QUESTION 88
An online gaming company is using an Amazon Kinesis Data Analytics SQL application with a Kinesis data stream as its source. The source sends three non-null fields to the application: player_id, score, and us_5_digit_zip_code.
A data analyst has a .csv mapping file that maps a small number of us_5_digit_zip_code values to a territory code. The data analyst needs to include the territory code, if one exists, as an additional output of the Kinesis Data Analytics application.
How should the data analyst meet this requirement while minimizing costs?

A. Store the contents of the mapping file in an Amazon DynamoDB table.
Preprocess the records as they arrive in the Kinesis Data Analytics application with an AWS Lambda function that fetches the mapping and supplements each record to include the territory code, if one exists.
Change the SQL query in the application to include the new field in the SELECT statement.
B. Store the mapping file in an Amazon S3 bucket and configure the reference data column headers for the .csv file in the Kinesis Data Analytics application.
Change the SQL query in the application to include a join to the file's S3 Amazon Resource Name (ARN), and add the territory code field to the SELECT columns.
C. Store the mapping file in an Amazon S3 bucket and configure it as a reference data source for the Kinesis Data Analytics application.
Change the SQL query in the application to include a join to the reference table and add the territory code field to the SELECT columns.
D. Store the contents of the mapping file in an Amazon DynamoDB table.
Change the Kinesis Data Analytics application to send its output to an AWS Lambda function that fetches the mapping and supplements each record to include the territory code, if one exists.
Forward the record from the Lambda function to the original application destination.

Answer: C

QUESTION 89
A company has collected more than 100 TB of log files in the last 24 months. The files are stored as raw text in a dedicated Amazon S3 bucket. Each object has a key of the form year-month- day_log_HHmmss.txt where HHmmss represents the time the log file was initially created. A table was created in Amazon Athena that points to the S3 bucket. One-time queries are run against a subset of columns in the table several times an hour.
A data analyst must make changes to reduce the cost of running these queries. Management wants a solution with minimal maintenance overhead.
Which combination of steps should the data analyst take to meet these requirements? (Choose three.)

A. Convert the log files to Apace Avro format.
B. Add a key prefix of the form date=year-month-day/ to the S3 objects to partition the data.
C. Convert the log files to Apache Parquet format.
D. Add a key prefix of the form year-month-day/ to the S3 objects to partition the data.

E. Drop and recreate the table with the PARTITIONED BY clause. Run the ALTER TABLE ADD PARTITION statement.
F. Drop and recreate the table with the PARTITIONED BY clause. Run the MSCK REPAIR TABLE statement.

Answer: BCF

QUESTION 90
A company has an application that ingests streaming data. The company needs to analyze this stream over a 5-minute timeframe to evaluate the stream for anomalies with Random Cut Forest (RCF) and summarize the current count of status codes. The source and summarized data should be persisted for future use.
Which approach would enable the desired outcome while keeping data persistence costs low?

A. Ingest the data stream with Amazon Kinesis Data Streams.
Have an AWS Lambda consumer evaluate the stream, collect the number status codes, and evaluate the data against a previously trained RCF model.
Persist the source and results as a time series to Amazon DynamoDB.
B. Ingest the data stream with Amazon Kinesis Data Streams.
Have a Kinesis Data Analytics application evaluate the stream over a 5-minute window using the RCF function and summarize the count of status codes.
Persist the source and results to Amazon S3 through output delivery to Kinesis Data Firehouse.
C. Ingest the data stream with Amazon Kinesis Data Firehose with a delivery frequency of 1 minute or 1 MB in Amazon S3.
Ensure Amazon S3 triggers an event to invoke an AWS Lambda consumer that evaluates the batch data, collects the number status codes, and evaluates the data against a previously trained RCF model.
Persist the source and results as a time series to Amazon DynamoDB.
D. Ingest the data stream with Amazon Kinesis Data Firehose with a delivery frequency of 5 minutes or 1 MB into Amazon S3.
Have a Kinesis Data Analytics application evaluate the stream over a 1-minute window using the RCF function and summarize the count of status codes.
Persist the results to Amazon S3 through a Kinesis Data Analytics output to an AWS Lambda integration.

Answer: B

QUESTION 91
An online retailer needs to deploy a product sales reporting solution. The source data is exported from an external online transaction processing (OLTP) system for reporting. Roll-up data is calculated each day for the previous day's activities. The reporting system has the following requirements:
- Have the daily roll-up data readily available for 1 year.
- After 1 year, archive the daily roll-up data for occasional but immediate access.
- The source data exports stored in the reporting system must be retained for 5 years. Query access will be needed only for re-evaluation, which may occur within the first 90 days.
Which combination of actions will meet these requirements while keeping storage costs to a minimum? (Choose two.)

A. Store the source data initially in the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class.
Apply a lifecycle configuration that changes the storage class to Amazon S3 Glacier Deep Archive 90 days after creation, and then deletes the data 5 years after creation.
B. Store the source data initially in the Amazon S3 Glacier storage class.
Apply a lifecycle configuration that changes the storage class from Amazon S3 Glacier to Amazon S3 Glacier Deep Archive 90 days after creation, and then deletes the data 5 years after creation.
C. Store the daily roll-up data initially in the Amazon S3 Standard storage class.
Apply a lifecycle configuration that changes the storage class to Amazon S3 Glacier Deep Archive 1 year after data creation.

D. Store the daily roll-up data initially in the Amazon S3 Standard storage class.
Apply a lifecycle configuration that changes the storage class to Amazon S3 Standard-Infrequent Access (S3 Standard- IA) 1 year after data creation.
E. Store the daily roll-up data initially in the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class.
Apply a lifecycle configuration that changes the storage class to Amazon S3 Glacier 1 year after data creation.

Answer: BE

QUESTION 92
A company needs to store objects containing log data in JSON format. The objects are generated by eight applications running in AWS. Six of the applications generate a total of 500 KiB of data per second, and two of the applications can generate up to 2 MiB of data per second.
A data engineer wants to implement a scalable solution to capture and store usage data in an Amazon S3 bucket. The usage data objects need to be reformatted, converted to .csv format, and then compressed before they are stored in Amazon S3. The company requires the solution to include the least custom code possible and has authorized the data engineer to request a service quota increase if needed.
Which solution meets these requirements?

A. Configure an Amazon Kinesis Data Firehose delivery stream for each application.
Write AWS Lambda functions to read log data objects from the stream for each application.
Have the function perform reformatting and .csv conversion.
Enable compression on all the delivery streams.
B. Configure an Amazon Kinesis data stream with one shard per application.
Write an AWS Lambda function to read usage data objects from the shards.
Have the function perform .csv conversion, reformatting, and compression of the data.
Have the function store the output in Amazon S3.
C. Configure an Amazon Kinesis data stream for each application.
Write an AWS Lambda function to read usage data objects from the stream for each application.
Have the function perform .csv conversion, reformatting, and compression of the data.
Have the function store the output in Amazon S3.
D. Store usage data objects in an Amazon DynamoDB table.
Configure a DynamoDB stream to copy the objects to an S3 bucket.
Configure an AWS Lambda function to be triggered when objects are written to the S3 bucket.
Have the function convert the objects into .csv format.

Answer: B

QUESTION 93
A data analytics specialist is building an automated ETL ingestion pipeline using AWS Glue to ingest compressed files that have been uploaded to an Amazon S3 bucket. The ingestion pipeline should support incremental data processing.
Which AWS Glue feature should the data analytics specialist use to meet this requirement?

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

Answer: B

QUESTION 94
A telecommunications company is looking for an anomaly-detection solution to identify fraudulent calls. The company currently uses Amazon Kinesis to stream voice call records in a JSON format from its on- premises database to Amazon S3. The existing dataset contains voice call records with 200 columns. To detect fraudulent calls, the solution would need to look at 5 of these columns only.

The company is interested in a cost-effective solution using AWS that requires minimal effort and experience in anomaly-detection algorithms.
Which solution meets these requirements?

A. Use an AWS Glue job to transform the data from JSON to Apache Parquet.
Use AWS Glue crawlers to discover the schema and build the AWS Glue Data Catalog.
Use Amazon Athena to create a table with a subset of columns.
Use Amazon QuickSight to visualize the data and then use Amazon QuickSight machine learning-powered anomaly detection.
B. Use Kinesis Data Firehose to detect anomalies on a data stream from Kinesis by running SQL queries, which compute an anomaly score for all calls and store the output in Amazon RDS.
Use Amazon Athena to build a dataset and Amazon QuickSight to visualize the results.
C. Use an AWS Glue job to transform the data from JSON to Apache Parquet.
Use AWS Glue crawlers to discover the schema and build the AWS Glue Data Catalog.
Use Amazon SageMaker to build an anomaly detection model that can detect fraudulent calls by ingesting data from Amazon S3.
D. Use Kinesis Data Analytics to detect anomalies on a data stream from Kinesis by running SQL queries, which compute an anomaly score for all calls.
Connect Amazon QuickSight to Kinesis Data Analytics to visualize the anomaly scores.

Answer: A

QUESTION 95
An online retailer is rebuilding its inventory management system and inventory reordering system to automatically reorder products by using Amazon Kinesis Data Streams. The inventory management system uses the Kinesis Producer Library (KPL) to publish data to a stream. The inventory reordering system uses the Kinesis Client Library (KCL) to consume data from the stream. The stream has been configured to scale as needed. Just before production deployment, the retailer discovers that the inventory reordering system is receiving duplicated data.
Which factors could be causing the duplicated data? (Choose two.)

A. The producer has a network-related timeout.
B. The stream's value for the IteratorAgeMilliseconds metric is 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 is too high.

Answer: BD

QUESTION 96
A large retailer has successfully migrated to an Amazon S3 data lake architecture. The company's marketing team is using Amazon Redshift and Amazon QuickSight to analyze data, and derive and visualize insights. To ensure the marketing team has the most up-to-date actionable information, a data analyst implements nightly refreshes of Amazon Redshift using terabytes of updates from the previous day.
After the first nightly refresh, users report that half of the most popular dashboards that had been running correctly before the refresh are now running much slower. Amazon CloudWatch does not show any alerts.
What is the MOST likely cause for the performance degradation?

A. The dashboards are suffering from inefficient SQL queries.
B. The cluster is undersized for the queries being run by the dashboards.
C. The nightly data refreshes are causing a lingering transaction that cannot be automatically closed by Amazon Redshift due to ongoing user workloads.
D. The nightly data refreshes left the dashboard tables in need of a vacuum operation that could not be automatically performed by Amazon Redshift due to ongoing user workloads.

Answer: B

QUESTION 97
A marketing company is storing its campaign response data in Amazon S3. A consistent set of sources has generated the data for each campaign. The data is saved into Amazon S3 as .csv files. A business analyst will use Amazon Athena to analyze each campaign's data. The company needs the cost of ongoing data analysis with Athena to be minimized.
Which combination of actions should a data analytics specialist take to meet these requirements? (Choose two.)

A. Convert the .csv files to Apache Parquet.
B. Convert the .csv files to Apache Avro.
C. Partition the data by campaign.
D. Partition the data by source.
E. Compress the .csv files.

Answer: BC

QUESTION 98
An online retail company is migrating its reporting system to AWS. The company's legacy system runs data processing on online transactions using a complex series of nested Apache Hive queries. Transactional data is exported from the online system to the reporting system several times a day.
Schemas in the files are stable between updates.
A data analyst wants to quickly migrate the data processing to AWS, so any code changes should be minimized. To keep storage costs low, the data analyst decides to store the data in Amazon S3. It is vital that the data from the reports and associated analytics is completely up to date based on the data in Amazon S3.
Which solution meets these requirements?

A. Create an AWS Glue Data Catalog to manage the Hive metadata.
Create an AWS Glue crawler over Amazon S3 that runs when data is refreshed to ensure that data changes are updated.
Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
B. Create an AWS Glue Data Catalog to manage the Hive metadata.
Create an Amazon EMR cluster with consistent view enabled.
Run emrfs sync before each analytics step to ensure data changes are updated.
Create an EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
C. Create an Amazon Athena table with CREATE TABLE AS SELECT (CTAS) to ensure data is refreshed from underlying queries against the raw dataset.
Create an AWS Glue Data Catalog to manage the Hive metadata over the CTAS table.
Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
D. Use an S3 Select query to ensure that the data is properly updated.
Create an AWS Glue Data Catalog to manage the Hive metadata over the S3 Select table.
Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.

Answer: A

QUESTION 99
A media company is using Amazon QuickSight dashboards to visualize its national sales data. The dashboard is using a dataset with these fields: ID, date, time_zone, city, state, country, longitude, latitude, sales_volume, and number_of_items.
To modify ongoing campaigns, the company wants an interactive and intuitive visualization of which states across the country recorded a significantly lower sales volume compared to the national average.
Which addition to the company's QuickSight dashboard will meet this requirement?

A. A geospatial color-coded chart of sales volume data across the country.

B. A pivot table of sales volume data summed up at the state level.
C. A drill-down layer for state-level sales volume data.
D. A drill through to other dashboards containing state-level sales volume data.

Answer: B

QUESTION 100
A company hosts an on-premises PostgreSQL database that contains historical data. An internal legacy application uses the database for read-only activities. The company's business team wants to move the data to a data lake in Amazon S3 as soon as possible and enrich the data for analytics.
The company has set up an AWS Direct Connect connection between its VPC and its on-premises network. A data analytics specialist must design a solution that achieves the business team's goals with the least operational overhead.
Which solution meets these requirements?

A. Upload the data from the on-premises PostgreSQL database to Amazon S3 by using a customized batch upload process.
Use the AWS Glue crawler to catalog the data in Amazon S3.
Use an AWS Glue job to enrich and store the result in a separate S3 bucket in Apache Parquet format. Use Amazon Athena to query the data.
B. Create an Amazon RDS for PostgreSQL database and use AWS Database Migration Service (AWS DMS) to migrate the data into Amazon RDS.
Use AWS Data Pipeline to copy and enrich the data from the Amazon RDS for PostgreSQL table and move the data to Amazon S3.
Use Amazon Athena to query the data.
C. Configure an AWS Glue crawler to use a JDBC connection to catalog the data in the on-premises database.
Use an AWS Glue job to enrich the data and save the result to Amazon S3 in Apache Parquet format. Create an Amazon Redshift cluster and use Amazon Redshift Spectrum to query the data.
D. Configure an AWS Glue crawler to use a JDBC connection to catalog the data in the on-premises database.
Use an AWS Glue job to enrich the data and save the result to Amazon S3 in Apache Parquet format.
Use Amazon Athena to query the data.

Answer: B

QUESTION 101
A medical company has a system with sensor devices that read metrics and send them in real time to an Amazon Kinesis data stream. The Kinesis data stream has multiple shards. The company needs to calculate the average value of a numeric metric every second and set an alarm for whenever the value is above one threshold or below another threshold. The alarm must be sent to Amazon Simple Notification Service (Amazon SNS) in less than 30 seconds.
Which architecture meets these requirements?

A. Use an Amazon Kinesis Data Firehose delivery stream to read the data from the Kinesis data stream with an AWS Lambda transformation function that calculates the average per second and sends the alarm to Amazon SNS.
B. Use an AWS Lambda function to read from the Kinesis data stream to calculate the average per second and sent the alarm to Amazon SNS.
C. Use an Amazon Kinesis Data Firehose deliver stream to read the data from the Kinesis data stream and store it on Amazon S3.
Have Amazon S3 trigger an AWS Lambda function that calculates the average per second and sends the alarm to Amazon SNS.
D. Use an Amazon Kinesis Data Analytics application to read from the Kinesis data stream and calculate the average per second.
Send the results to an AWS Lambda function that sends the alarm to Amazon SNS.

Answer: C

QUESTION 102
An IoT company wants to release a new device that will collect data to track sleep overnight on an intelligent mattress. Sensors will send data that will be uploaded to an Amazon S3 bucket. About 2 MB of data is generated each night for each bed. Data must be processed and summarized for each user, and the results need to be available as soon as possible. Part of the process consists of time windowing and other functions. Based on tests with a Python script, every run will require about 1 GB of memory and will complete within a couple of minutes.
Which solution will run the script in the MOST cost-effective way?

A. AWS Lambda with a Python script
B. AWS Glue with a Scala job
C. Amazon EMR with an Apache Spark script
D. AWS Glue with a PySpark job

Answer: A

QUESTION 103
A company wants to provide its data analysts with uninterrupted access to the data in its Amazon Redshift cluster. All data is streamed to an Amazon S3 bucket with Amazon Kinesis Data Firehose. An AWS Glue job that is scheduled to run every 5 minutes issues a COPY command to move the data into Amazon Redshift.
The amount of data delivered is uneven throughout then day, and cluster utilization is high during certain periods. The COPY command usually completes within a couple of seconds. However, when load spike occurs, locks can exist and data can be missed. Currently, the AWS Glue job is configured to run without retries, with timeout at 5 minutes and concurrency at 1.
How should a data analytics specialist configure the AWS Glue job to optimize fault tolerance and improve data availability in the Amazon Redshift cluster?

A. Increase the number of retries. Decrease the timeout value. Increase the job concurrency.
B. Keep the number of retries at 0. Decrease the timeout value. Increase the job concurrency.
C. Keep the number of retries at 0. Decrease the timeout value. Keep the job concurrency at 1.
D. Keep the number of retries at 0. Increase the timeout value. Keep the job concurrency at 1.

Answer: B

QUESTION 104
A retail company leverages Amazon Athena for ad-hoc queries against an AWS Glue Data Catalog. The data analytics team manages the data catalog and data access for the company. The data analytics team wants to separate queries and manage the cost of running those queries by different workloads and teams. Ideally, the data analysts want to group the queries run by different users within a team, store the query results in individual Amazon S3 buckets specific to each team, and enforce cost constraints on the queries run against the Data Catalog.
Which solution meets these requirements?

A. Create IAM groups and resource tags for each team within the company.
Set up IAM policies that control user access and actions on the Data Catalog resources.
B. Create Athena resource groups for each team within the company and assign users to these groups.
Add S3 bucket names and other query configurations to the properties list for the resource groups.
C. Create Athena workgroups for each team within the company.
Set up IAM workgroup policies that control user access and actions on the workgroup resources.
D. Create Athena query groups for each team within the company and assign users to the groups.

Answer: A

QUESTION 105
A manufacturing company uses Amazon S3 to store its data. The company wants to use AWS Lake Formation to provide granular-level security on those data assets. The data is in Apache Parquet format. The company has set a deadline for a consultant to build a data lake.
How should the consultant create the MOST cost-effective solution that meets these requirements?

A. Run Lake Formation blueprints to move the data to Lake Formation.
Once Lake Formation has the data, apply permissions on Lake Formation.
B. To create the data catalog, run an AWS Glue crawler on the existing Parquet data.
Register the Amazon S3 path and then apply permissions through Lake Formation to provide granular-level security.
C. Install Apache Ranger on an Amazon EC2 instance and integrate with Amazon EMR.
Using Ranger policies, create role-based access control for the existing data assets in Amazon S3.
D. Create multiple IAM roles for different users and groups.
Assign IAM roles to different data assets in Amazon S3 to create table-based and column-based access controls.

Answer: C

QUESTION 106
A company has an application that uses the Amazon Kinesis Client Library (KCL) to read records from a Kinesis data stream.
After a successful marketing campaign, the application experienced a significant increase in usage. As a result, a data analyst had to split some shards in the data stream. When the shards were split, the application started throwing an ExpiredIteratorExceptions error sporadically.
What should the data analyst do to resolve this?

A. Increase the number of threads that process the stream records.
B. Increase the provisioned read capacity units assigned to the stream's Amazon DynamoDB table.
C. Increase the provisioned write capacity units assigned to the stream's Amazon DynamoDB table.
D. Decrease the provisioned write capacity units assigned to the stream's Amazon DynamoDB table.

Answer: C

QUESTION 107
A company is building a service to monitor fleets of vehicles. The company collects IoT data from a device in each vehicle and loads the data into Amazon Redshift in near-real time. Fleet owners upload .csv files containing vehicle reference data into Amazon S3 at different times throughout the day. A nightly process loads the vehicle reference data from Amazon S3 into Amazon Redshift. The company joins the IoT data from the device and the vehicle reference data to power reporting and dashboards. Fleet owners are frustrated by waiting a day for the dashboards to update.
Which solution would provide the SHORTEST delay between uploading reference data to Amazon S3 and the change showing up in the owners' dashboards?

A. Use S3 event notifications to trigger an AWS Lambda function to copy the vehicle reference data into Amazon Redshift immediately when the reference data is uploaded to Amazon S3.
B. Create and schedule an AWS Glue Spark job to run every 5 minutes.
The job inserts reference data into Amazon Redshift.
C. Send reference data to Amazon Kinesis Data Streams.
Configure the Kinesis data stream to directly load the reference data into Amazon Redshift in real time.
D. Send the reference data to an Amazon Kinesis Data Firehose delivery stream.
Configure Kinesis with a buffer interval of 60 seconds and to directly load the data into Amazon Redshift.

Answer: A

QUESTION 108
A company is migrating from an on-premises Apache Hadoop cluster to an Amazon EMR cluster. The cluster runs only during business hours. Due to a company requirement to avoid intraday cluster failures, the EMR cluster must be highly available. When the cluster is terminated at the end of each business day, the data must persist.
Which configurations would enable the EMR cluster to meet these requirements? (Choose three.)

A. EMR File System (EMRFS) for storage
B. Hadoop Distributed File System (HDFS) for storage
C. AWS Glue Data Catalog as the metastore for Apache Hive
D. MySQL database on the master node as the metastore for Apache Hive
E. Multiple master nodes in a single Availability Zone
F. Multiple master nodes in multiple Availability Zones

Answer: BCF

QUESTION 109
A retail company wants to use Amazon QuickSight to generate dashboards for web and in-store sales. A group of 50 business intelligence professionals will develop and use the dashboards. Once ready, the dashboards will be shared with a group of 1,000 users.
The sales data comes from different stores and is uploaded to Amazon S3 every 24 hours. The data is partitioned by year and month, and is stored in Apache Parquet format. The company is using the AWS Glue Data Catalog as its main data catalog and Amazon Athena for querying. The total size of the uncompressed data that the dashboards query from at any point is 200 GB.
Which configuration will provide the MOST cost-effective solution that meets these requirements?

A. Load the data into an Amazon Redshift cluster by using the COPY command.
Configure 50 author users and 1,000 reader users.
Use QuickSight Enterprise edition.
Configure an Amazon Redshift data source with a direct query option.
B. Use QuickSight Standard edition.
Configure 50 author users and 1,000 reader users.
Configure an Athena data source with a direct query option.
C. Use QuickSight Enterprise edition.
Configure 50 author users and 1,000 reader users.
Configure an Athena data source and import the data into SPICE.
Automatically refresh every 24 hours.
D. Use QuickSight Enterprise edition.
Configure 1 administrator and 1,000 reader users.
Configure an S3 data source and import the data into SPICE.
Automatically refresh every 24 hours.

Answer: C

QUESTION 110
A central government organization is collecting events from various internal applications using Amazon Managed Streaming for Apache Kafka (Amazon MSK). The organization has configured a separate Kafka topic for each application to separate the data. For security reasons, the Kafka cluster has been configured to only allow TLS encrypted data and it encrypts the data at rest.
A recent application update showed that one of the applications was configured incorrectly, resulting in writing data to a Kafka topic that belongs to another application. This resulted in multiple errors in the analytics pipeline as data from different applications appeared on the same topic. After this incident, the organization wants to prevent applications from writing to a topic different than the one they should write to.
Which solution meets these requirements with the least amount of effort?

A. Create a different Amazon EC2 security group for each application.

Configure each security group to have access to a specific topic in the Amazon MSK cluster.
Attach the security group to each application based on the topic that the applications should read and write to.

B. Install Kafka Connect on each application instance and configure each Kafka Connect instance to write to a specific topic only.

C. Use Kafka ACLs and configure read and write permissions for each topic.
Use the distinguished name of the clients' TLS certificates as the principal of the ACL.

D. Create a different Amazon EC2 security group for each application.
Create an Amazon MSK cluster and Kafka topic for each application.
Configure each security group to have access to the specific cluster.

Answer: B

QUESTION 111
A company wants to collect and process events data from different departments in near-real time. Before storing the data in Amazon S3, the company needs to clean the data by standardizing the format of the address and timestamp columns. The data varies in size based on the overall load at each particular point in time. A single data record can be 100 KB-10 MB.
How should a data analytics specialist design the solution for data ingestion?

A. Use Amazon Kinesis Data Streams.
Configure a stream for the raw data.
Use a Kinesis Agent to write data to the stream.
Create an Amazon Kinesis Data Analytics application that reads data from the raw stream, cleanses it, and stores the output to Amazon S3.

B. Use Amazon Kinesis Data Firehose.
Configure a Firehose delivery stream with a preprocessing AWS Lambda function for data cleansing.
Use a Kinesis Agent to write data to the delivery stream. Configure Kinesis Data Firehose to deliver the data to Amazon S3.

C. Use Amazon Managed Streaming for Apache Kafka.
Configure a topic for the raw data.
Use a Kafka producer to write data to the topic.
Create an application on Amazon EC2 that reads data from the topic by using the Apache Kafka consumer API, cleanses the data, and writes to Amazon S3.

D. Use Amazon Simple Queue Service (Amazon SQS).
Configure an AWS Lambda function to read events from the SQS queue and upload the events to Amazon S3.

Answer: B

QUESTION 112
An operations team notices that a few AWS Glue jobs for a given ETL application are failing. The AWS Glue jobs read a large number of small JOSN files from an Amazon S3 bucket and write the data to a different S3 bucket in Apache Parquet format with no major transformations. Upon initial investigation, a data engineer notices the following error message in the History tab on the AWS Glue console: "Command Failed with Exit Code 1."
Upon further investigation, the data engineer notices that the driver memory profile of the failed jobs crosses the safe threshold of 50% usage quickly and reaches 90?5% soon after. The average memory usage across all executors continues to be less than 4%.
The data engineer also notices the following error while examining the related Amazon CloudWatch Logs.

```
# java.lang.OutOfMemoryError: Java heap space
# -XX: OnOutOfMemoryError= "kill -9 %p"
# Executing /bin/sh -c "kill -9 12039"…
```

What should the data engineer do to solve the failure in the MOST cost-effective way?

A. Change the worker type from Standard to G.2X.
B. Modify the AWS Glue ETL code to use the `groupFiles': `inPartition' feature.
C. Increase the fetch size setting by using AWS Glue dynamics frame.
D. Modify maximum capacity to increase the total maximum data processing units (DPUs) used.

Answer: D

QUESTION 113
A transport company wants to track vehicular movements by capturing geolocation records. The records are 10 B in size and up to 10,000 records are captured each second. Data transmission delays of a few minutes are acceptable, considering unreliable network conditions. The transport company decided to use Amazon Kinesis Data Streams to ingest the data. The company is looking for a reliable mechanism to send data to Kinesis Data Streams while maximizing the throughput efficiency of the Kinesis shards.
Which solution will meet the company's requirements?

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

Answer: B