs a solution that will prevent user access to rows for customers who are in Canada.
Which solution will meet this requirement with the LEAST operational effort?

Options
A.Set a row-level filter to prevent user access to a row where the country is Canada.
B.Create an IAM role that restricts user access to an address where the country is Canada.
C.Set a column-level filter to prevent user access to a row where the country is Canada.
D.Apply a tag to all rows where Canada is the country. Prevent user access where the tag is equal to “Canada”.

Answer: A

Explanation:
Lake Formation supports row-level filtering, which is the simplest and most direct way to block access to rows where country = Canada.

- B: IAM roles cannot apply fine-grained row-level restrictions.
- C: Column-level filters restrict entire columns, not rows.
- D: Row tagging at scale adds complexity and requires custom enforcement.

Question: 220
A company has implemented a lake house architecture in Amazon Redshift. The company needs to give users the ability to authenticate into Redshift query editor by using a third-party identity provider (IdP).
A data engineer must set up the authentication mechanism.
What is the first step the data engineer should take to meet this requirement?

Options
A.Register the third-party IdP as an identity provider in the configuration settings of the Redshift cluster.
B.Register the third-party IdP as an identity provider from within Amazon Redshift.
C.Register the third-party IdP as an identity provider for AVS Secrets Manager. Configure Amazon Redshift to use Secrets Manager to manage user credentials.
D.Register the third-party IdP as an identity provider for AWS Certificate Manager (ACM). Configure Amazon Redshift to use ACM to manage user credentials.

Answer: A

Explanation:
Redshift supports authentication via federated SAML-based identity providers. The first step is registering the IdP in the Redshift cluster configuration.

- B: There is no direct in-console configuration from within Redshift itself.
- C: Secrets Manager manages credentials, not IdP integration.
- D: ACM handles SSL/TLS certificates, not identity providers.

Question: 221
A company uses Amazon Redshift as its data warehouse service. A data engineer needs to design a physical data model. The data engineer encounters a de-normalized table that is growing in size. The table does not have a suitable column to use as the distribution key.
Which distribution style should the data engineer use to meet these requirements with the LEAST maintenance overhead?
A.ALL distribution
B.EVEN distribution
C.AUTO distribution
D.KEY distribution
Answer: C

Explanation:
AUTO distribution lets Redshift choose and evolve the optimal distribution style as the table grows, minimizing manual tuning and maintenance overhead.

- A: ALL replicates the table to every node; high storage/network cost and suited only for very small dimension tables.
- B: EVEN spreads rows round-robin but doesn't adapt over time; more management than AUTO.
- D: KEY needs a good dist key column, which the table lacks in this scenario.

Question: 222
A retail company is expanding its operations globally. The company needs to use Amazon QuickSight to accurately
calculate currency exchange rates for financial reports. The company has an existing dashboard that includes a
visual that is based on an analysis of a dataset that contains global currency values and exchange rates.
A data engineer needs to ensure that exchange rates are calculated with a precision of four decimal places. The
calculations must be precomputed. The data engineer must materialize results in QuickSight super-fast, parallel,
in-memory calculation engine (SPICE).
Which solution will meet these requirements?
A.Define and create the calculated field in the dataset.
B.Define and create the calculated field in the analysis.
C.Define and create the calculated field in the visual.
D.Define and create the calculated field in the dashboard.
Answer: A

Explanation:
Dataset-level calculated fields are materialized in SPICE during dataset refresh, ensuring precomputed results with the required precision for all analyses/visuals using that dataset.

- B: Analysis-level calcs are computed at analysis time, not precomputed into SPICE.
- C: Visual-level calcs are evaluated at render time, not precomputed.
- D: Dashboard does not introduce new calc materialization; it reflects the analysis.

Question: 223
A company has three subsidiaries. Each subsidiary uses a different data warehousing solution. The first subsidiary hosts its data warehouse in Amazon Redshift. The second subsidiary uses Teradata Vantage on AWS. The third subsidiary uses Google BigQuery.
The company wants to aggregate all the data into a central Amazon S3 data lake. The company wants to use Apache Iceberg as the table format. A data engineer needs to build a new pipeline to connect to all the data sources, run transformations by using each source engine, join the data, and write the data to Iceberg.
Which solution will meet these requirements with the LEAST operational effort?
A.Use native Amazon Redshift, Teradata, and BigQuery connectors to build the pipeline in AWS Glue. Use native
AWS Glue transforms to join the data. Run a Merge operation on the data lake Iceberg table.
B.Use the Amazon Athena federated query connectors for Amazon Redshift, Teradata, and BigQuery to build
the pipeline in Athena. Write a SQL query to read from all the data sources, join the data, and run a Merge
operation on the data lake Iceberg table.
C.Use the native Amazon Redshift connector, the Java Database Connectivity (JDBC) connector for Teradata,
and the open source Apache Spark BigQuery connector to build the pipeline in Amazon EMR. Write code in
PySpark to join the data. Run a Merge operation on the data lake Iceberg table.
D.Use the native Amazon Redshift, Teradata, and BigQuery connectors in Amazon Appflow to write data to
Amazon S3 and AWS Glue Data Catalog. Use Amazon Athena to join the data. Run a Merge operation on the
data lake Iceberg table.
Answer: A

Explanation:
AWS Glue provides managed connectors and built-in transforms to integrate multiple warehouses with minimal code, and supports writing/merging into Iceberg tables on S3.

- B: Athena federations may not cover all needed transforms/merge semantics easily across sources.
- C: EMR requires cluster management and more coding effort in PySpark.
- D: AppFlow targets app/SaaS data movement; not ideal for complex joins/merges across warehouses.

Question: 224
A company is building a data stream processing application. The application runs in an Amazon Elastic Kubernetes Service (Amazon EKS) cluster. The application stores processed data in an Amazon DynamoDB table. The company needs the application containers in the EKS cluster to have secure access to the DynamoDB table. The company does not want to embed AWS credentials in the containers. Which solution will meet these requirements?

A.Store the AWS credentials in an Amazon S3 bucket. Grant the EKS containers access to the S3 bucket to
retrieve the credentials.
B.Attach an IAM role to the EKS worker nodes, Grant the IAM role access to DynamoDUse the IAM role to set up
IAM roles service accounts (IRSA) functionality.
C.Create an IAM user that has an access key to access the DynamoDB table. Use environment variables in the
EKS containers to store the IAM user access key data.
D.Create an IAM user that has an access key to access the DynamoDB table. Use Kubernetes secrets that are
mounted in a volume of the EKS duster nodes to store the user access key data.
Answer: B

Explanation:
Using IAM Roles for Service Accounts (IRSA) assigns fine-grained IAM permissions to pods via OpenID Connect without embedding credentials, enabling secure DynamoDB access.

- A: Storing credentials in S3 is insecure and still embeds secrets.
- C: IAM user keys in environment variables are long-lived secrets and insecure.
- D: Kubernetes secrets still store static credentials; higher risk and management overhead.

Question: 225
A data engineer needs to onboard a new data producer into AWS. The data producer needs to migrate data products to AWS. The data producer maintains many data pipelines that support a business application. Each pipeline must have service accounts and their corresponding credentials. The data engineer must establish a secure connection from
the data producer's on-premises data center to AWS. The data engineer must not use the public internet to transfer data from an on-premises data center to AWS.

Which solution will meet these requirements?
A.Instruct the new data producer to create Amazon Machine Images (AMIs) on Amazon Elastic Container
Service (Amazon ECS) to store the code base of the application. Create security groups in a public subnet that
allow connections only to the on-premises data center.
B.Create an AWS Direct Connect connection to the on-premises data center. Store the service account
credentials in AWS Secrets manager.
C.Create a security group in a public subnet. Configure the security group to allow only connections from the
CIDR blocks that correspond to the data producer. Create Amazon S3 buckets than contain presigned URLS
that have one-day expiration dates.
D.Create an AWS Direct Connect connection to the on-premises data center. Store the application keys in AWS
Secrets Manager. Create Amazon S3 buckets that contain presigned URLS that have one-day expiration dates.
Answer: B

Explanation:
AWS Direct Connect provides private connectivity that avoids the public internet. Storing service account credentials in AWS Secrets Manager centralizes and secures secret management.

- A: Uses public subnets and does not provide private network connectivity.
- C: Presigned URLs traverse the public internet and don’t meet the constraint.
- D: Mixing presigned URLs still uses public internet, violating requirements.

Question: 226
A data engineer configured an AWS Glue Data Catalog for data that is stored in Amazon S3 buckets. The data engineer needs to configure the Data Catalog to receive incremental updates. The data engineer sets up event notifications for the S3 bucket and creates an Amazon Simple Queue Service (Amazon SQS) queue to receive the S3 events.
Which combination of steps should the data engineer take to meet these requirements with LEAST operational overhead? (Choose two.)

A.Create an S3 event-based AWS Glue crawler to consume events from the SQS queue.
B.Define a time-based schedule to run the AWS Glue crawler, and perform incremental updates to the Data
Catalog.
C.Use an AWS Lambda function to directly update the Data Catalog based on S3 events that the SQS queue
receives.
D.Manually initiate the AWS Glue crawler to perform updates to the Data Catalog when there is a change in the
S3 bucket.
E.Use AWS Step Functions to orchestrate the process of updating the Data Catalog based on S3 events that
the SQS queue receives.
Answer: BC

Explanation:
A scheduled Glue crawler (B) can perform incremental updates with low ops effort, and a lightweight Lambda (C) can react to SQS events to keep the catalog current without orchestration overhead.

- A: Event-based crawler via SQS isn't a native Glue feature; adds complexity.
- D: Manual runs increase operational burden.
- E: Step Functions orchestration is unnecessary for simple catalog updates.

Question: 227
A company uses AWS Glue Data Catalog to index data that is uploaded to an Amazon S3 bucket every day. The company uses a daily batch process in an extract, transform, and load (ETL) pipeline to upload data from external sources into the S3 bucket.

The company runs a daily report on the S3 data. Some days, the company runs the report before all the daily data has been uploaded to the S3 bucket. A data engineer must be able to send a message that identifies any incomplete data to an existing Amazon Simple Notification Service (Amazon SNS) topic.

Which solution will meet this requirement with the LEAST operational overhead?

Options
A. Create data quality checks for the source datasets that the daily reports use. Create a new AWS managed Apache Airflow cluster. Run the data quality checks by using Airflow tasks that run data quality queries on the columns data type and the presence of null values. Configure Airflow Directed Acyclic Graphs (DAGs) to send an email notification that informs the data engineer about the incomplete datasets to the SNS topic.
B. Create data quality checks on the source datasets that the daily reports use. Create a new Amazon EMR cluster. Use Apache Spark SQL to create Apache Spark jobs in the EMR cluster that run data quality queries on the columns data type and the presence of null values. Orchestrate the ETL pipeline by using an AWS Step Functions workflow. Configure the workflow to send an email notification that informs the data engineer about the incomplete datasets to the SNS topic.
C. Create data quality checks on the source datasets that the daily reports use. Create data quality actions by using AWS Glue workflows to confirm the completeness and consistency of the datasets. Configure the data quality actions to create an event in Amazon EventBridge if a dataset is incomplete. Configure EventBridge to send the event that informs the data engineer about the incomplete datasets to the Amazon SNS topic.
D. Create AWS Lambda functions that run data quality queries on the columns data type and the presence of null values. Orchestrate the ETL pipeline by using an AWS Step Functions workflow that runs the Lambda functions. Configure the Step Functions workflow to send an email notification that informs the data engineer about the incomplete datasets to the SNS topic.

Answer: C

Explanation:
Glue Data Quality integrated with Glue workflows (C) provides built-in checks and actions. EventBridge can route failures directly to SNS with minimal infrastructure and ops.

- A: Managed Airflow adds extra service and orchestration overhead.
- B: EMR + Step Functions is heavier to operate for simple completeness checks.
- D: Custom Lambda + Step Functions requires more code and orchestration.

Question: 228
A company stores customer data that contains personally identifiable information (PII) in an Amazon Redshift cluster. The company's marketing, claims, and analytics teams need to be able to access the customer data.

The marketing team should have access to obfuscated claim information but should have full access to customer contact information. The claims team should have access to customer information for each claim that the team processes. The analytics team should have access only to obfuscated PII data.

Which solution will enforce these data access requirements with the LEAST administrative overhead?

Options
A. Create a separate Redshift cluster for each team. Load only the required data for each team. Restrict access to clusters based on the teams.
B. Create views that include required fields for each of the data requirements. Grant the teams access only to the view that each team requires.
C. Create a separate Amazon Redshift database role for each team. Define masking policies that apply for each team separately. Attach appropriate masking policies to each team role.
D. Move the customer data to an Amazon S3 bucket. Use AWS Lake Formation to create a data lake. Use fine-grained security capabilities to grant each team appropriate permissions to access the data.

Answer: C

Explanation:
Redshift column-level data masking policies tied to roles allow obfuscation and conditional access per team with minimal maintenance versus duplicating data or clusters.

- A: Multiple clusters greatly increase cost/administration.
- B: Views alone don’t enforce dynamic masking needs per user/team.
- D: Moving to S3/Lake Formation changes architecture and adds migration overhead.

Question: 229
A financial company recently added more features to its mobile app. The new features required the company to create a new topic in an existing Amazon Managed Streaming for Apache Kafka (Amazon MSK) cluster.

A few days after the company added the new topic, Amazon CloudWatch raised an alarm on the RootDiskUsed metric for the MSK cluster.

How should the company address the CloudWatch alarm?

Options
A. Expand the storage of the MSK broker. Configure the MSK cluster storage to expand automatically.
B. Expand the storage of the Apache ZooKeeper nodes.
C. Update the MSK broker instance to a larger instance type. Restart the MSK cluster.
D. Specify the Target Volume-in-GiB parameter for the existing topic.

Answer: A

Explanation:
RootDiskUsed indicates broker disk usage. Expanding broker storage (and enabling auto-scaling of storage) addresses the condition without interrupting producers/consumers.

- B: ZooKeeper storage is not the limiting factor for topic data.
- C: Larger instances don’t solve insufficient disk capacity.
- D: Topic-level target volume setting is not an MSK control for storage scaling.

Question: 230
A data engineer needs to build an enterprise data catalog based on the company's Amazon S3 buckets and Amazon RDS databases. The data catalog must include storage format metadata for the data in the catalog.

Which solution will meet these requirements with the LEAST effort?

Options
A. Use an AWS Glue crawler to scan the S3 buckets and RDS databases and build a data catalog. Use data stewards to inspect the data and update the data catalog with the data format.
B. Use an AWS Glue crawler to build a data catalog. Use AWS Glue crawler classifiers to recognize the format of data and store the format in the catalog.
C. Use Amazon Macie to build a data catalog and to identify sensitive data elements. Collect the data format information from Macie.
D. Use scripts to scan data elements and to assign data classifications based on the format of the data.

Answer: B

Explanation:
Glue crawlers with built-in and custom classifiers automatically detect file/storage formats for S3 and RDS sources and store them in the Data Catalog with minimal effort.

- A: Manual steward updates add unnecessary overhead.
- C: Macie focuses on sensitive data discovery, not comprehensive cataloging/format metadata.
- D: Custom scripts increase development and maintenance burden.

Question: 231
A company is building an analytics solution. The solution uses Amazon S3 for data lake storage and Amazon Redshift for a data warehouse. The company wants to use Amazon Redshift Spectrum to query the data that is in Amazon S3. Which actions will provide the FASTEST queries? (Choose two.)

Options
A.Use gzip compression to compress individual files to sizes that are between 1 GB and 5 GB.
B.Use a columnar storage file format.
C.Partition the data based on the most common query predicates.
D.Split the data into files that are less than 10 KB.
E.Use file formats that are not splittable.

Answer: BC

Explanation:
Columnar file formats (like Parquet/ORC) are optimized for analytics and allow Spectrum to scan only needed columns. Partitioning data by common predicates reduces the amount of scanned data.

- A is incorrect because gzip is row-based and not splittable; large gzip files can slow performance.
- D is incorrect because very small files cause overhead during queries.
- E is incorrect because non-splittable formats reduce parallelism and slow down queries.

Question: 232
A company uses Amazon RDS to store transactional data. The company runs an RDS DB instance in a private subnet. A developer wrote an AWS Lambda function with default settings to insert, update, or delete data in the DB instance. The developer needs to give the Lambda function the ability to connect to the DB instance privately without using the public internet. Which combination of steps will meet this requirement with the LEAST operational overhead? (Choose two.)

Options
A.Turn on the public access setting for the DB instance.
B.Update the security group of the DB instance to allow only Lambda function invocations on the database port.
C.Configure the Lambda function to run in the same subnet that the DB instance uses.
D.Attach the same security group to the Lambda function and the DB instance. Include a self-referencing rule that allows access through the database port.
E.Update the network ACL of the private subnet to include a self-referencing rule that allows access through the database port.

Answer: CD

Explanation:
Running Lambda inside the same VPC/subnet as the RDS instance ensures private connectivity. Sharing the same security group with a self-referencing rule enables secure communication.

- A is incorrect because enabling public access exposes the DB to the internet.
- B is incorrect because security groups cannot filter on Lambda invocations; they filter IP/ports.

- E is incorrect because NACLs are stateless and unnecessary here; security groups suffice.

Question: 233
A company has a frontend ReactJS website that uses Amazon API Gateway to invoke REST APIs. The APIs perform the functionality of the website. A data engineer needs to write a Python script that can be occasionally invoked through API Gateway. The code must return results to API Gateway. Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Deploy a custom Python script on an Amazon Elastic Container Service (Amazon ECS) cluster.
B.Create an AWS Lambda Python function with provisioned concurrency.
C.Deploy a custom Python script that can integrate with API Gateway on Amazon Elastic Kubernetes Service (Amazon EKS).
D.Create an AWS Lambda function. Ensure that the function is warm by scheduling an Amazon EventBridge rule to invoke the Lambda function every 5 minutes by using mock events.

Answer: B

Explanation:
An AWS Lambda function integrates natively with API Gateway. Provisioned concurrency ensures the function is pre-initialized to reduce cold starts with minimal overhead.

- A is incorrect because ECS requires cluster management.
- C is incorrect because EKS has high operational overhead.
- D is incorrect because scheduled warmups add unnecessary invocations; provisioned concurrency is simpler.

Question: 234
A company has a production AWS account that runs company workloads. The company's security team created a security AWS account to store and analyze security logs from the production AWS account. The security logs in the production AWS account are stored in Amazon CloudWatch Logs. The company needs to use Amazon Kinesis Data Streams to deliver the security logs to the security AWS account. Which solution will meet these requirements?

Options
A.Create a destination data stream in the production AWS account. In the security AWS account, create an IAM role that has cross-account permissions to Kinesis Data Streams in the production AWS account.
B.Create a destination data stream in the security AWS account. Create an IAM role and a trust policy to grant CloudWatch Logs the permission to put data into the stream. Create a subscription filter in the security AWS account.
C.Create a destination data stream in the production AWS account. In the production AWS account, create an IAM role that has cross-account permissions to Kinesis Data Streams in the security AWS account.
D.Create a destination data stream in the security AWS account. Create an IAM role and a trust policy to grant CloudWatch Logs the permission to put data into the stream. Create a subscription filter in the production AWS account.

Answer: D

Explanation:
The correct design is to create the Kinesis stream in the security account and set a subscription filter in the production account that pushes data. The IAM role with trust allows CloudWatch Logs to write across accounts.

- A is incorrect because the subscription must be in the source account, not the destination.
- B is incorrect because filters must be created in the production account, not the security account.
- C is incorrect because the destination should be in the security account, not production.

Question: 235
A company uses Amazon S3 to store semi-structured data in a transactional data lake. Some of the data files are small, but other data files are tens of terabytes. A data engineer must perform a change data capture (CDC) operation to identify changed data from the data source. The data source sends a full snapshot as a JSON file every day and ingests the changed data into the data lake. Which solution will capture the changed data MOST cost-effectively?

Options
A.Create an AWS Lambda function to identify the changes between the previous data and the current data. Configure the Lambda function to ingest the changes into the data lake.
B.Ingest the data into Amazon RDS for MySQL. Use AWS Database Migration Service (AWS DMS) to write the changed data to the data lake.
C.Use an open source data lake format to merge the data source with the S3 data lake to insert the new data and update the existing data.
D.Ingest the data into an Amazon Aurora MySQL DB instance that runs Aurora Serverless. Use AWS Database Migration Service (AWS DMS) to write the changed data to the data lake.

Answer: C

Explanation:
Open-source data lake formats (Apache Hudi, Delta Lake, Iceberg) support incremental merges directly on S3, enabling cost-effective CDC without moving data to a database.

- A is incorrect because Lambda is inefficient for large datasets (terabytes).
- B is incorrect because ingesting to RDS adds cost and overhead.
- D is incorrect because Aurora + DMS adds unnecessary complexity and cost.

Question: 236
A data engineer runs Amazon Athena queries on data that is in an Amazon S3 bucket. The Athena queries use AWS Glue Data Catalog as a metadata table. The data engineer notices that the Athena query plans are experiencing a performance bottleneck. The data engineer determines that the cause of the performance bottleneck is the large number of partitions that are in the S3 bucket. The data engineer must resolve the performance bottleneck and reduce Athena query planning time. Which solutions will meet these requirements? (Choose two.)

Options
A.Create an AWS Glue partition index. Enable partition filtering.
B.Bucket the data based on a column that the data have in common in a WHERE clause of the user query.
C.Use Athena partition projection based on the S3 bucket prefix.
D.Transform the data that is in the S3 bucket to Apache Parquet format.
E.Use the Amazon EMR S3DistCP utility to combine smaller objects in the S3 bucket into larger objects.

Answer: AC

Explanation:
Partition indexes and partition projection both optimize query planning by reducing metadata lookups. They improve query plan generation significantly.

- B is incorrect because bucketing improves query performance but doesn't reduce partition planning overhead.
- D is incorrect because Parquet improves scan efficiency but not planning time.
- E is incorrect because combining files reduces read overhead, not partition planning time.

Question: 237
A data engineer must manage the ingestion of real-time streaming data into AWS. The data engineer wants to perform real-time analytics on the incoming streaming data by using time-based aggregations over a window of up to 30 minutes. The data engineer needs a solution that is highly fault tolerant. Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Use an AWS Lambda function that includes both the business and the analytics logic to perform time-based aggregations over a window of up to 30 minutes for the data in Amazon Kinesis Data Streams.
B.Use Amazon Managed Service for Apache Flink (previously known as Amazon Kinesis Data Analytics) to analyze the data that might occasionally contain duplicates by using multiple types of aggregations.
C.Use an AWS Lambda function that includes both the business and the analytics logic to perform aggregations for a tumbling window of up to 30 minutes, based on the event timestamp.
D.Use Amazon Managed Service for Apache Flink (previously known as Amazon Kinesis Data Analytics) to analyze the data by using multiple types of aggregations to perform time-based analytics over a window of up to 30 minutes.

Answer: D

Explanation:
Managed Service for Apache Flink is designed for stateful, fault-tolerant, time-based aggregations and can handle large-scale streaming with minimal operational overhead.

- A is incorrect because Lambda is not ideal for maintaining 30-minute windows with large state.
- B is incorrect because duplicates handling is irrelevant; the focus is time-based window aggregation.
- C is incorrect because Lambda cannot efficiently maintain state over a 30-minute window.

Question: 238
A company is planning to upgrade its Amazon Elastic Block Store (Amazon EBS) General Purpose SSD storage from gp2 to gp3. The company wants to prevent any interruptions in its Amazon EC2 instances that will cause data loss during the migration to the upgraded storage. Which solution will meet these requirements with the LEAST operational overhead?

Options
A.Create snapshots of the gp2 volumes. Create new gp3 volumes from the snapshots. Attach the new gp3 volumes to the EC2 instances.
B.Create new gp3 volumes. Gradually transfer the data to the new gp3 volumes. When the transfer is complete, mount the new gp3 volumes to the EC2 instances to replace the gp2 volumes.
C.Change the volume type of the existing gp2 volumes to gp3. Enter new values for volume size, IOPS, and throughput.
D.Use AWS DataSync to create new gp3 volumes. Transfer the data from the original gp2 volumes to the new gp3 volumes.

Answer: C

Explanation:
EBS supports changing volume types directly (gp2 → gp3) without detaching or data loss. This is the simplest, least overhead method.

- A is incorrect because creating new volumes and attaching them involves downtime.
- B is incorrect because gradual transfers require manual intervention and downtime.
- D is incorrect because DataSync is unnecessary for in-place EBS upgrades.

Question: 239
A company is migrating its database servers from Amazon EC2 instances that run Microsoft SQL Server to Amazon RDS for Microsoft SQL Server DB instances. The company's analytics team must export large data elements every day until the migration is complete. The data elements are the result of SQL joins across multiple tables. The data must be in Apache Parquet format. The analytics team must store the data in Amazon S3. Which solution will meet these requirements in the MOST operationally efficient way?

Options
A.Create a view in the EC2 instance-based SQL Server databases that contains the required data elements. Create an AWS Glue job that selects the data directly from the view and transfers the data in Parquet format to an S3 bucket. Schedule the AWS Glue job to run every day.
B.Schedule SQL Server Agent to run a daily SQL query that selects the desired data elements from the EC2 instance-based SQL Server databases. Configure the query to direct the output .csv objects to an S3 bucket. Create an S3 event that invokes an AWS Lambda function to transform the output format from .csv to Parquet.
C.Use a SQL query to create a view in the EC2 instance-based SQL Server databases that contains the required data elements. Create and run an AWS Glue crawler to read the view. Create an AWS Glue job that retrieves the data and transfers the data in Parquet format to an S3 bucket. Schedule the AWS Glue job to run every day.
D.Create an AWS Lambda function that queries the EC2 instance-based databases by using Java Database Connectivity (JDBC). Configure the Lambda function to retrieve the required data, transform the data into Parquet format, and transfer the data into an S3 bucket. Use Amazon EventBridge to schedule the Lambda function to run every day.

Answer: C

Explanation:
Using AWS Glue to read from SQL Server views and export to S3 in Parquet is automated and efficient. Glue jobs can be scheduled easily for daily runs.

- A is incorrect because Glue cannot directly query without crawling/connection setup.
- B is incorrect because outputting CSV then converting to Parquet is inefficient.
- D is incorrect because Lambda is not suitable for large data exports.

Question: 240
A data engineering team is using an Amazon Redshift data warehouse for operational reporting. The team wants to prevent performance issues that might result from long-running queries. A data engineer must choose a system table in Amazon Redshift to record anomalies when a query optimizer identifies conditions that might indicate performance issues. Which table views should the data engineer use to meet this requirement?

Options
A.STL_USAGE_CONTROL
B.STL_ALERT_EVENT_LOG
C.STL_QUERY_METRICS
D.STL_PLAN_INFO

Answer: B

Explanation:
STL_ALERT_EVENT_LOG records alerts generated by the query optimizer when it detects conditions that may cause performance issues (e.g., missing stats, skew).

- A is incorrect because STL_USAGE_CONTROL does not track optimizer alerts.
- C is incorrect because STL_QUERY_METRICS shows runtime metrics, not optimizer warnings.
- D is incorrect because STL_PLAN_INFO provides execution plan info, not alerts.