EXAMTOPICS
- Expert Verified, Online, **Free**.

⚙ Custom View Settings

Topic 1 - Single Topic

Question #1

Topic 1

A financial services company needs to aggregate daily stock trade data from the exchanges into a data store. The company requires that data be streamed directly into the data store, but also occasionally allows data to be modified using SQL. The solution should integrate complex, analytic queries running with minimal latency. The solution must provide a business intelligence dashboard that enables viewing of the top contributors to anomalies in stock prices.
Which solution meets the company's requirements?

A. Use Amazon Kinesis Data Firehose to stream data to Amazon S3. Use Amazon Athena as a data source for Amazon QuickSight to create a business intelligence dashboard.

B. Use Amazon Kinesis Data Streams to stream data to Amazon Redshift. Use Amazon Redshift as a data source for Amazon QuickSight to create a business intelligence dashboard.

C. Use Amazon Kinesis Data Firehose to stream data to Amazon Redshift. Use Amazon Redshift as a data source for Amazon QuickSight to create a business intelligence dashboard.

D. Use Amazon Kinesis Data Streams to stream data to Amazon S3. Use Amazon Athena as a data source for Amazon QuickSight to create a business intelligence dashboard.

Question #2

Topic 1

A financial company hosts a data lake in Amazon S3 and a data warehouse on an Amazon Redshift cluster. The company uses Amazon QuickSight to build dashboards and wants to secure access from its on-premises Active Directory to Amazon QuickSight.
How should the data be secured?

A. Use an Active Directory connector and single sign-on (SSO) in a corporate network environment.

B. Use a VPC endpoint to connect to Amazon S3 from Amazon QuickSight and an IAM role to authenticate Amazon Redshift.

C. Establish a secure connection by creating an S3 endpoint to connect Amazon QuickSight and a VPC endpoint to connect to Amazon Redshift.

D. Place Amazon QuickSight and Amazon Redshift in the security group and use an Amazon S3 endpoint to connect Amazon QuickSight to Amazon S3.

Question #3 *Topic 1*

A real estate company has a mission-critical application using Apache HBase in Amazon EMR. Amazon EMR is configured with a single master node. The company has over 5 TB of data stored on an Hadoop Distributed File System (HDFS). The company wants a cost-effective solution to make its HBase data highly available.
Which architectural pattern meets company's requirements?

A. Use Spot Instances for core and task nodes and a Reserved Instance for the EMR master node. Configure the EMR cluster with multiple master nodes. Schedule automated snapshots using Amazon EventBridge.

B. Store the data on an EMR File System (EMRFS) instead of HDFS. Enable EMRFS consistent view. Create an EMR HBase cluster with multiple master nodes. Point the HBase root directory to an Amazon S3 bucket.

C. Store the data on an EMR File System (EMRFS) instead of HDFS and enable EMRFS consistent view. Run two separate EMR clusters in two different Availability Zones. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.

D. Store the data on an EMR File System (EMRFS) instead of HDFS and enable EMRFS consistent view. Create a primary EMR HBase cluster with multiple master nodes. Create a secondary EMR HBase read-replica cluster in a separate Availability Zone. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.

Question #4 *Topic 1*

A software company hosts an application on AWS, and new features are released weekly. As part of the application testing process, a solution must be developed that analyzes logs from each Amazon EC2 instance to ensure that the application is working as expected after each deployment. The collection and analysis solution should be highly available with the ability to display new information with minimal delays.
Which method should the company use to collect and analyze the logs?

A. Enable detailed monitoring on Amazon EC2, use Amazon CloudWatch agent to store logs in Amazon S3, and use Amazon Athena for fast, interactive log analytics.

B. Use the Amazon Kinesis Producer Library (KPL) agent on Amazon EC2 to collect and send data to Kinesis Data Streams to further push the data to Amazon OpenSearch Service (Amazon Elasticsearch Service) and visualize using Amazon QuickSight.

C. Use the Amazon Kinesis Producer Library (KPL) agent on Amazon EC2 to collect and send data to Kinesis Data Firehose to further push the data to Amazon OpenSearch Service (Amazon Elasticsearch Service) and OpenSearch Dashboards (Kibana).

D. Use Amazon CloudWatch subscriptions to get access to a real-time feed of logs and have the logs delivered to Amazon Kinesis Data Streams to further push the data to Amazon OpenSearch Service (Amazon Elasticsearch Service) and OpenSearch Dashboards (Kibana).

Question #5 *Topic 1*

A data analyst is using AWS Glue to organize, cleanse, validate, and format a 200 GB dataset. The data analyst triggered the job to run with the Standard worker type. After 3 hours, the AWS Glue job status is still RUNNING. Logs from the job run show no error codes. The data analyst wants to improve the job execution time without overprovisioning.
Which actions should the data analyst take?

A. Enable job bookmarks in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the executor- cores job parameter.

B. Enable job metrics in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the maximum capacity job parameter.

C. Enable job metrics in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the spark.yarn.executor.memoryOverhead job parameter.

D. Enable job bookmarks in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the num- executors job parameter.

Question #6 *Topic 1*

A company has a business unit uploading .csv files to an Amazon S3 bucket. The company's data platform team has set up an AWS Glue crawler to do discovery, and create tables and schemas. An AWS Glue job writes processed data from the created tables to an Amazon Redshift database. The AWS Glue job handles column mapping and creating the Amazon Redshift table appropriately. When the AWS Glue job is rerun for any reason in a day, duplicate records are introduced into the Amazon Redshift table.
Which solution will update the Redshift table without duplicates when jobs are rerun?

A. Modify the AWS Glue job to copy the rows into a staging table. Add SQL commands to replace the existing rows in the main table as postactions in the DynamicFrameWriter class.

B. Load the previously inserted data into a MySQL database in the AWS Glue job. Perform an upsert operation in MySQL, and copy the results to the Amazon Redshift table.

C. Use Apache Spark's DataFrame dropDuplicates() API to eliminate duplicates and then write the data to Amazon Redshift.

D. Use the AWS Glue ResolveChoice built-in transform to select the most recent value of the column.

Question #7 *Topic 1*

A streaming application is reading data from Amazon Kinesis Data Streams and immediately writing the data to an Amazon S3 bucket every 10 seconds. The application is reading data from hundreds of shards. The batch interval cannot be changed due to a separate requirement. The data is being accessed by Amazon
Athena. Users are seeing degradation in query performance as time progresses.
Which action can help improve query performance?

A. Merge the files in Amazon S3 to form larger files.

B. Increase the number of shards in Kinesis Data Streams.

C. Add more memory and CPU capacity to the streaming application.

D. Write the files to multiple S3 buckets.

Question #8 *Topic 1*

A company uses Amazon OpenSearch Service (Amazon Elasticsearch Service) to store and analyze its website clickstream data. The company ingests 1 TB of data daily using Amazon Kinesis Data Firehose and stores one day's worth of data in an Amazon ES cluster.
The company has very slow query performance on the Amazon ES index and occasionally sees errors from Kinesis Data Firehose when attempting to write to the index. The Amazon ES cluster has 10 nodes running a single index and 3 dedicated master nodes. Each data node has 1.5 TB of Amazon EBS storage attached and the cluster is configured with 1,000 shards. Occasionally, JVMMemoryPressure errors are found in the cluster logs.
Which solution will improve the performance of Amazon ES?

A. Increase the memory of the Amazon ES master nodes.

B. Decrease the number of Amazon ES data nodes.

C. Decrease the number of Amazon ES shards for the index.

D. Increase the number of Amazon ES shards for the index.

Question #9

Topic 1

A manufacturing company has been collecting IoT sensor data from devices on its factory floor for a year and is storing the data in Amazon Redshift for daily analysis. A data analyst has determined that, at an expected ingestion rate of about 2 TB per day, the cluster will be undersized in less than 4 months. A long-term solution is needed. The data analyst has indicated that most queries only reference the most recent 13 months of data, yet there are also quarterly reports that need to query all the data generated from the past 7 years. The chief technology officer (CTO) is concerned about the costs, administrative effort, and performance of a long-term solution.
Which solution should the data analyst use to meet these requirements?

A. Create a daily job in AWS Glue to UNLOAD records older than 13 months to Amazon S3 and delete those records from Amazon Redshift. Create an external table in Amazon Redshift to point to the S3 location. Use Amazon Redshift Spectrum to join to data that is older than 13 months.

B. Take a snapshot of the Amazon Redshift cluster. Restore the cluster to a new cluster using dense storage nodes with additional storage capacity.

C. Execute a CREATE TABLE AS SELECT (CTAS) statement to move records that are older than 13 months to quarterly partitioned data in Amazon Redshift Spectrum backed by Amazon S3.

D. Unload all the tables in Amazon Redshift to an Amazon S3 bucket using S3 Intelligent-Tiering. Use AWS Glue to crawl the S3 bucket location to create external tables in an AWS Glue Data Catalog. Create an Amazon EMR cluster using Auto Scaling for any daily analytics needs, and use Amazon Athena for the quarterly reports, with both using the same AWS Glue Data Catalog.

Question #10

Topic 1

An insurance company has raw data in JSON format that is sent without a predefined schedule through an Amazon Kinesis Data Firehose delivery stream to an
Amazon S3 bucket. An AWS Glue crawler is scheduled to run every 8 hours to update the schema in the data catalog of the tables stored in the S3 bucket. Data analysts analyze the data using Apache Spark SQL on Amazon EMR set up with AWS Glue Data Catalog as the metastore. Data analysts say that, occasionally, the data they receive is stale. A data engineer needs to provide access to the most up-to-date data.
Which solution meets these requirements?

A. Create an external schema based on the AWS Glue Data Catalog on the existing Amazon Redshift cluster to query new data in Amazon S3 with Amazon Redshift Spectrum.

B. Use Amazon CloudWatch Events with the rate (1 hour) expression to execute the AWS Glue crawler every hour.

C. Using the AWS CLI, modify the execution schedule of the AWS Glue crawler from 8 hours to 1 minute.

D. Run the AWS Glue crawler from an AWS Lambda function triggered by an S3:ObjectCreated:* event notification on the S3 bucket.

Question #11

Topic 1

A company that produces network devices has millions of users. Data is collected from the devices on an hourly basis and stored in an Amazon S3 data lake.
The company runs analyses on the last 24 hours of data flow logs for abnormality detection and to troubleshoot and resolve user issues. The company also analyzes historical logs dating back 2 years to discover patterns and look for improvement opportunities.
The data flow logs contain many metrics, such as date, timestamp, source IP, and target IP. There are about 10 billion events every day.
How should this data be stored for optimal performance?

A. In Apache ORC partitioned by date and sorted by source IP

B. In compressed .csv partitioned by date and sorted by source IP

C. In Apache Parquet partitioned by source IP and sorted by date

D. In compressed nested JSON partitioned by source IP and sorted by date

Question #12 *Topic 1*

A banking company is currently using an Amazon Redshift cluster with dense storage (DS) nodes to store sensitive data. An audit found that the cluster is unencrypted. Compliance requirements state that a database with sensitive data must be encrypted through a hardware security module (HSM) with automated key rotation.
Which combination of steps is required to achieve compliance? (Choose two.)

A. Set up a trusted connection with HSM using a client and server certificate with automatic key rotation.

B. Modify the cluster with an HSM encryption option and automatic key rotation.

C. Create a new HSM-encrypted Amazon Redshift cluster and migrate the data to the new cluster.

D. Enable HSM with key rotation through the AWS CLI.

E. Enable Elliptic Curve Diffie-Hellman Ephemeral (ECDHE) encryption in the HSM.

Question #13 *Topic 1*

A company is planning to do a proof of concept for a machine learning (ML) project using Amazon SageMaker with a subset of existing on-premises data hosted in the company's 3 TB data warehouse. For part of the project, AWS Direct Connect is established and tested. To prepare the data for ML, data analysts are performing data curation. The data analysts want to perform multiple step, including mapping, dropping null fields, resolving choice, and splitting fields. The company needs the fastest solution to curate the data for this project.
Which solution meets these requirements?

A. Ingest data into Amazon S3 using AWS DataSync and use Apache Spark scrips to curate the data in an Amazon EMR cluster. Store the curated data in Amazon S3 for ML processing.

B. Create custom ETL jobs on-premises to curate the data. Use AWS DMS to ingest data into Amazon S3 for ML processing.

C. Ingest data into Amazon S3 using AWS DMS. Use AWS Glue to perform data curation and store the data in Amazon S3 for ML processing.

D. Take a full backup of the data store and ship the backup files using AWS Snowball. Upload Snowball data into Amazon S3 and schedule data curation jobs using AWS Batch to prepare the data for ML.

Question #14 *Topic 1*

A US-based sneaker retail company launched its global website. All the transaction data is stored in Amazon RDS and curated historic transaction data is stored in Amazon Redshift in the us-east-1 Region. The business intelligence (BI) team wants to enhance the user experience by providing a dashboard for sneaker trends.
The BI team decides to use Amazon QuickSight to render the website dashboards. During development, a team in Japan provisioned Amazon QuickSight in ap- northeast-1. The team is having difficulty connecting Amazon QuickSight from ap-northeast-1 to Amazon Redshift in us-east-1.
Which solution will solve this issue and meet the requirements?

A. In the Amazon Redshift console, choose to configure cross-Region snapshots and set the destination Region as ap-northeast-1. Restore the Amazon Redshift Cluster from the snapshot and connect to Amazon QuickSight launched in ap-northeast-1.

B. Create a VPC endpoint from the Amazon QuickSight VPC to the Amazon Redshift VPC so Amazon QuickSight can access data from Amazon Redshift.

C. Create an Amazon Redshift endpoint connection string with Region information in the string and use this connection string in Amazon QuickSight to connect to Amazon Redshift.

D. Create a new security group for Amazon Redshift in us-east-1 with an inbound rule authorizing access from the appropriate IP address range for the Amazon QuickSight servers in ap-northeast-1.

Question #15 *Topic 1*

An airline has .csv-formatted data stored in Amazon S3 with an AWS Glue Data Catalog. Data analysts want to join this data with call center data stored in
Amazon Redshift as part of a dally batch process. The Amazon Redshift cluster is already under a heavy load. The solution must be managed, serverless, well- functioning, and minimize the load on the existing Amazon Redshift cluster. The solution should also require minimal effort and development activity.
Which solution meets these requirements?

A. Unload the call center data from Amazon Redshift to Amazon S3 using an AWS Lambda function. Perform the join with AWS Glue ETL scripts.

B. Export the call center data from Amazon Redshift using a Python shell in AWS Glue. Perform the join with AWS Glue ETL scripts.

C. Create an external table using Amazon Redshift Spectrum for the call center data and perform the join with Amazon Redshift.

D. Export the call center data from Amazon Redshift to Amazon EMR using Apache Sqoop. Perform the join with Apache Hive.

Question #16 *Topic 1*

A data analyst is using Amazon QuickSight for data visualization across multiple datasets generated by applications. Each application stores files within a separate Amazon S3 bucket. AWS Glue Data Catalog is used as a central catalog across all application data in Amazon S3. A new application stores its data within a separate S3 bucket. After updating the catalog to include the new application data source, the data analyst created a new Amazon QuickSight data source from an Amazon Athena table, but the import into SPICE failed.
How should the data analyst resolve the issue?

A. Edit the permissions for the AWS Glue Data Catalog from within the Amazon QuickSight console.

B. Edit the permissions for the new S3 bucket from within the Amazon QuickSight console.

C. Edit the permissions for the AWS Glue Data Catalog from within the AWS Glue console.

D. Edit the permissions for the new S3 bucket from within the S3 console.

Question #17 *Topic 1*

A team of data scientists plans to analyze market trend data for their company's new investment strategy. The trend data comes from five different data sources in large volumes. The team wants to utilize Amazon Kinesis to support their use case. The team uses SQL-like queries to analyze trends and wants to send notifications based on certain significant patterns in the trends. Additionally, the data scientists want to save the data to Amazon S3 for archival and historical re- processing, and use AWS managed services wherever possible. The team wants to implement the lowest-cost solution.
Which solution meets these requirements?

A. Publish data to one Kinesis data stream. Deploy a custom application using the Kinesis Client Library (KCL) for analyzing trends, and send notifications using Amazon SNS. Configure Kinesis Data Firehose on the Kinesis data stream to persist data to an S3 bucket.

B. Publish data to one Kinesis data stream. Deploy Kinesis Data Analytic to the stream for analyzing trends, and configure an AWS Lambda function as an output to send notifications using Amazon SNS. Configure Kinesis Data Firehose on the Kinesis data stream to persist data to an S3 bucket.

C. Publish data to two Kinesis data streams. Deploy Kinesis Data Analytics to the first stream for analyzing trends, and configure an AWS Lambda function as an output to send notifications using Amazon SNS. Configure Kinesis Data Firehose on the second Kinesis data stream to persist data to an S3 bucket.

D. Publish data to two Kinesis data streams. Deploy a custom application using the Kinesis Client Library (KCL) to the first stream for analyzing trends, and send notifications using Amazon SNS. Configure Kinesis Data Firehose on the second Kinesis data stream to persist data to an S3 bucket.

Question #18

Topic 1

A company currently uses Amazon Athena to query its global datasets. The regional data is stored in Amazon S3 in the us-east-1 and us-west-2 Regions. The data is not encrypted. To simplify the query process and manage it centrally, the company wants to use Athena in us-west-2 to query data from Amazon S3 in both
Regions. The solution should be as low-cost as possible.
What should the company do to achieve this goal?

A. Use AWS DMS to migrate the AWS Glue Data Catalog from us-east-1 to us-west-2. Run Athena queries in us-west-2.

B. Run the AWS Glue crawler in us-west-2 to catalog datasets in all Regions. Once the data is crawled, run Athena queries in us-west-2.

C. Enable cross-Region replication for the S3 buckets in us-east-1 to replicate data in us-west-2. Once the data is replicated in us-west-2, run the AWS Glue crawler there to update the AWS Glue Data Catalog in us-west-2 and run Athena queries.

D. Update AWS Glue resource policies to provide us-east-1 AWS Glue Data Catalog access to us-west-2. Once the catalog in us-west-2 has access to the catalog in us-east-1, run Athena queries in us-west-2.

Question #19

Topic 1

A large company receives files from external parties in Amazon EC2 throughout the day. At the end of the day, the files are combined into a single file, compressed into a gzip file, and uploaded to Amazon S3. The total size of all the files is close to 100 GB daily. Once the files are uploaded to Amazon S3, an
AWS Batch program executes a COPY command to load the files into an Amazon Redshift cluster.
Which program modification will accelerate the COPY process?

A. Upload the individual files to Amazon S3 and run the COPY command as soon as the files become available.

B. Split the number of files so they are equal to a multiple of the number of slices in the Amazon Redshift cluster. Gzip and upload the files to Amazon S3. Run the COPY command on the files.

C. Split the number of files so they are equal to a multiple of the number of compute nodes in the Amazon Redshift cluster. Gzip and upload the files to Amazon S3. Run the COPY command on the files.

D. Apply sharding by breaking up the files so the distkey columns with the same values go to the same file. Gzip and upload the sharded files to Amazon S3. Run the COPY command on the files.

Question #20

Topic 1

A large ride-sharing company has thousands of drivers globally serving millions of unique customers every day. The company has decided to migrate an existing data mart to Amazon Redshift. The existing schema includes the following tables.
⚯ A trips fact table for information on completed rides.
⚯ A drivers dimension table for driver profiles.
⚯ A customers fact table holding customer profile information.
The company analyzes trip details by date and destination to examine profitability by region. The drivers data rarely changes. The customers data frequently changes.
What table design provides optimal query performance?

A. Use DISTSTYLE KEY (destination) for the trips table and sort by date. Use DISTSTYLE ALL for the drivers and customers tables.

B. Use DISTSTYLE EVEN for the trips table and sort by date. Use DISTSTYLE ALL for the drivers table. Use DISTSTYLE EVEN for the customers table.

C. Use DISTSTYLE KEY (destination) for the trips table and sort by date. Use DISTSTYLE ALL for the drivers table. Use DISTSTYLE EVEN for the customers table.

D. Use DISTSTYLE EVEN for the drivers table and sort by date. Use DISTSTYLE ALL for both fact tables.

Question #21 *Topic 1*

Three teams of data analysts use Apache Hive on an Amazon EMR cluster with the EMR File System (EMRFS) to query data stored within each teams Amazon
S3 bucket. The EMR cluster has Kerberos enabled and is configured to authenticate users from the corporate Active Directory. The data is highly sensitive, so access must be limited to the members of each team.
Which steps will satisfy the security requirements?

A. For the EMR cluster Amazon EC2 instances, create a service role that grants no access to Amazon S3. Create three additional IAM roles, each granting access to each team's specific bucket. Add the additional IAM roles to the cluster's EMR role for the EC2 trust policy. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.

B. For the EMR cluster Amazon EC2 instances, create a service role that grants no access to Amazon S3. Create three additional IAM roles, each granting access to each team's specific bucket. Add the service role for the EMR cluster EC2 instances to the trust policies for the additional IAM roles. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.

C. For the EMR cluster Amazon EC2 instances, create a service role that grants full access to Amazon S3. Create three additional IAM roles, each granting access to each team's specific bucket. Add the service role for the EMR cluster EC2 instances to the trust polices for the additional IAM roles. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.

D. For the EMR cluster Amazon EC2 instances, create a service role that grants full access to Amazon S3. Create three additional IAM roles, each granting access to each team's specific bucket. Add the service role for the EMR cluster EC2 instances to the trust polices for the base IAM roles. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.

Question #22 *Topic 1*

A company is planning to create a data lake in Amazon S3. The company wants to create tiered storage based on access patterns and cost objectives. The solution must include support for JDBC connections from legacy clients, metadata management that allows federation for access control, and batch-based ETL using PySpark and Scala. Operational management should be limited.
Which combination of components can meet these requirements? (Choose three.)

A. AWS Glue Data Catalog for metadata management

B. Amazon EMR with Apache Spark for ETL

C. AWS Glue for Scala-based ETL

D. Amazon EMR with Apache Hive for JDBC clients

E. Amazon Athena for querying data in Amazon S3 using JDBC drivers

F. Amazon EMR with Apache Hive, using an Amazon RDS with MySQL-compatible backed metastore

Question #23 *Topic 1*

A company wants to optimize the cost of its data and analytics platform. The company is ingesting a number of .csv and JSON files in Amazon S3 from various data sources. Incoming data is expected to be 50 GB each day. The company is using Amazon Athena to query the raw data in Amazon S3 directly. Most queries aggregate data from the past 12 months, and data that is older than 5 years is infrequently queried. The typical query scans about 500 MB of data and is expected to return results in less than 1 minute. The raw data must be retained indefinitely for compliance requirements.
Which solution meets the company's requirements?

A. Use an AWS Glue ETL job to compress, partition, and convert the data into a columnar data format. Use Athena to query the processed dataset. Configure a lifecycle policy to move the processed data into the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class 5 years after object creation. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after object creation.

B. Use an AWS Glue ETL job to partition and convert the data into a row-based data format. Use Athena to query the processed dataset. Configure a lifecycle policy to move the data into the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class 5 years after object creation. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after object creation.

C. Use an AWS Glue ETL job to compress, partition, and convert the data into a columnar data format. Use Athena to query the processed dataset. Configure a lifecycle policy to move the processed data into the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class 5 years after the object was last accessed. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after the last date the object was accessed.

D. Use an AWS Glue ETL job to partition and convert the data into a row-based data format. Use Athena to query the processed dataset. Configure a lifecycle policy to move the data into the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class 5 years after the object was last accessed. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after the last date the object was accessed.

Question #24 *Topic 1*

An energy company collects voltage data in real time from sensors that are attached to buildings. The company wants to receive notifications when a sequence of two voltage drops is detected within 10 minutes of a sudden voltage increase at the same building. All notifications must be delivered as quickly as possible. The system must be highly available. The company needs a solution that will automatically scale when this monitoring feature is implemented in other cities. The notification system is subscribed to an Amazon Simple Notification Service (Amazon SNS) topic for remediation.
Which solution will meet these requirements?

A. Create an Amazon Managed Streaming for Apache Kafka cluster to ingest the data. Use an Apache Spark Streaming with Apache Kafka consumer API in an automatically scaled Amazon EMR cluster to process the incoming data. Use the Spark Streaming application to detect the known event sequence and send the SNS message.

B. Create a REST-based web service by using Amazon API Gateway in front of an AWS Lambda function. Create an Amazon RDS for PostgreSQL database with sufficient Provisioned IOPS to meet current demand. Configure the Lambda function to store incoming events in the RDS for PostgreSQL database, query the latest data to detect the known event sequence, and send the SNS message.

C. Create an Amazon Kinesis Data Firehose delivery stream to capture the incoming sensor data. Use an AWS Lambda transformation function to detect the known event sequence and send the SNS message.

D. Create an Amazon Kinesis data stream to capture the incoming sensor data. Create another stream for notifications. Set up AWS Application Auto Scaling on both streams. Create an Amazon Kinesis Data Analytics for Java application to detect the known event sequence, and add a message to the message stream Configure an AWS Lambda function to poll the message stream and publish to the SNS topic.

Question #25 *Topic 1*

A media company has a streaming playback application. The company needs to collect and analyze data to provide near-real-time feedback on playback issues within 30 seconds. The company requires a consumer application to identify playback issues, such as decreased quality during a specified time frame. The data will be streamed in JSON format. The schema can change over time.
Which solution will meet these requirements?

A. Send the data to Amazon Kinesis Data Firehose with delivery to Amazon S3. Configure an S3 event to invoke an AWS Lambda function to process and analyze the data.

B. Send the data to Amazon Managed Streaming for Apache Kafka. Configure Amazon Kinesis Data Analytics for SQL Application as the consumer application to process and analyze the data.

C. Send the data to Amazon Kinesis Data Firehose with delivery to Amazon S3. Configure Amazon S3 to initiate an event for AWS Lambda to process and analyze the data.

D. Send the data to Amazon Kinesis Data Streams. Configure an Amazon Kinesis Data Analytics for Apache Flink application as the consumer application to process and analyze the data.

Question #26 *Topic 1*

An ecommerce company stores customer purchase data in Amazon RDS. The company wants a solution to store and analyze historical data. The most recent 6 months of data will be queried frequently for analytics workloads. This data is several terabytes large. Once a month, historical data for the last 5 years must be accessible and will be joined with the more recent data. The company wants to optimize performance and cost.
Which storage solution will meet these requirements?

A. Create a read replica of the RDS database to store the most recent 6 months of data. Copy the historical data into Amazon S3. Create an AWS Glue Data Catalog of the data in Amazon S3 and Amazon RDS. Run historical queries using Amazon Athena.

B. Use an ETL tool to incrementally load the most recent 6 months of data into an Amazon Redshift cluster. Run more frequent queries against this cluster. Create a read replica of the RDS database to run queries on the historical data.

C. Incrementally copy data from Amazon RDS to Amazon S3. Create an AWS Glue Data Catalog of the data in Amazon S3. Use Amazon Athena to query the data.

D. Incrementally copy data from Amazon RDS to Amazon S3. Load and store the most recent 6 months of data in Amazon Redshift. Configure an Amazon Redshift Spectrum table to connect to all historical data.

Question #27 *Topic 1*

A company leverages Amazon Athena for ad-hoc queries against data stored in Amazon S3. The company wants to implement additional controls to separate query execution and query history among users, teams, or applications running in the same AWS account to comply with internal security policies.
Which solution meets these requirements?

A. Create an S3 bucket for each given use case, create an S3 bucket policy that grants permissions to appropriate individual IAM users. and apply the S3 bucket policy to the S3 bucket.

B. Create an Athena workgroup for each given use case, apply tags to the workgroup, and create an IAM policy using the tags to apply appropriate permissions to the workgroup.

C. Create an IAM role for each given use case, assign appropriate permissions to the role for the given use case, and add the role to associate the role with Athena.

D. Create an AWS Glue Data Catalog resource policy for each given use case that grants permissions to appropriate individual IAM users, and apply the resource policy to the specific tables used by Athena.

Question #28 *Topic 1*

A company wants to use an automatic machine learning (ML) Random Cut Forest (RCF) algorithm to visualize complex real-world scenarios, such as detecting seasonality and trends, excluding outers, and imputing missing values.
The team working on this project is non-technical and is looking for an out-of-the-box solution that will require the LEAST amount of management overhead.
Which solution will meet these requirements?

A. Use an AWS Glue ML transform to create a forecast and then use Amazon QuickSight to visualize the data.

B. Use Amazon QuickSight to visualize the data and then use ML-powered forecasting to forecast the key business metrics.

C. Use a pre-build ML AMI from the AWS Marketplace to create forecasts and then use Amazon QuickSight to visualize the data.

D. Use calculated fields to create a new forecast and then use Amazon QuickSight to visualize the data.

Question #29 *Topic 1*

A retail company's data analytics team recently created multiple product sales analysis dashboards for the average selling price per product using Amazon
QuickSight. The dashboards were created from .csv files uploaded to Amazon S3. The team is now planning to share the dashboards with the respective external product owners by creating individual users in Amazon QuickSight. For compliance and governance reasons, restricting access is a key requirement. The product owners should view only their respective product analysis in the dashboard reports.
Which approach should the data analytics team take to allow product owners to view only their products in the dashboard?

A. Separate the data by product and use S3 bucket policies for authorization.

B. Separate the data by product and use IAM policies for authorization.

C. Create a manifest file with row-level security.

D. Create dataset rules with row-level security.

Question #30 *Topic 1*

A company has developed an Apache Hive script to batch process data stared in Amazon S3. The script needs to run once every day and store the output in
Amazon S3. The company tested the script, and it completes within 30 minutes on a small local three-node cluster.
Which solution is the MOST cost-effective for scheduling and executing the script?

A. Create an AWS Lambda function to spin up an Amazon EMR cluster with a Hive execution step. Set KeepJobFlowAliveWhenNoSteps to false and disable the termination protection flag. Use Amazon CloudWatch Events to schedule the Lambda function to run daily.

B. Use the AWS Management Console to spin up an Amazon EMR cluster with Python Hue. Hive, and Apache Oozie. Set the termination protection flag to true and use Spot Instances for the core nodes of the cluster. Configure an Oozie workflow in the cluster to invoke the Hive script daily.

C. Create an AWS Glue job with the Hive script to perform the batch operation. Configure the job to run once a day using a time-based schedule.

D. Use AWS Lambda layers and load the Hive runtime to AWS Lambda and copy the Hive script. Schedule the Lambda function to run daily by creating a workflow using AWS Step Functions.

Question #31 *Topic 1*

A company wants to improve the data load time of a sales data dashboard. Data has been collected as .csv files and stored within an Amazon S3 bucket that is partitioned by date. The data is then loaded to an Amazon Redshift data warehouse for frequent analysis. The data volume is up to 500 GB per day.
Which solution will improve the data loading performance?

A. Compress .csv files and use an INSERT statement to ingest data into Amazon Redshift.

B. Split large .csv files, then use a COPY command to load data into Amazon Redshift.

C. Use Amazon Kinesis Data Firehose to ingest data into Amazon Redshift.

D. Load the .csv files in an unsorted key order and vacuum the table in Amazon Redshift.

Question #32 *Topic 1*

A company has a data warehouse in Amazon Redshift that is approximately 500 TB in size. New data is imported every few hours and read-only queries are run throughout the day and evening. There is a particularly heavy load with no writes for several hours each morning on business days. During those hours, some queries are queued and take a long time to execute. The company needs to optimize query execution and avoid any downtime.
What is the MOST cost-effective solution?

A. Enable concurrency scaling in the workload management (WLM) queue.

B. Add more nodes using the AWS Management Console during peak hours. Set the distribution style to ALL.

C. Use elastic resize to quickly add nodes during peak times. Remove the nodes when they are not needed.

D. Use a snapshot, restore, and resize operation. Switch to the new target cluster.

Question #33 *Topic 1*

A company analyzes its data in an Amazon Redshift data warehouse, which currently has a cluster of three dense storage nodes. Due to a recent business acquisition, the company needs to load an additional 4 TB of user data into Amazon Redshift. The engineering team will combine all the user data and apply complex calculations that require I/O intensive resources. The company needs to adjust the cluster's capacity to support the change in analytical and storage requirements.
Which solution meets these requirements?

A. Resize the cluster using elastic resize with dense compute nodes.

B. Resize the cluster using classic resize with dense compute nodes.

C. Resize the cluster using elastic resize with dense storage nodes.

D. Resize the cluster using classic resize with dense storage nodes.

Question #34 *Topic 1*

A company stores its sales and marketing data that includes personally identifiable information (PII) in Amazon S3. The company allows its analysts to launch their own Amazon EMR cluster and run analytics reports with the data. To meet compliance requirements, the company must ensure the data is not publicly accessible throughout this process. A data engineer has secured Amazon S3 but must ensure the individual EMR clusters created by the analysts are not exposed to the public internet.
Which solution should the data engineer to meet this compliance requirement with LEAST amount of effort?

A. Create an EMR security configuration and ensure the security configuration is associated with the EMR clusters when they are created.

B. Check the security group of the EMR clusters regularly to ensure it does not allow inbound traffic from IPv4 0.0.0.0/0 or IPv6 ::/0.

C. Enable the block public access setting for Amazon EMR at the account level before any EMR cluster is created.

D. Use AWS WAF to block public internet access to the EMR clusters across the board.

Question #35 *Topic 1*

A financial company uses Amazon S3 as its data lake and has set up a data warehouse using a multi-node Amazon Redshift cluster. The data files in the data lake are organized in folders based on the data source of each data file. All the data files are loaded to one table in the Amazon Redshift cluster using a separate
COPY command for each data file location. With this approach, loading all the data files into Amazon Redshift takes a long time to complete.
Users want a faster solution with little or no increase in cost while maintaining the segregation of the data files in the S3 data lake.
Which solution meets these requirements?

A. Use Amazon EMR to copy all the data files into one folder and issue a COPY command to load the data into Amazon Redshift.

B. Load all the data files in parallel to Amazon Aurora, and 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 and issue a COPY command to load the data into Amazon Redshift.

D. Create a manifest file that contains the data file locations and issue a COPY command to load the data into Amazon Redshift.

Question #36 *Topic 1*

A company's marketing team has asked for help in identifying a high performing long-term storage service for their data based on the following requirements:
⊶ The data size is approximately 32 TB uncompressed.
⊶ There is a low volume of single-row inserts each day.
⊶ There is a high volume of aggregation queries each day.
⊶ Multiple complex joins are performed.
⊶ The queries typically involve a small subset of the columns in a table.
Which storage service will provide the MOST performant solution?

A. Amazon Aurora MySQL

B. Amazon Redshift

C. Amazon Neptune

D. Amazon Elasticsearch

Question #37 *Topic 1*

A technology company is creating a dashboard that will visualize and analyze time-sensitive data. The data will come in through Amazon Kinesis Data Firehose with the butter interval set to 60 seconds. The dashboard must support near-real-time data.
Which visualization solution will meet these requirements?

A. Select Amazon OpenSearch Service (Amazon Elasticsearch Service) as the endpoint for Kinesis Data Firehose. Set up an OpenSearch Dashboards (Kibana) using the data in Amazon OpenSearch Service (Amazon ES) with the desired analyses and visualizations.

B. Select Amazon S3 as the endpoint for Kinesis Data Firehose. Read data into an Amazon SageMaker Jupyter notebook and carry out the desired analyses and visualizations.

C. Select Amazon Redshift as the endpoint for Kinesis Data Firehose. Connect Amazon QuickSight with SPICE to Amazon Redshift to create the desired analyses and visualizations.

D. Select Amazon S3 as the endpoint for Kinesis Data Firehose. Use AWS Glue to catalog the data and Amazon Athena to query it. Connect Amazon QuickSight with SPICE to Athena to create the desired analyses and visualizations.

Question #38 *Topic 1*

A financial company uses Apache Hive on Amazon EMR for ad-hoc queries. Users are complaining of sluggish performance.
A data analyst notes the following:
✑ Approximately 90% of queries are submitted 1 hour after the market opens.
Hadoop Distributed File System (HDFS) utilization never exceeds 10%.
▪
Which solution would help address the performance issues?

A. Create instance fleet configurations for core and task nodes. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch CapacityRemainingGB metric. Create an automatic scaling policy to scale in the instance fleet based on the CloudWatch CapacityRemainingGB metric.

B. Create instance fleet configurations for core and task nodes. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch YARNMemoryAvailablePercentage metric. Create an automatic scaling policy to scale in the instance fleet based on the CloudWatch YARNMemoryAvailablePercentage metric.

C. Create instance group configurations for core and task nodes. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch CapacityRemainingGB metric. Create an automatic scaling policy to scale in the instance groups based on the CloudWatch CapacityRemainingGB metric.

D. Create instance group configurations for core and task nodes. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch YARNMemoryAvailablePercentage metric. Create an automatic scaling policy to scale in the instance groups based on the CloudWatch YARNMemoryAvailablePercentage metric.

Question #39 *Topic 1*

A media company has been performing analytics on log data generated by its applications. There has been a recent increase in the number of concurrent analytics jobs running, and the overall performance of existing jobs is decreasing as the number of new jobs is increasing. The partitioned data is stored in
Amazon S3 One Zone-Infrequent Access (S3 One Zone-IA) and the analytic processing is performed on Amazon EMR clusters using the EMR File System
(EMRFS) with consistent view enabled. A data analyst has determined that it is taking longer for the EMR task nodes to list objects in Amazon S3.
Which action would MOST likely increase the performance of accessing log data in Amazon S3?

A. Use a hash function to create a random string and add that to the beginning of the object prefixes when storing the log data in Amazon S3.

B. Use a lifecycle policy to change the S3 storage class to S3 Standard for the log data.

C. Increase the read capacity units (RCUs) for the shared Amazon DynamoDB table.

D. Redeploy the EMR clusters that are running slowly to a different Availability Zone.

Question #40 *Topic 1*

A company has developed several AWS Glue jobs to validate and transform its data from Amazon S3 and load it into Amazon RDS for MySQL in batches once every day. The ETL jobs read the S3 data using a DynamicFrame. Currently, the ETL developers are experiencing challenges in processing only the incremental data on every run, as the AWS Glue job processes all the S3 input data on each run.
Which approach would allow the developers to solve the issue with minimal coding effort?

A. Have the ETL jobs read the data from Amazon S3 using a DataFrame.

B. Enable job bookmarks on the AWS Glue jobs.

C. Create custom logic on the ETL jobs to track the processed S3 objects.

D. Have the ETL jobs delete the processed objects or data from Amazon S3 after each run.

Question #41 *Topic 1*

A mortgage company has a microservice for accepting payments. This microservice uses the Amazon DynamoDB encryption client with AWS KMS managed keys to encrypt the sensitive data before writing the data to DynamoDB. The finance team should be able to load this data into Amazon Redshift and aggregate the values within the sensitive fields. The Amazon Redshift cluster is shared with other data analysts from different business units.
Which steps should a data analyst take to accomplish this task efficiently and securely?

A. Create an AWS Lambda function to process the DynamoDB stream. Decrypt the sensitive data using the same KMS key. Save the output to a restricted S3 bucket for the finance team. Create a finance table in Amazon Redshift that is accessible to the finance team only. Use the COPY command to load the data from Amazon S3 to the finance table.

B. Create an AWS Lambda function to process the DynamoDB stream. Save the output to a restricted S3 bucket for the finance team. Create a finance table in Amazon Redshift that is accessible to the finance team only. Use the COPY command with the IAM role that has access to the KMS key to load the data from S3 to the finance table.

C. Create an Amazon EMR cluster with an EMR_EC2_DefaultRole role that has access to the KMS key. Create Apache Hive tables that reference the data stored in DynamoDB and the finance table in Amazon Redshift. In Hive, select the data from DynamoDB and then insert the output to the finance table in Amazon Redshift.

D. Create an Amazon EMR cluster. Create Apache Hive tables that reference the data stored in DynamoDB. Insert the output to the restricted Amazon S3 bucket for the finance team. Use the COPY command with the IAM role that has access to the KMS key to load the data from Amazon S3 to the finance table in Amazon Redshift.

Question #42 *Topic 1*

A company is building a data lake and needs to ingest data from a relational database that has time-series data. The company wants to use managed services to accomplish this. The process needs to be scheduled daily and bring incremental data only from the source into Amazon S3.
What is the MOST cost-effective approach to meet these requirements?

A. Use AWS Glue to connect to the data source using JDBC Drivers. Ingest incremental records only using job bookmarks.

B. Use AWS Glue to connect to the data source using JDBC Drivers. Store the last updated key in an Amazon DynamoDB table and ingest the data using the updated key as a filter.

C. Use AWS Glue to connect to the data source using JDBC Drivers and ingest the entire dataset. Use appropriate Apache Spark libraries to compare the dataset, and find the delta.

D. Use AWS Glue to connect to the data source using JDBC Drivers and ingest the full data. Use AWS DataSync to ensure the delta only is written into Amazon S3.

Question #43 *Topic 1*

An Amazon Redshift database contains sensitive user data. Logging is necessary to meet compliance requirements. The logs must contain database authentication attempts, connections, and disconnections. The logs must also contain each query run against the database and record which database user ran each query.
Which steps will create the required logs?

A. Enable Amazon Redshift Enhanced VPC Routing. Enable VPC Flow Logs to monitor traffic.

B. Allow access to the Amazon Redshift database using AWS IAM only. Log access using AWS CloudTrail.

C. Enable audit logging for Amazon Redshift using the AWS Management Console or the AWS CLI.

D. Enable and download audit reports from AWS Artifact.

Question #44 *Topic 1*

A company that monitors weather conditions from remote construction sites is setting up a solution to collect temperature data from the following two weather stations.
⊶ Station A, which has 10 sensors
⊶ Station B, which has five sensors
These weather stations were placed by onsite subject-matter experts.
Each sensor has a unique ID. The data collected from each sensor will be collected using Amazon Kinesis Data Streams.
Based on the total incoming and outgoing data throughput, a single Amazon Kinesis data stream with two shards is created. Two partition keys are created based on the station names. During testing, there is a bottleneck on data coming from Station A, but not from Station B. Upon review, it is confirmed that the total stream throughput is still less than the allocated Kinesis Data Streams throughput.
How can this bottleneck be resolved without increasing the overall cost and complexity of the solution, while retaining the data collection quality requirements?

A. Increase the number of shards in Kinesis Data Streams to increase the level of parallelism.

B. Create a separate Kinesis data stream for Station A with two shards, and stream Station A sensor data to the new stream.

C. Modify the partition key to use the sensor ID instead of the station name.

D. Reduce the number of sensors in Station A from 10 to 5 sensors.

Question #45 *Topic 1*

Once a month, a company receives a 100 MB .csv file compressed with gzip. The file contains 50,000 property listing records and is stored in Amazon S3 Glacier.

The company needs its data analyst to query a subset of the data for a specific vendor.

What is the most cost-effective solution?

A. Load the data into Amazon S3 and query it with Amazon S3 Select.

B. Query the data from Amazon S3 Glacier directly with Amazon Glacier Select.

C. Load the data to Amazon S3 and query it with Amazon Athena.

D. Load the data to Amazon S3 and query it with Amazon Redshift Spectrum.

Question #46 *Topic 1*

A retail company is building its data warehouse solution using Amazon Redshift. As a part of that effort, the company is loading hundreds of files into the fact table created in its Amazon Redshift cluster. The company wants the solution to achieve the highest throughput and optimally use cluster resources when loading data into the company's fact table.

How should the company meet these requirements?

A. Use multiple COPY commands to load the data into the Amazon Redshift cluster.

B. Use S3DistCp to load multiple files into the Hadoop Distributed File System (HDFS) and use an HDFS connector to ingest the data into the Amazon Redshift cluster.

C. Use LOAD commands equal to the number of Amazon Redshift cluster nodes and load the data in parallel into each node.

D. Use a single COPY command to load the data into the Amazon Redshift cluster.

Question #47 *Topic 1*

A data analyst is designing a solution to interactively query datasets with SQL using a JDBC connection. Users will join data stored in Amazon S3 in Apache ORC format with data stored in Amazon OpenSearch Service (Amazon Elasticsearch Service) and Amazon Aurora MySQL.

Which solution will provide the MOST up-to-date results?

A. Use AWS Glue jobs to ETL data from Amazon ES and Aurora MySQL to Amazon S3. Query the data with Amazon Athena.

B. Use Amazon DMS to stream data from Amazon ES and Aurora MySQL to Amazon Redshift. Query the data with Amazon Redshift.

C. Query all the datasets in place with Apache Spark SQL running on an AWS Glue developer endpoint.

D. Query all the datasets in place with Apache Presto running on Amazon EMR.

Question #48

Topic 1

A company developed a new elections reporting website that uses Amazon Kinesis Data Firehose to deliver full logs from AWS WAF to an Amazon S3 bucket.

The company is now seeking a low-cost option to perform this infrequent data analysis with visualizations of logs in a way that requires minimal development effort.

Which solution meets these requirements?

A. Use an AWS Glue crawler to create and update a table in the Glue data catalog from the logs. Use Athena to perform ad-hoc analyses and use Amazon QuickSight to develop data visualizations.

B. Create a second Kinesis Data Firehose delivery stream to deliver the log files to Amazon OpenSearch Service (Amazon Elasticsearch Service). Use Amazon ES to perform text-based searches of the logs for ad-hoc analyses and use OpenSearch Dashboards (Kibana) for data visualizations.

C. Create an AWS Lambda function to convert the logs into .csv format. Then add the function to the Kinesis Data Firehose transformation configuration. Use Amazon Redshift to perform ad-hoc analyses of the logs using SQL queries and use Amazon QuickSight to develop data visualizations.

D. Create an Amazon EMR cluster and use Amazon S3 as the data source. Create an Apache Spark job to perform ad-hoc analyses and use Amazon QuickSight to develop data visualizations.

Question #49

Topic 1

A large company has a central data lake to run analytics across different departments. Each department uses a separate AWS account and stores its data in an

Amazon S3 bucket in that account. Each AWS account uses the AWS Glue Data Catalog as its data catalog. There are different data lake access requirements based on roles. Associate analysts should only have read access to their departmental data. Senior data analysts can have access in multiple departments including theirs, but for a subset of columns only.

Which solution achieves these required access patterns to minimize costs and administrative tasks?

A. Consolidate all AWS accounts into one account. Create different S3 buckets for each department and move all the data from every account to the central data lake account. Migrate the individual data catalogs into a central data catalog and apply fine-grained permissions to give to each user the required access to tables and databases in AWS Glue and Amazon S3.

B. Keep the account structure and the individual AWS Glue catalogs on each account. Add a central data lake account and use AWS Glue to catalog data from various accounts. Configure cross-account access for AWS Glue crawlers to scan the data in each departmental S3 bucket to identify the schema and populate the catalog. Add the senior data analysts into the central account and apply highly detailed access controls in the Data Catalog and Amazon S3.

C. Set up an individual AWS account for the central data lake. Use AWS Lake Formation to catalog the cross-account locations. On each individual S3 bucket, modify the bucket policy to grant S3 permissions to the Lake Formation service-linked role. Use Lake Formation permissions to add fine-grained access controls to allow senior analysts to view specific tables and columns.

D. Set up an individual AWS account for the central data lake and configure a central S3 bucket. Use an AWS Lake Formation blueprint to move the data from the various buckets into the central S3 bucket. On each individual bucket, modify the bucket policy to grant S3 permissions to the Lake Formation service-linked role. Use Lake Formation permissions to add fine-grained access controls for both associate and senior analysts to view specific tables and columns.

Question #50 *Topic 1*

A company wants to improve user satisfaction for its smart home system by adding more features to its recommendation engine. Each sensor asynchronously pushes its nested JSON data into Amazon Kinesis Data Streams using the Kinesis Producer Library (KPL) in Java. Statistics from a set of failed sensors showed that, when a sensor is malfunctioning, its recorded data is not always sent to the cloud.
The company needs a solution that offers near-real-time analytics on the data from the most updated sensors.
Which solution enables the company to meet these requirements?

A. Set the RecordMaxBufferedTime property of the KPL to "1’ˆג" to disable the buffering on the sensor side. Use Kinesis Data Analytics to enrich the data based on a company-developed anomaly detection SQL script. Push the enriched data to a fleet of Kinesis data streams and enable the data transformation feature to flatten the JSON file. Instantiate a dense storage Amazon Redshift cluster and use it as the destination for the Kinesis Data Firehose delivery stream.

B. Update the sensors code to use the PutRecord/PutRecords call from the Kinesis Data Streams API with the AWS SDK for Java. Use Kinesis Data Analytics to enrich the data based on a company-developed anomaly detection SQL script. Direct the output of KDA application to a Kinesis Data Firehose delivery stream, enable the data transformation feature to flatten the JSON file, and set the Kinesis Data Firehose destination to an Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster.

C. Set the RecordMaxBufferedTime property of the KPL to "0" to disable the buffering on the sensor side. Connect for each stream a dedicated Kinesis Data Firehose delivery stream and enable the data transformation feature to flatten the JSON file before sending it to an Amazon S3 bucket. Load the S3 data into an Amazon Redshift cluster.

D. Update the sensors code to use the PutRecord/PutRecords call from the Kinesis Data Streams API with the AWS SDK for Java. Use AWS Glue to fetch and process data from the stream using the Kinesis Client Library (KCL). Instantiate an Amazon Elasticsearch Service cluster and use AWS Lambda to directly push data into it.

Question #51 *Topic 1*

A global company has different sub-organizations, and each sub-organization sells its products and services in various countries. The company's senior leadership wants to quickly identify which sub-organization is the strongest performer in each country. All sales data is stored in Amazon S3 in Parquet format.
Which approach can provide the visuals that senior leadership requested with the least amount of effort?

A. Use Amazon QuickSight with Amazon Athena as the data source. Use heat maps as the visual type.

B. Use Amazon QuickSight with Amazon S3 as the data source. Use heat maps as the visual type.

C. Use Amazon QuickSight with Amazon Athena as the data source. Use pivot tables as the visual type.

D. Use Amazon QuickSight with Amazon S3 as the data source. Use pivot tables as the visual type.

Question #52 *Topic 1*

A company has 1 million scanned documents stored as image files in Amazon S3. The documents contain typewritten application forms with information including the applicant first name, applicant last name, application date, application type, and application text. The company has developed a machine learning algorithm to extract the metadata values from the scanned documents. The company wants to allow internal data analysts to analyze and find applications using the applicant name, application date, or application text. The original images should also be downloadable. Cost control is secondary to query performance.
Which solution organizes the images and metadata to drive insights while meeting the requirements?

A. For each image, use object tags to add the metadata. Use Amazon S3 Select to retrieve the files based on the applicant name and application date.

B. Index the metadata and the Amazon S3 location of the image file in Amazon OpenSearch Service (Amazon Elasticsearch Service). Allow the data analysts to use OpenSearch Dashboards (Kibana) to submit queries to the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster.

C. Store the metadata and the Amazon S3 location of the image file in an Amazon Redshift table. Allow the data analysts to run ad-hoc queries on the table.

D. Store the metadata and the Amazon S3 location of the image files in an Apache Parquet file in Amazon S3, and define a table in the AWS Glue Data Catalog. Allow data analysts to use Amazon Athena to submit custom queries.

Question #53 *Topic 1*

A mobile gaming company wants to capture data from its gaming app and make the data available for analysis immediately. The data record size will be approximately 20 KB. The company is concerned about achieving optimal throughput from each device. Additionally, the company wants to develop a data stream processing application with dedicated throughput for each consumer.
Which solution would achieve this goal?

A. Have the app call the PutRecords API to send data to Amazon Kinesis Data Streams. Use the enhanced fan-out feature while consuming the data.

B. Have the app call the PutRecordBatch API to send data to Amazon Kinesis Data Firehose. Submit a support case to enable dedicated throughput on the account.

C. Have the app use Amazon Kinesis Producer Library (KPL) to send data to Kinesis Data Firehose. Use the enhanced fan-out feature while consuming the data.

D. Have the app call the PutRecords API to send data to Amazon Kinesis Data Streams. Host the stream-processing application on Amazon EC2 with Auto Scaling.

Question #54 *Topic 1*

A marketing company wants to improve its reporting and business intelligence capabilities. During the planning phase, the company interviewed the relevant stakeholders and discovered that:
⊶ The operations team reports are run hourly for the current month's data.
⊶ The sales team wants to use multiple Amazon QuickSight dashboards to show a rolling view of the last 30 days based on several categories. The sales team also wants to view the data as soon as it reaches the reporting backend.
⊶ The finance team's reports are run daily for last month's data and once a month for the last 24 months of data.
Currently, there is 400 TB of data in the system with an expected additional 100 TB added every month. The company is looking for a solution that is as cost- effective as possible.
Which solution meets the company's requirements?

A. Store the last 24 months of data in Amazon Redshift. Configure Amazon QuickSight with Amazon Redshift as the data source.

B. Store the last 2 months of data in Amazon Redshift and the rest of the months in Amazon S3. Set up an external schema and table for Amazon Redshift Spectrum. Configure Amazon QuickSight with Amazon Redshift as the data source.

C. Store the last 24 months of data in Amazon S3 and query it using Amazon Redshift Spectrum. Configure Amazon QuickSight with Amazon Redshift Spectrum as the data source.

D. Store the last 2 months of data in Amazon Redshift and the rest of the months in Amazon S3. Use a long-running Amazon EMR with Apache Spark cluster to query the data as needed. Configure Amazon QuickSight with Amazon EMR as the data source.

Question #55 *Topic 1*

A media company wants to perform machine learning and analytics on the data residing in its Amazon S3 data lake. There are two data transformation requirements that will enable the consumers within the company to create reports:
⊶ Daily transformations of 300 GB of data with different file formats landing in Amazon S3 at a scheduled time.
⊶ One-time transformations of terabytes of archived data residing in the S3 data lake.
Which combination of solutions cost-effectively meets the company's requirements for transforming the data? (Choose three.)

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 incoming data, use AWS Glue workflows with AWS Glue jobs to perform transformations.

E. For archived data, use Amazon EMR to perform data transformations.

F. For archived data, use Amazon SageMaker to perform data transformations.

Question #56 *Topic 1*

A hospital uses wearable medical sensor devices to collect data from patients. The hospital is architecting a near-real-time solution that can ingest the data securely at scale. The solution should also be able to remove the patient's protected health information (PHI) from the streaming data and store the data in durable storage.
Which solution meets these requirements with the least operational overhead?

A. Ingest the data using Amazon Kinesis Data Streams, which invokes an AWS Lambda function using Kinesis Client Library (KCL) to remove all PHI. Write the data in Amazon S3.

B. Ingest the data using Amazon Kinesis Data Firehose to write the data to Amazon S3. Have Amazon S3 trigger an AWS Lambda function that parses the sensor data to remove all PHI in Amazon S3.

C. Ingest the data using Amazon Kinesis Data Streams to write the data to Amazon S3. Have the data stream launch an AWS Lambda function that parses the sensor data and removes all PHI in Amazon S3.

D. Ingest the data using Amazon Kinesis Data Firehose to write the data to Amazon S3. Implement a transformation AWS Lambda function that parses the sensor data to remove all PHI.

Question #57 *Topic 1*

A company is migrating its existing on-premises ETL jobs to Amazon EMR. The code consists of a series of jobs written in Java. The company needs to reduce overhead for the system administrators without changing the underlying code. Due to the sensitivity of the data, compliance requires that the company use root device volume encryption on all nodes in the cluster. Corporate standards require that environments be provisioned though AWS CloudFormation when possible.
Which solution satisfies these requirements?

A. Install open-source Hadoop on Amazon EC2 instances with encrypted root device volumes. Configure the cluster in the CloudFormation template.

B. Use a CloudFormation template to launch an EMR cluster. In the configuration section of the cluster, define a bootstrap action to enable TLS.

C. Create a custom AMI with encrypted root device volumes. Configure Amazon EMR to use the custom AMI using the CustomAmild property in the CloudFormation template.

D. Use a CloudFormation template to launch an EMR cluster. In the configuration section of the cluster, define a bootstrap action to encrypt the root device volume of every node.

Question #58

Topic 1

A transportation company uses IoT sensors attached to trucks to collect vehicle data for its global delivery fleet. The company currently sends the sensor data in small .csv files to Amazon S3. The files are then loaded into a 10-node Amazon Redshift cluster with two slices per node and queried using both Amazon Athena and Amazon Redshift. The company wants to optimize the files to reduce the cost of querying and also improve the speed of data loading into the Amazon
Redshift cluster.
Which solution meets these requirements?

A. Use AWS Glue to convert all the files from .csv to a single large Apache Parquet file. COPY the file into Amazon Redshift and query the file with Athena from Amazon S3.

B. Use Amazon EMR to convert each .csv file to Apache Avro. COPY the files into Amazon Redshift and query the file with Athena from Amazon S3.

C. Use AWS Glue to convert the files from .csv to a single large Apache ORC file. COPY the file into Amazon Redshift and query the file with Athena from Amazon S3.

D. Use AWS Glue to convert the files from .csv to Apache Parquet to create 20 Parquet files. COPY the files into Amazon Redshift and query the files with Athena from Amazon S3.

Question #59

Topic 1

An online retail company with millions of users around the globe wants to improve its ecommerce analytics capabilities. Currently, clickstream data is uploaded directly to Amazon S3 as compressed files. Several times each day, an application running on Amazon EC2 processes the data and makes search options and reports available for visualization by editors and marketers. The company wants to make website clicks and aggregated data available to editors and marketers in minutes to enable them to connect with users more effectively.
Which options will help meet these requirements in the MOST efficient way? (Choose two.)

A. Use Amazon Kinesis Data Firehose to upload compressed and batched clickstream records to Amazon OpenSearch Service (Amazon Elasticsearch Service).

B. Upload clickstream records to Amazon S3 as compressed files. Then use AWS Lambda to send data to Amazon OpenSearch Service (Amazon Elasticsearch Service) from Amazon S3.

C. Use Amazon OpenSearch Service (Amazon Elasticsearch Service) deployed on Amazon EC2 to aggregate, filter, and process the data. Refresh content performance dashboards in near-real time.

D. Use OpenSearch Dashboards (Kibana) to aggregate, filter, and visualize the data stored in Amazon OpenSearch Service (Amazon Elasticsearch Service). Refresh content performance dashboards in near-real time.

E. Upload clickstream records from Amazon S3 to Amazon Kinesis Data Streams and use a Kinesis Data Streams consumer to send records to Amazon OpenSearch Service (Amazon Elasticsearch Service).

Question #60 *Topic 1*

A company is streaming its high-volume billing data (100 MBps) to Amazon Kinesis Data Streams. A data analyst partitioned the data on account_id to ensure that all records belonging to an account go to the same Kinesis shard and order is maintained. While building a custom consumer using the Kinesis Java SDK, the data analyst notices that, sometimes, the messages arrive out of order for account_id. Upon further investigation, the data analyst discovers the messages that are out of order seem to be arriving from different shards for the same account_id and are seen when a stream resize runs.
What is an explanation for this behavior and what is the solution?

A. There are multiple shards in a stream and order needs to be maintained in the shard. The data analyst needs to make sure there is only a single shard in the stream and no stream resize runs.

B. The hash key generation process for the records is not working correctly. The data analyst should generate an explicit hash key on the producer side so the records are directed to the appropriate shard accurately.

C. The records are not being received by Kinesis Data Streams in order. The producer should use the PutRecords API call instead of the PutRecord API call with the SequenceNumberForOrdering parameter.

D. The consumer is not processing the parent shard completely before processing the child shards after a stream resize. The data analyst should process the parent shard completely first before processing the child shards.

Question #61 *Topic 1*

A media analytics company consumes a stream of social media posts. The posts are sent to an Amazon Kinesis data stream partitioned on user_id. An AWS
Lambda function retrieves the records and validates the content before loading the posts into an Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster. The validation process needs to receive the posts for a given user in the order they were received by the Kinesis data stream.
During peak hours, the social media posts take more than an hour to appear in the Amazon OpenSearch Service (Amazon ES) cluster. A data analytics specialist must implement a solution that reduces this latency with the least possible operational overhead.
Which solution meets these requirements?

A. Migrate the validation process from Lambda to AWS Glue.

B. Migrate the Lambda consumers from standard data stream iterators to an HTTP/2 stream consumer.

C. Increase the number of shards in the Kinesis data stream.

D. Send the posts stream to Amazon Managed Streaming for Apache Kafka instead of the Kinesis data stream.

Question #62 *Topic 1*

A company launched a service that produces millions of messages every day and uses Amazon Kinesis Data Streams as the streaming service.
The company uses the Kinesis SDK to write data to Kinesis Data Streams. A few months after launch, a data analyst found that write performance is significantly reduced. The data analyst investigated the metrics and determined that Kinesis is throttling the write requests. The data analyst wants to address this issue without significant changes to the architecture.
Which actions should the data analyst take to resolve this issue? (Choose two.)

A. Increase the Kinesis Data Streams retention period to reduce throttling.

B. Replace the Kinesis API-based data ingestion mechanism with Kinesis Agent.

C. Increase the number of shards in the stream using the UpdateShardCount API.

D. Choose partition keys in a way that results in a uniform record distribution across shards.

E. Customize the application code to include retry logic to improve performance.

Question #63 *Topic 1*

A smart home automation company must efficiently ingest and process messages from various connected devices and sensors. The majority of these messages are comprised of a large number of small files. These messages are ingested using Amazon Kinesis Data Streams and sent to Amazon S3 using a Kinesis data stream consumer application. The Amazon S3 message data is then passed through a processing pipeline built on Amazon EMR running scheduled PySpark jobs.

The data platform team manages data processing and is concerned about the efficiency and cost of downstream data processing. They want to continue to use

PySpark.

Which solution improves the efficiency of the data processing jobs and is well architected?

A. Send the sensor and devices data directly to a Kinesis Data Firehose delivery stream to send the data to Amazon S3 with Apache Parquet record format conversion enabled. Use Amazon EMR running PySpark to process the data in Amazon S3.

B. Set up an AWS Lambda function with a Python runtime environment. Process individual Kinesis data stream messages from the connected devices and sensors using Lambda.

C. Launch an Amazon Redshift cluster. Copy the collected data from Amazon S3 to Amazon Redshift and move the data processing jobs from Amazon EMR to Amazon Redshift.

D. Set up AWS Glue Python jobs to merge the small data files in Amazon S3 into larger files and transform them to Apache Parquet format. Migrate the downstream PySpark jobs from Amazon EMR to AWS Glue.

Question #64 *Topic 1*

A large financial company is running its ETL process. Part of this process is to move data from Amazon S3 into an Amazon Redshift cluster. The company wants to use the most cost-efficient method to load the dataset into Amazon Redshift.

Which combination of steps would meet these requirements? (Choose two.)

A. Use the COPY command with the manifest file to load data into Amazon Redshift.

B. Use S3DistCp to load files into Amazon Redshift.

C. Use temporary staging tables during the loading process.

D. Use the UNLOAD command to upload data into Amazon Redshift.

E. Use Amazon Redshift Spectrum to query files from Amazon S3.

Question #65 *Topic 1*

A university intends to use Amazon Kinesis Data Firehose to collect JSON-formatted batches of water quality readings in Amazon S3. The readings are from 50 sensors scattered across a local lake. Students will query the stored data using Amazon Athena to observe changes in a captured metric over time, such as water temperature or acidity. Interest has grown in the study, prompting the university to reconsider how data will be stored.

Which data format and partitioning choices will MOST significantly reduce costs? (Choose two.)

A. Store the data in Apache Avro format using Snappy compression.

B. Partition the data by year, month, and day.

C. Store the data in Apache ORC format using no compression.

D. Store the data in Apache Parquet format using Snappy compression.

E. Partition the data by sensor, year, month, and day.

Question #66 *Topic 1*

A healthcare company uses AWS data and analytics tools to collect, ingest, and store electronic health record (EHR) data about its patients. The raw EHR data is stored in Amazon S3 in JSON format partitioned by hour, day, and year and is updated every hour. The company wants to maintain the data catalog and metadata in an AWS Glue Data Catalog to be able to access the data using Amazon Athena or Amazon Redshift Spectrum for analytics.

When defining tables in the Data Catalog, the company has the following requirements:

➾ Choose the catalog table name and do not rely on the catalog table naming algorithm.

➾ Keep the table updated with new partitions loaded in the respective S3 bucket prefixes.

Which solution meets these requirements with minimal effort?

A. Run an AWS Glue crawler that connects to one or more data stores, determines the data structures, and writes tables in the Data Catalog.

B. Use the AWS Glue console to manually create a table in the Data Catalog and schedule an AWS Lambda function to update the table partitions hourly.

C. Use the AWS Glue API CreateTable operation to create a table in the Data Catalog. Create an AWS Glue crawler and specify the table as the source.

D. Create an Apache Hive catalog in Amazon EMR with the table schema definition in Amazon S3, and update the table partition with a scheduled job. Migrate the Hive catalog to the Data Catalog.

Question #67 *Topic 1*

A large university has adopted a strategic goal of increasing diversity among enrolled students. The data analytics team is creating a dashboard with data visualizations to enable stakeholders to view historical trends. All access must be authenticated using Microsoft Active Directory. All data in transit and at rest must be encrypted.

Which solution meets these requirements?

A. Amazon QuickSight Standard edition configured to perform identity federation using SAML 2.0. and the default encryption settings.

B. Amazon QuickSight Enterprise edition configured to perform identity federation using SAML 2.0 and the default encryption settings.

C. Amazon QuckSight Standard edition using AD Connector to authenticate using Active Directory. Configure Amazon QuickSight to use customer-provided keys imported into AWS KMS.

D. Amazon QuickSight Enterprise edition using AD Connector to authenticate using Active Directory. Configure Amazon QuickSight to use customer-provided keys imported into AWS KMS.

Question #68 *Topic 1*

An airline has been collecting metrics on flight activities for analytics. A recently completed proof of concept demonstrates how the company provides insights to data analysts to improve on-time departures. The proof of concept used objects in Amazon S3, which contained the metrics in .csv format, and used Amazon
Athena for querying the data. As the amount of data increases, the data analyst wants to optimize the storage solution to improve query performance.
Which options should the data analyst use to improve performance as the data lake grows? (Choose three.)

A. Add a randomized string to the beginning of the keys in S3 to get more throughput across partitions.

B. Use an S3 bucket in the same account as Athena.

C. Compress the objects to reduce the data transfer I/O.

D. Use an S3 bucket in the same Region as Athena.

E. Preprocess the .csv data to JSON to reduce I/O by fetching only the document keys needed by the query.

F. Preprocess the .csv data to Apache Parquet to reduce I/O by fetching only the data blocks needed for predicates.

Question #69 *Topic 1*

A company uses the Amazon Kinesis SDK to write data to Kinesis Data Streams. Compliance requirements state that the data must be encrypted at rest using a key that can be rotated. The company wants to meet this encryption requirement with minimal coding effort.
How can these requirements be met?

A. Create a customer master key (CMK) in AWS KMS. Assign the CMK an alias. Use the AWS Encryption SDK, providing it with the key alias to encrypt and decrypt the data.

B. Create a customer master key (CMK) in AWS KMS. Assign the CMK an alias. Enable server-side encryption on the Kinesis data stream using the CMK alias as the KMS master key.

C. Create a customer master key (CMK) in AWS KMS. Create an AWS Lambda function to encrypt and decrypt the data. Set the KMS key ID in the function's environment variables.

D. Enable server-side encryption on the Kinesis data stream using the default KMS key for Kinesis Data Streams.

Question #70 *Topic 1*

A company wants to enrich application logs in near-real-time and use the enriched dataset for further analysis. The application is running on Amazon EC2 instances across multiple Availability Zones and storing its logs using Amazon CloudWatch Logs. The enrichment source is stored in an Amazon DynamoDB table.
Which solution meets the requirements for the event collection and enrichment?

A. Use a CloudWatch Logs subscription to send the data to Amazon Kinesis Data Firehose. Use AWS Lambda to transform the data in the Kinesis Data Firehose delivery stream and enrich it with the data in the DynamoDB table. Configure Amazon S3 as the Kinesis Data Firehose delivery destination.

B. Export the raw logs to Amazon S3 on an hourly basis using the AWS CLI. Use AWS Glue crawlers to catalog the logs. Set up an AWS Glue connection for the DynamoDB table and set up an AWS Glue ETL job to enrich the data. Store the enriched data in Amazon S3.

C. Configure the application to write the logs locally and use Amazon Kinesis Agent to send the data to Amazon Kinesis Data Streams. Configure a Kinesis Data Analytics SQL application with the Kinesis data stream as the source. Join the SQL application input stream with DynamoDB records, and then store the enriched output stream in Amazon S3 using Amazon Kinesis Data Firehose.

D. Export the raw logs to Amazon S3 on an hourly basis using the AWS CLI. Use Apache Spark SQL on Amazon EMR to read the logs from Amazon S3 and enrich the records with the data from DynamoDB. Store the enriched data in Amazon S3.

Question #71 *Topic 1*

A company uses Amazon Redshift as its data warehouse. A new table has columns that contain sensitive data. The data in the table will eventually be referenced by several existing queries that run many times a day.

A data analyst needs to load 100 billion rows of data into the new table. Before doing so, the data analyst must ensure that only members of the auditing group can read the columns containing sensitive data.

How can the data analyst meet these requirements with the lowest maintenance overhead?

A. Load all the data into the new table and grant the auditing group permission to read from the table. Load all the data except for the columns containing sensitive data into a second table. Grant the appropriate users read-only permissions to the second table.

B. Load all the data into the new table and grant the auditing group permission to read from the table. Use the GRANT SQL command to allow read-only access to a subset of columns to the appropriate users.

C. Load all the data into the new table and grant all users read-only permissions to non-sensitive columns. Attach an IAM policy to the auditing group with explicit ALLOW access to the sensitive data columns.

D. Load all the data into the new table and grant the auditing group permission to read from the table. Create a view of the new table that contains all the columns, except for those considered sensitive, and grant the appropriate users read-only permissions to the table.

Question #72 *Topic 1*

A banking company wants to collect large volumes of transactional data using Amazon Kinesis Data Streams for real-time analytics. The company uses

PutRecord to send data to Amazon Kinesis, and has observed network outages during certain times of the day. The company wants to obtain exactly once semantics for the entire processing pipeline.

What should the company do to obtain these characteristics?

A. Design the application so it can remove duplicates during processing be embedding a unique ID in each record.

B. Rely on the processing semantics of Amazon Kinesis Data Analytics to avoid duplicate processing of events.

C. Design the data producer so events are not ingested into Kinesis Data Streams multiple times.

D. Rely on the exactly one processing semantics of Apache Flink and Apache Spark Streaming included in Amazon EMR.

Question #73 *Topic 1*

A company's data analyst needs to ensure that queries run in Amazon Athena cannot scan more than a prescribed amount of data for cost control purposes.

Queries that exceed the prescribed threshold must be canceled immediately.

What should the data analyst do to achieve this?

A. Configure Athena to invoke an AWS Lambda function that terminates queries when the prescribed threshold is crossed.

B. For each workgroup, set the control limit for each query to the prescribed threshold.

C. Enforce the prescribed threshold on all Amazon S3 bucket policies

D. For each workgroup, set the workgroup-wide data usage control limit to the prescribed threshold.

Question #74 *Topic 1*

A marketing company is using Amazon EMR clusters for its workloads. The company manually installs third-party libraries on the clusters by logging in to the master nodes. A data analyst needs to create an automated solution to replace the manual process.
Which options can fulfill these requirements? (Choose two.)

A. Place the required installation scripts in Amazon S3 and execute them using custom bootstrap actions.

B. Place the required installation scripts in Amazon S3 and execute them through Apache Spark in Amazon EMR.

C. Install the required third-party libraries in the existing EMR master node. Create an AMI out of that master node and use that custom AMI to re-create the EMR cluster.

D. Use an Amazon DynamoDB table to store the list of required applications. Trigger an AWS Lambda function with DynamoDB Streams to install the software.

E. Launch an Amazon EC2 instance with Amazon Linux and install the required third-party libraries on the instance. Create an AMI and use that AMI to create the EMR cluster.

Question #75 *Topic 1*

A data engineering team within a shared workspace company wants to build a centralized logging system for all weblogs generated by the space reservation system. The company has a fleet of Amazon EC2 instances that process requests for shared space reservations on its website. The data engineering team wants to ingest all weblogs into a service that will provide a near-real-time search engine. The team does not want to manage the maintenance and operation of the logging system.
Which solution allows the data engineering team to efficiently set up the web logging system within AWS?

A. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis data stream to CloudWatch. Choose Amazon OpenSearch Service (Amazon Elasticsearch Service) as the end destination of the weblogs.

B. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis Data Firehose delivery stream to CloudWatch. Choose Amazon OpenSearch Service (Amazon Elasticsearch Service) as the end destination of the weblogs.

C. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis data stream to CloudWatch. Configure Splunk as the end destination of the weblogs.

D. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis Firehose delivery stream to CloudWatch. Configure Amazon DynamoDB as the end destination of the weblogs.

Question #76 *Topic 1*

A company wants to research user turnover by analyzing the past 3 months of user activities. With millions of users, 1.5 TB of uncompressed data is generated each day. A 30-node Amazon Redshift cluster with 2.56 TB of solid state drive (SSD) storage for each node is required to meet the query performance goals.

The company wants to run an additional analysis on a year's worth of historical data to examine trends indicating which features are most popular. This analysis will be done once a week.

What is the MOST cost-effective solution?

A. Increase the size of the Amazon Redshift cluster to 120 nodes so it has enough storage capacity to hold 1 year of data. Then use Amazon Redshift for the additional analysis.

B. Keep the data from the last 90 days in Amazon Redshift. Move data older than 90 days to Amazon S3 and store it in Apache Parquet format partitioned by date. Then use Amazon Redshift Spectrum for the additional analysis.

C. Keep the data from the last 90 days in Amazon Redshift. Move data older than 90 days to Amazon S3 and store it in Apache Parquet format partitioned by date. Then provision a persistent Amazon EMR cluster and use Apache Presto for the additional analysis.

D. Resize the cluster node type to the dense storage node type (DS2) for an additional 16 TB storage capacity on each individual node in the Amazon Redshift cluster. Then use Amazon Redshift for the additional analysis.

Question #77 *Topic 1*

A bank operates in a regulated environment. The compliance requirements for the country in which the bank operates say that customer data for each state should only be accessible by the bank's employees located in the same state. Bank employees in one state should NOT be able to access data for customers who have provided a home address in a different state.

The bank's marketing team has hired a data analyst to gather insights from customer data for a new campaign being launched in certain states. Currently, data linking each customer account to its home state is stored in a tabular .csv file within a single Amazon S3 folder in a private S3 bucket. The total size of the S3 folder is 2 GB uncompressed. Due to the country's compliance requirements, the marketing team is not able to access this folder.

The data analyst is responsible for ensuring that the marketing team gets one-time access to customer data for their campaign analytics project, while being subject to all the compliance requirements and controls.

Which solution should the data analyst implement to meet the desired requirements with the LEAST amount of setup effort?

A. Re-arrange data in Amazon S3 to store customer data about each state in a different S3 folder within the same bucket. Set up S3 bucket policies to provide marketing employees with appropriate data access under compliance controls. Delete the bucket policies after the project.

B. Load tabular data from Amazon S3 to an Amazon EMR cluster using s3DistCp. Implement a custom Hadoop-based row-level security solution on the Hadoop Distributed File System (HDFS) to provide marketing employees with appropriate data access under compliance controls. Terminate the EMR cluster after the project.

C. Load tabular data from Amazon S3 to Amazon Redshift with the COPY command. Use the built-in row-level security feature in Amazon Redshift to provide marketing employees with appropriate data access under compliance controls. Delete the Amazon Redshift tables after the project.

D. Load tabular data from Amazon S3 to Amazon QuickSight Enterprise edition by directly importing it as a data source. Use the built-in row-level security feature in Amazon QuickSight to provide marketing employees with appropriate data access under compliance controls. Delete Amazon QuickSight data sources after the project is complete.

Question #78 *Topic 1*

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.

Question #79 *Topic 1*

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.

Question #80 *Topic 1*

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.

Question #81 *Topic 1*

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.

Question #82 *Topic 1*

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.

Question #83 *Topic 1*

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

Question #84

Topic 1

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.

Question #85

Topic 1

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.

Question #86

Topic 1

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.

Question #87

Topic 1

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.

Question #88 *Topic 1*

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.

Question #89 *Topic 1*

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.

Question #90 *Topic 1*

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.

Question #91 *Topic 1*

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.

Question #92

Topic 1

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

Question #93

Topic 1

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 the 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.

Question #94

Topic 1

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.

Question #95 *Topic 1*

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.

Question #96 *Topic 1*

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.

Question #97 *Topic 1*

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.

Question #98 *Topic 1*

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

Question #99 *Topic 1*

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.

Question #100 *Topic 1*

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.

Question #101 *Topic 1*

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.

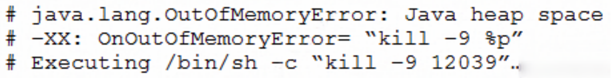
B. Use Amazon Kinesis Data Firehose. Configure a Firehose delivery stream with a preprocessing AWS Lambda function for data cleansing. Use a Kinesis Agent to write data to the delivery stream. Configure Kinesis Data Firehose to deliver the data to Amazon S3.

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.

Question #102 Topic 1

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 JSON 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`"95% 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.

Question #103 Topic 1

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

Question #104 Topic 1

A retail company has 15 stores across 6 cities in the United States. Once a month, the sales team requests a visualization in Amazon QuickSight that provides the ability to easily identify revenue trends across cities and stores. The visualization also helps identify outliers that need to be examined with further analysis.
Which visual type in QuickSight meets the sales team's requirements?

A. Geospatial chart

B. Line chart

C. Heat map

D. Tree map

Question #105

Topic 1

A marketing company has data in Salesforce, MySQL, and Amazon S3. The company wants to use data from these three locations and create mobile dashboards for its users. The company is unsure how it should create the dashboards and needs a solution with the least possible customization and coding.
Which solution meets these requirements?

A. Use Amazon Athena federated queries to join the data sources. Use Amazon QuickSight to generate the mobile dashboards.

B. Use AWS Lake Formation to migrate the data sources into Amazon S3. Use Amazon QuickSight to generate the mobile dashboards.

C. Use Amazon Redshift federated queries to join the data sources. Use Amazon QuickSight to generate the mobile dashboards.

D. Use Amazon QuickSight to connect to the data sources and generate the mobile dashboards.

Question #106

Topic 1

A company uses Amazon Redshift for its data warehousing needs. ETL jobs run every night to load data, apply business rules, and create aggregate tables for reporting. The company's data analysis, data science, and business intelligence teams use the data warehouse during regular business hours. The workload management is set to auto, and separate queues exist for each team with the priority set to NORMAL.
Recently, a sudden spike of read queries from the data analysis team has occurred at least twice daily, and queries wait in line for cluster resources. The company needs a solution that enables the data analysis team to avoid query queuing without impacting latency and the query times of other teams.
Which solution meets these requirements?

A. Increase the query priority to HIGHEST for the data analysis queue.

B. Configure the data analysis queue to enable concurrency scaling.

C. Create a query monitoring rule to add more cluster capacity for the data analysis queue when queries are waiting for resources.

D. Use workload management query queue hopping to route the query to the next matching queue.

Question #107

Topic 1

A company owns facilities with IoT devices installed across the world. The company is using Amazon Kinesis Data Streams to stream data from the devices to
Amazon S3. The company's operations team wants to get insights from the IoT data to monitor data quality at ingestion. The insights need to be derived in near- real time, and the output must be logged to Amazon DynamoDB for further analysis.
Which solution meets these requirements?

A. Connect Amazon Kinesis Data Analytics to analyze the stream data. Save the output to DynamoDB by using the default output from Kinesis Data Analytics.

B. Connect Amazon Kinesis Data Analytics to analyze the stream data. Save the output to DynamoDB by using an AWS Lambda function.

C. Connect Amazon Kinesis Data Firehose to analyze the stream data by using an AWS Lambda function. Save the output to DynamoDB by using the default output from Kinesis Data Firehose.

D. Connect Amazon Kinesis Data Firehose to analyze the stream data by using an AWS Lambda function. Save the data to Amazon S3. Then run an AWS Glue job on schedule to ingest the data into DynamoDB.

Question #108 *Topic 1*

A company has a data lake on AWS that ingests sources of data from multiple business units and uses Amazon Athena for queries. The storage layer is Amazon
S3 using the AWS Glue Data Catalog. The company wants to make the data available to its data scientists and business analysts. However, the company first needs to manage data access for Athena based on user roles and responsibilities.
What should the company do to apply these access controls with the LEAST operational overhead?

A. Define security policy-based rules for the users and applications by role in AWS Lake Formation.

B. Define security policy-based rules for the users and applications by role in AWS Identity and Access Management (IAM).

C. Define security policy-based rules for the tables and columns by role in AWS Glue.

D. Define security policy-based rules for the tables and columns by role in AWS Identity and Access Management (IAM).

Question #109 *Topic 1*

A company has an encrypted Amazon Redshift cluster. The company recently enabled Amazon Redshift audit logs and needs to ensure that the audit logs are also encrypted at rest. The logs are retained for 1 year. The auditor queries the logs once a month.
What is the MOST cost-effective way to meet these requirements?

A. Encrypt the Amazon S3 bucket where the logs are stored by using AWS Key Management Service (AWS KMS). Copy the data into the Amazon Redshift cluster from Amazon S3 on a daily basis. Query the data as required.

B. Disable encryption on the Amazon Redshift cluster, configure audit logging, and encrypt the Amazon Redshift cluster. Use Amazon Redshift Spectrum to query the data as required.

C. Enable default encryption on the Amazon S3 bucket where the logs are stored by using AES-256 encryption. Copy the data into the Amazon Redshift cluster from Amazon S3 on a daily basis. Query the data as required.

D. Enable default encryption on the Amazon S3 bucket where the logs are stored by using AES-256 encryption. Use Amazon Redshift Spectrum to query the data as required.

Question #110 *Topic 1*

A data analytics specialist is setting up workload management in manual mode for an Amazon Redshift environment. The data analytics specialist is defining query monitoring rules to manage system performance and user experience of an Amazon Redshift cluster.
Which elements must each query monitoring rule include?

A. A unique rule name, a query runtime condition, and an AWS Lambda function to resubmit any failed queries in off hours

B. A queue name, a unique rule name, and a predicate-based stop condition

C. A unique rule name, one to three predicates, and an action

D. A workload name, a unique rule name, and a query runtime-based condition

Question #111 *Topic 1*

A market data company aggregates external data sources to create a detailed view of product consumption in different countries. The company wants to sell this data to external parties through a subscription. To achieve this goal, the company needs to make its data securely available to external parties who are also AWS users.
What should the company do to meet these requirements with the LEAST operational overhead?

A. Store the data in Amazon S3. Share the data by using presigned URLs for security.

B. Store the data in Amazon S3. Share the data by using S3 bucket ACLs.

C. Upload the data to AWS Data Exchange for storage. Share the data by using presigned URLs for security.

D. Upload the data to AWS Data Exchange for storage. Share the data by using the AWS Data Exchange sharing wizard.

Question #112 *Topic 1*

A company has a marketing department and a finance department. The departments are storing data in Amazon S3 in their own AWS accounts in AWS
Organizations. Both departments use AWS Lake Formation to catalog and secure their data. The departments have some databases and tables that share common names.
The marketing department needs to securely access some tables from the finance department.
Which two steps are required for this process? (Choose two.)

A. The finance department grants Lake Formation permissions for the tables to the external account for the marketing department.

B. The finance department creates cross-account IAM permissions to the table for the marketing department role.

C. The marketing department creates an IAM role that has permissions to the Lake Formation tables.

Question #113 *Topic 1*

A human resources company maintains a 10-node Amazon Redshift cluster to run analytics queries on the company's data. The Amazon Redshift cluster contains a product table and a transactions table, and both tables have a product_sku column. The tables are over 100 GB in size. The majority of queries run on both tables.
Which distribution style should the company use for the two tables to achieve optimal query performance?

A. An EVEN distribution style for both tables

B. A KEY distribution style for both tables

C. An ALL distribution style for the product table and an EVEN distribution style for the transactions table

D. An EVEN distribution style for the product table and an KEY distribution style for the transactions table

Question #114 *Topic 1*

A company receives data from its vendor in JSON format with a timestamp in the file name. The vendor uploads the data to an Amazon S3 bucket, and the data is registered into the company's data lake for analysis and reporting. The company has configured an S3 Lifecycle policy to archive all files to S3 Glacier after 5 days.

The company wants to ensure that its AWS Glue crawler catalogs data only from S3 Standard storage and ignores the archived files. A data analytics specialist must implement a solution to achieve this goal without changing the current S3 bucket configuration.

Which solution meets these requirements?

A. Use the exclude patterns feature of AWS Glue to identify the S3 Glacier files for the crawler to exclude.

B. Schedule an automation job that uses AWS Lambda to move files from the original S3 bucket to a new S3 bucket for S3 Glacier storage.

C. Use the excludeStorageClasses property in the AWS Glue Data Catalog table to exclude files on S3 Glacier storage.

D. Use the include patterns feature of AWS Glue to identify the S3 Standard files for the crawler to include.

Question #115 *Topic 1*

A company analyzes historical data and needs to query data that is stored in Amazon S3. New data is generated daily as .csv files that are stored in Amazon S3.

The company's analysts are using Amazon Athena to perform SQL queries against a recent subset of the overall data. The amount of data that is ingested into

Amazon S3 has increased substantially over time, and the query latency also has increased.

Which solutions could the company implement to improve query performance? (Choose two.)

A. Use MySQL Workbench on an Amazon EC2 instance, and connect to Athena by using a JDBC or ODBC connector. Run the query from MySQL Workbench instead of Athena directly.

B. Use Athena to extract the data and store it in Apache Parquet format on a daily basis. Query the extracted data.

C. Run a daily AWS Glue ETL job to convert the data files to Apache Parquet and to partition the converted files. Create a periodic AWS Glue crawler to automatically crawl the partitioned data on a daily basis.

D. Run a daily AWS Glue ETL job to compress the data files by using the .gzip format. Query the compressed data.

E. Run a daily AWS Glue ETL job to compress the data files by using the .lzo format. Query the compressed data.

Question #116 *Topic 1*

A company is sending historical datasets to Amazon S3 for storage. A data engineer at the company wants to make these datasets available for analysis using
Amazon Athena. The engineer also wants to encrypt the Athena query results in an S3 results location by using AWS solutions for encryption. The requirements for encrypting the query results are as follows:
⮚ Use custom keys for encryption of the primary dataset query results.
⮚ Use generic encryption for all other query results.
⮚ Provide an audit trail for the primary dataset queries that shows when the keys were used and by whom.
Which solution meets these requirements?

A. Use server-side encryption with S3 managed encryption keys (SSE-S3) for the primary dataset. Use SSE-S3 for the other datasets.

B. Use server-side encryption with customer-provided encryption keys (SSE-C) for the primary dataset. Use server-side encryption with S3 managed encryption keys (SSE-S3) for the other datasets.

C. Use server-side encryption with AWS KMS managed customer master keys (SSE-KMS CMKs) for the primary dataset. Use server-side encryption with S3 managed encryption keys (SSE-S3) for the other datasets.

D. Use client-side encryption with AWS Key Management Service (AWS KMS) customer managed keys for the primary dataset. Use S3 client-side encryption with client-side keys for the other datasets.

Question #117 *Topic 1*

A large telecommunications company is planning to set up a data catalog and metadata management for multiple data sources running on AWS. The catalog will be used to maintain the metadata of all the objects stored in the data stores. The data stores are composed of structured sources like Amazon RDS and Amazon
Redshift, and semistructured sources like JSON and XML files stored in Amazon S3. The catalog must be updated on a regular basis, be able to detect the changes to object metadata, and require the least possible administration.
Which solution meets these requirements?

A. Use Amazon Aurora as the data catalog. Create AWS Lambda functions that will connect and gather the metadata information from multiple sources and update the data catalog in Aurora. Schedule the Lambda functions periodically.

B. Use the AWS Glue Data Catalog as the central metadata repository. Use AWS Glue crawlers to connect to multiple data stores and update the Data Catalog with metadata changes. Schedule the crawlers periodically to update the metadata catalog.

C. Use Amazon DynamoDB as the data catalog. Create AWS Lambda functions that will connect and gather the metadata information from multiple sources and update the DynamoDB catalog. Schedule the Lambda functions periodically.

D. Use the AWS Glue Data Catalog as the central metadata repository. Extract the schema for RDS and Amazon Redshift sources and build the Data Catalog. Use AWS crawlers for data stored in Amazon S3 to infer the schema and automatically update the Data Catalog.

Question #118 *Topic 1*

An ecommerce company is migrating its business intelligence environment from on premises to the AWS Cloud. The company will use Amazon Redshift in a public subnet and Amazon QuickSight. The tables already are loaded into Amazon Redshift and can be accessed by a SQL tool.
The company starts QuickSight for the first time. During the creation of the data source, a data analytics specialist enters all the information and tries to validate the connection. An error with the following message occurs: `Creating a connection to your data source timed out.`
How should the data analytics specialist resolve this error?

A. Grant the SELECT permission on Amazon Redshift tables.

B. Add the QuickSight IP address range into the Amazon Redshift security group.

C. Create an IAM role for QuickSight to access Amazon Redshift.

D. Use a QuickSight admin user for creating the dataset.

Question #119 *Topic 1*

A power utility company is deploying thousands of smart meters to obtain real-time updates about power consumption. The company is using Amazon Kinesis
Data Streams to collect the data streams from smart meters. The consumer application uses the Kinesis Client Library (KCL) to retrieve the stream data. The company has only one consumer application.
The company observes an average of 1 second of latency from the moment that a record is written to the stream until the record is read by a consumer application. The company must reduce this latency to 500 milliseconds.
Which solution meets these requirements?

A. Use enhanced fan-out in Kinesis Data Streams.

B. Increase the number of shards for the Kinesis data stream.

C. Reduce the propagation delay by overriding the KCL default settings.

D. Develop consumers by using Amazon Kinesis Data Firehose.

Question #120 *Topic 1*

A company needs to collect streaming data from several sources and store the data in the AWS Cloud. The dataset is heavily structured, but analysts need to perform several complex SQL queries and need consistent performance. Some of the data is queried more frequently than the rest. The company wants a solution that meets its performance requirements in a cost-effective manner.
Which solution meets these requirements?

A. Use Amazon Managed Streaming for Apache Kafka to ingest the data to save it to Amazon S3. Use Amazon Athena to perform SQL queries over the ingested data.

B. Use Amazon Managed Streaming for Apache Kafka to ingest the data to save it to Amazon Redshift. Enable Amazon Redshift workload management (WLM) to prioritize workloads.

C. Use Amazon Kinesis Data Firehose to ingest the data to save it to Amazon Redshift. Enable Amazon Redshift workload management (WLM) to prioritize workloads.

D. Use Amazon Kinesis Data Firehose to ingest the data to save it to Amazon S3. Load frequently queried data to Amazon Redshift using the COPY command. Use Amazon Redshift Spectrum for less frequently queried data.

Question #121 *Topic 1*

A manufacturing company uses Amazon Connect to manage its contact center and Salesforce to manage its customer relationship management (CRM) data. The data engineering team must build a pipeline to ingest data from the contact center and CRM system into a data lake that is built on Amazon S3.
What is the MOST efficient way to collect data in the data lake with the LEAST operational overhead?

A. Use Amazon Kinesis Data Streams to ingest Amazon Connect data and Amazon AppFlow to ingest Salesforce data.

B. Use Amazon Kinesis Data Firehose to ingest Amazon Connect data and Amazon Kinesis Data Streams to ingest Salesforce data.

C. Use Amazon Kinesis Data Firehose to ingest Amazon Connect data and Amazon AppFlow to ingest Salesforce data.

D. Use Amazon AppFlow to ingest Amazon Connect data and Amazon Kinesis Data Firehose to ingest Salesforce data.

Question #122 *Topic 1*

A manufacturing company wants to create an operational analytics dashboard to visualize metrics from equipment in near-real time. The company uses Amazon
Kinesis Data Streams to stream the data to other applications. The dashboard must automatically refresh every 5 seconds. A data analytics specialist must design a solution that requires the least possible implementation effort.
Which solution meets these requirements?

A. Use Amazon Kinesis Data Firehose to store the data in Amazon S3. Use Amazon QuickSight to build the dashboard.

B. Use Apache Spark Streaming on Amazon EMR to read the data in near-real time. Develop a custom application for the dashboard by using D3.js.

C. Use Amazon Kinesis Data Firehose to push the data into an Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster. Visualize the data by using an OpenSearch Dashboards (Kibana).

D. Use AWS Glue streaming ETL to store the data in Amazon S3. Use Amazon QuickSight to build the dashboard.

Question #123 *Topic 1*

A data analyst is designing an Amazon QuickSight dashboard using centralized sales data that resides in Amazon Redshift. The dashboard must be restricted so that a salesperson in Sydney, Australia, can see only the Australia view and that a salesperson in New York can see only United States (US) data.
What should the data analyst do to ensure the appropriate data security is in place?

A. Place the data sources for Australia and the US into separate SPICE capacity pools.

B. Set up an Amazon Redshift VPC security group for Australia and the US.

C. Deploy QuickSight Enterprise edition to implement row-level security (RLS) to the sales table.

D. Deploy QuickSight Enterprise edition and set up different VPC security groups for Australia and the US.

Question #124 *Topic 1*

A company wants to run analytics on its Elastic Load Balancing logs stored in Amazon S3. A data analyst needs to be able to query all data from a desired year, month, or day. The data analyst should also be able to query a subset of the columns. The company requires minimal operational overhead and the most cost- effective solution.
Which approach meets these requirements for optimizing and querying the log data?

A. Use an AWS Glue job nightly to transform new log files into .csv format and partition by year, month, and day. Use AWS Glue crawlers to detect new partitions. Use Amazon Athena to query data.

B. Launch a long-running Amazon EMR cluster that continuously transforms new log files from Amazon S3 into its Hadoop Distributed File System (HDFS) storage and partitions by year, month, and day. Use Apache Presto to query the optimized format.

C. Launch a transient Amazon EMR cluster nightly to transform new log files into Apache ORC format and partition by year, month, and day. Use Amazon Redshift Spectrum to query the data.

D. Use an AWS Glue job nightly to transform new log files into Apache Parquet format and partition by year, month, and day. Use AWS Glue crawlers to detect new partitions. Use Amazon Athena to query data.

Question #125 *Topic 1*

An education provider's learning management system (LMS) is hosted in a 100 TB data lake that is built on Amazon S3. The provider's LMS supports hundreds of schools. The provider wants to build an advanced analytics reporting platform using Amazon Redshift to handle complex queries with optimal performance.
System users will query the most recent 4 months of data 95% of the time while 5% of the queries will leverage data from the previous 12 months.
Which solution meets these requirements in the MOST cost-effective way?

A. Store the most recent 4 months of data in the Amazon Redshift cluster. Use Amazon Redshift Spectrum to query data in the data lake. Use S3 lifecycle management rules to store data from the previous 12 months in Amazon S3 Glacier storage.

B. Leverage DS2 nodes for the Amazon Redshift cluster. Migrate all data from Amazon S3 to Amazon Redshift. Decommission the data lake.

C. Store the most recent 4 months of data in the Amazon Redshift cluster. Use Amazon Redshift Spectrum to query data in the data lake. Ensure the S3 Standard storage class is in use with objects in the data lake.

D. Store the most recent 4 months of data in the Amazon Redshift cluster. Use Amazon Redshift federated queries to join cluster data with the data lake to reduce costs. Ensure the S3 Standard storage class is in use with objects in the data lake.

Question #126 *Topic 1*

A company is hosting an enterprise reporting solution with Amazon Redshift. The application provides reporting capabilities to three main groups: an executive group to access financial reports, a data analyst group to run long-running ad-hoc queries, and a data engineering group to run stored procedures and ETL processes. The executive team requires queries to run with optimal performance. The data engineering team expects queries to take minutes.
Which Amazon Redshift feature meets the requirements for this task?

A. Concurrency scaling

B. Short query acceleration (SQA)

C. Workload management (WLM)

D. Materialized views

Question #127 *Topic 1*

A global pharmaceutical company receives test results for new drugs from various testing facilities worldwide. The results are sent in millions of 1 KB-sized JSON objects to an Amazon S3 bucket owned by the company. The data engineering team needs to process those files, convert them into Apache Parquet format, and load them into Amazon Redshift for data analysts to perform dashboard reporting. The engineering team uses AWS Glue to process the objects, AWS Step
Functions for process orchestration, and Amazon CloudWatch for job scheduling.
More testing facilities were recently added, and the time to process files is increasing.
What will MOST efficiently decrease the data processing time?

A. Use AWS Lambda to group the small files into larger files. Write the files back to Amazon S3. Process the files using AWS Glue and load them into Amazon Redshift tables.

B. Use the AWS Glue dynamic frame file grouping option while ingesting the raw input files. Process the files and load them into Amazon Redshift tables.

C. Use the Amazon Redshift COPY command to move the files from Amazon S3 into Amazon Redshift tables directly. Process the files in Amazon Redshift.

D. Use Amazon EMR instead of AWS Glue to group the small input files. Process the files in Amazon EMR and load them into Amazon Redshift tables.

Question #128 *Topic 1*

A company operates toll services for highways across the country and collects data that is used to understand usage patterns. Analysts have requested the ability to run traffic reports in near-real time. The company is interested in building an ingestion pipeline that loads all the data into an Amazon Redshift cluster and alerts operations personnel when toll traffic for a particular toll station does not meet a specified threshold.
Station data and the corresponding threshold values are stored in Amazon S3.
Which approach is the MOST efficient way to meet these requirements?

A. Use Amazon Kinesis Data Firehose to collect data and deliver it to Amazon Redshift and Amazon Kinesis Data Analytics simultaneously. Create a reference data source in Kinesis Data Analytics to temporarily store the threshold values from Amazon S3 and compare the count of vehicles for a particular toll station against its corresponding threshold value. Use AWS Lambda to publish an Amazon Simple Notification Service (Amazon SNS) notification if the threshold is not met.

B. Use Amazon Kinesis Data Streams to collect all the data from toll stations. Create a stream in Kinesis Data Streams to temporarily store the threshold values from Amazon S3. Send both streams to Amazon Kinesis Data Analytics to compare the count of vehicles for a particular toll station against its corresponding threshold value. Use AWS Lambda to publish an Amazon Simple Notification Service (Amazon SNS) notification if the threshold is not met. Connect Amazon Kinesis Data Firehose to Kinesis Data Streams to deliver the data to Amazon Redshift.

C. Use Amazon Kinesis Data Firehose to collect data and deliver it to Amazon Redshift. Then, automatically trigger an AWS Lambda function that queries the data in Amazon Redshift, compares the count of vehicles for a particular toll station against its corresponding threshold values read from Amazon S3, and publishes an Amazon Simple Notification Service (Amazon SNS) notification if the threshold is not met.

D. Use Amazon Kinesis Data Firehose to collect data and deliver it to Amazon Redshift and Amazon Kinesis Data Analytics simultaneously. Use Kinesis Data Analytics to compare the count of vehicles against the threshold value for the station stored in a table as an in-application stream based on information stored in Amazon S3. Configure an AWS Lambda function as an output for the application that will publish an Amazon Simple Queue Service (Amazon SQS) notification to alert operations personnel if the threshold is not met.

Question #129 *Topic 1*

An online retail company uses Amazon Redshift to store historical sales transactions. The company is required to encrypt data at rest in the clusters to comply with the Payment Card Industry Data Security Standard (PCI DSS). A corporate governance policy mandates management of encryption keys using an on- premises hardware security module (HSM).
Which solution meets these requirements?

A. Create and manage encryption keys using AWS CloudHSM Classic. Launch an Amazon Redshift cluster in a VPC with the option to use CloudHSM Classic for key management.

B. Create a VPC and establish a VPN connection between the VPC and the on-premises network. Create an HSM connection and client certificate for the on- premises HSM. Launch a cluster in the VPC with the option to use the on-premises HSM to store keys.

C. Create an HSM connection and client certificate for the on-premises HSM. Enable HSM encryption on the existing unencrypted cluster by modifying the cluster. Connect to the VPC where the Amazon Redshift cluster resides from the on-premises network using a VPN.

D. Create a replica of the on-premises HSM in AWS CloudHSM. Launch a cluster in a VPC with the option to use CloudHSM to store keys.

Question #130 *Topic 1*

A hospital is building a research data lake to ingest data from electronic health records (EHR) systems from multiple hospitals and clinics. The EHR systems are independent of each other and do not have a common patient identifier. The data engineering team is not experienced in machine learning (ML) and has been asked to generate a unique patient identifier for the ingested records.
Which solution will accomplish this task?

A. An AWS Glue ETL job with the FindMatches transform

B. Amazon Kendra

C. Amazon SageMaker Ground Truth

D. An AWS Glue ETL job with the ResolveChoice transform

Question #131 *Topic 1*

A company is Running Apache Spark on an Amazon EMR cluster. The Spark job writes to an Amazon S3 bucket. The job fails and returns an HTTP 503 `Slow
Down` AmazonS3Exception error.
Which actions will resolve this error? (Choose two.)

A. Add additional prefixes to the S3 bucket

B. Reduce the number of prefixes in the S3 bucket

C. Increase the EMR File System (EMRFS) retry limit

D. Disable dynamic partition pruning in the Spark configuration for the cluster

E. Add more partitions in the Spark configuration for the cluster

Question #132 *Topic 1*

A company recently created a test AWS account to use for a development environment. The company also created a production AWS account in another AWS
Region. As part of its security testing, the company wants to send log data from Amazon CloudWatch Logs in its production account to an Amazon Kinesis data stream in its test account.
Which solution will allow the company to accomplish this goal?

A. Create a subscription filter in the production account's CloudWatch Logs to target the Kinesis data stream in the test account as its destination. In the test account, create an IAM role that grants access to the Kinesis data stream and the CloudWatch Logs resources in the production account.

B. In the test account, create an IAM role that grants access to the Kinesis data stream and the CloudWatch Logs resources in the production account. Create a destination data stream in Kinesis Data Streams in the test account with an IAM role and a trust policy that allow CloudWatch Logs in the production account to write to the test account.

C. In the test account, create an IAM role that grants access to the Kinesis data stream and the CloudWatch Logs resources in the production account. Create a destination data stream in Kinesis Data Streams in the test account with an IAM role and a trust policy that allow CloudWatch Logs in the production account to write to the test account.

D. Create a destination data stream in Kinesis Data Streams in the test account with an IAM role and a trust policy that allow CloudWatch Logs in the production account to write to the test account. Create a subscription filter in the production account's CloudWatch Logs to target the Kinesis data stream in the test account as its destination.

Question #133 *Topic 1*

A data architect is building an Amazon S3 data lake for a bank. The goal is to provide a single data repository for customer data needs, such as personalized recommendations. The bank uses Amazon Kinesis Data Firehose to ingest customers' personal information bank accounts, and transactions in near-real time from a transactional relational database. The bank requires all personally identifiable information (PII) that is stored in the AWS Cloud to be masked.
Which solution will meet these requirements?

A. Invoke an AWS Lambda function from Kinesis Data Firehose to mask PII before delivering the data into Amazon S3.

B. Use Amazon Made, and configure it to discover and mask PII.

C. Enable server-side encryption (SSE) in Amazon S3.

D. Invoke Amazon Comprehend from Kinesis Data Firehose to detect and mask PII before delivering the data into Amazon S3.

Question #134 *Topic 1*

An analytics software as a service (SaaS) provider wants to offer its customers business intelligence (BI) reporting capabilities that are self-service. The provider is using Amazon QuickSight to build these reports. The data for the reports resides in a multi-tenant database, but each customer should only be able to access their own data.
The provider wants to give customers two user role options:
⊸ Read-only users for individuals who only need to view dashboards.
⊸ Power users for individuals who are allowed to create and share new dashboards with other users.
Which QuickSught feature allows the provider to meet these requirements?

A. Embedded dashboards

B. Table calculations

C. Isolated namespaces

D. SPICE

Question #135 *Topic 1*

A company is providing analytics services to its sales and marketing departments. The departments can access the data only through their business intelligence
(BI) tools, which run queries on Amazon Redshift using an Amazon Redshift internal user to connect. Each department is assigned a user in the Amazon Redshift database with the permissions needed for that department. The marketing data analysts must be granted direct access to the advertising table, which is stored in
Apache Parquet format in the marketing S3 bucket of the company data lake. The company data lake is managed by AWS Lake Formation. Finally, access must be limited to the three promotion columns in the table.
Which combination of steps will meet these requirements? (Choose three.)

A. Grant permissions in Amazon Redshift to allow the marketing Amazon Redshift user to access the three promotion columns of the advertising external table.

B. Create an Amazon Redshift Spectrum IAM role with permissions for Lake Formation. Attach it to the Amazon Redshift cluster.

C. Create an Amazon Redshift Spectrum IAM role with permissions for the marketing S3 bucket. Attach it to the Amazon Redshift cluster.

D. Create an external schema in Amazon Redshift by using the Amazon Redshift Spectrum IAM role. Grant usage to the marketing Amazon Redshift user.

E. Grant permissions in Lake Formation to allow the Amazon Redshift Spectrum role to access the three promotion columns of the advertising table.

F. Grant permissions in Lake Formation to allow the marketing IAM group to access the three promotion columns of the advertising table.

Question #136 *Topic 1*

A web retail company wants to implement a near-real-time clickstream analytics solution. The company wants to analyze the data with an open-source package.
The analytics application will process the raw data only once, but other applications will need immediate access to the raw data for up to 1 year.
Which solution meets these requirements with the LEAST amount of operational effort?

A. Use Amazon Kinesis Data Streams to collect the data. Use Amazon EMR with Apache Flink to consume and process the data from the Kinesis data stream. Set the retention period of the Kinesis data stream to 8.760 hours.

B. Use Amazon Kinesis Data Streams to collect the data. Use Amazon Kinesis Data Analytics with Apache Flink to process the data in real time. Set the retention period of the Kinesis data stream to 8,760 hours.

C. Use Amazon Managed Streaming for Apache Kafka (Amazon MSK) to collect the data. Use Amazon EMR with Apache Flink to consume and process the data from the Amazon MSK stream. Set the log retention hours to 8,760.

D. Use Amazon Kinesis Data Streams to collect the data. Use Amazon EMR with Apache Flink to consume and process the data from the Kinesis data stream. Create an Amazon Kinesis Data Firehose delivery stream to store the data in Amazon S3. Set an S3 Lifecycle policy to delete the data after 365 days.

Question #137 *Topic 1*

A data analyst runs a large number of data manipulation language (DML) queries by using Amazon Athena with the JDBC driver. Recently, a query failed after it ran for 30 minutes. The query returned the following message: java.sql.SQLException: Query timeout
The data analyst does not immediately need the query results. However, the data analyst needs a long-term solution for this problem.
Which solution will meet these requirements?

A. Split the query into smaller queries to search smaller subsets of data

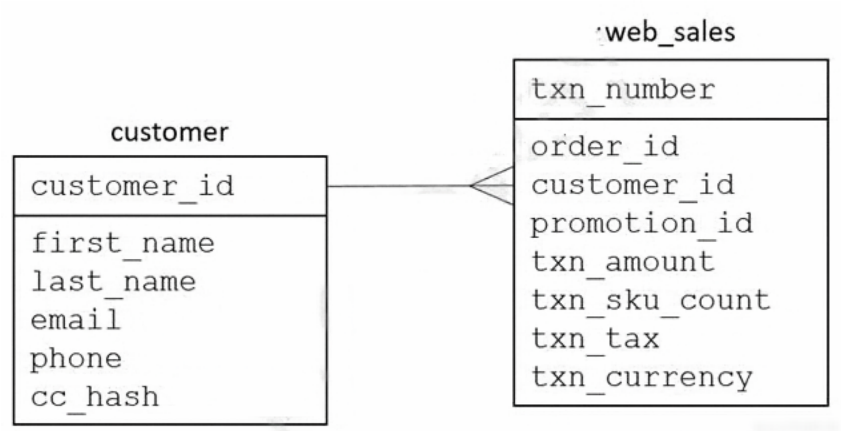
B. In the settings for Athena, adjust the DML query timeout limit

C. In the Service Quotas console, request an increase for the DML query timeout

D. Save the tables as compressed .csv files

Question #138 *Topic 1*

A retail company is using an Amazon S3 bucket to host an ecommerce data lake. The company is using AWS Lake Formation to manage the data lake.
A data analytics specialist must provide access to a new business analyst team. The team will use Amazon Athena from the AWS Management Console to query data from existing web_sales and customer tables in the ecommerce database. The team needs read-only access and the ability to uniquely identify customers by using first and last names. However, the team must not be able to see any other personally identifiable data. The table structure is as follows:



Which combination of steps should the data analytics specialist take to provide the required permission by using the principle of least privilege? (Choose three.)

A. In AWS Lake Formation, grant the business_analyst group SELECT and ALTER permissions for the web_sales table.

B. In AWS Lake Formation, grant the business_analyst group the SELECT permission for the web_sales table.

C. In AWS Lake Formation, grant the business_analyst group the SELECT permission for the customer table. Under columns, choose filter type λ€Include columnsλ€ with columns fisrt_name, last_name, and customer_id.

D. In AWS Lake Formation, grant the business_analyst group SELECT and ALTER permissions for the customer table. Under columns, choose filter type λ€Include columnsλ€ with columns fisrt_name and last_name.

E. Create users under a business_analyst IAM group. Create a policy that allows the lakeformation:GetDataAccess action, the athena:* action, and the glue:Get* action.

F. Create users under a business_analyst IAM group. Create a policy that allows the lakeformation:GetDataAccess action, the athena:* action, and the glue:Get* action. In addition, allow the s3:GetObject action, the s3:PutObject action, and the s3:GetBucketLocation action for the Athena query results S3 bucket.

Question #139 *Topic 1*

A company has multiple data workflows to ingest data from its operational databases into its data lake on Amazon S3. The workflows use AWS Glue and Amazon
EMR for data processing and ETL. The company wants to enhance its architecture to provide automated orchestration and minimize manual intervention.
Which solution should the company use to manage the data workflows to meet these requirements?

A. AWS Glue workflows

B. AWS Step Functions

C. AWS Lambda

D. AWS Batch

Question #140 *Topic 1*

An online retail company is using Amazon Redshift to run queries and perform analytics on customer shopping behavior. When multiple queries are running on the cluster, runtime for small queries increases significantly. The company's data analytics team to decrease the runtime of these small queries by prioritizing them ahead of large queries.
Which solution will meet these requirements?

A. Use Amazon Redshift Spectrum for small queries

B. Increase the concurrency limit in workload management (WLM)

C. Configure short query acceleration in workload management (WLM)

D. Add a dedicated compute node for small queries

Question #141 *Topic 1*

A company uses Amazon Redshift as its data warehouse. A new table includes some columns that contain sensitive data and some columns that contain non- sensitive data. The data in the table eventually will be referenced by several existing queries that run many times each day.
A data analytics specialist must ensure that only members of the company's auditing team can read the columns that contain sensitive data. All other users must have read-only access to the columns that contain non-sensitive data.
Which solution will meet these requirements with the LEAST operational overhead?

A. Grant the auditing team permission to read from the table. Load the columns that contain non-sensitive data into a second table. Grant the appropriate users read-only permissions to the second table.

B. Grant all users read-only permissions to the columns that contain non-sensitive data. Use the GRANT SELECT command to allow the auditing team to access the columns that contain sensitive data.

C. Grant all users read-only permissions to the columns that contain non-sensitive data. Attach an IAM policy to the auditing team with an explicit. Allow action that grants access to the columns that contain sensitive data.

D. Grant the auditing team permission to read from the table. Create a view of the table that includes the columns that contain non-sensitive data. Grant the appropriate users read-only permissions to that view.

Question #142

Topic 1

A company hosts an Apache Flink application on premises. The application processes data from several Apache Kafka clusters. The data originates from a variety of sources, such as web applications, mobile apps, and operational databases. The company has migrated some of these sources to AWS and now wants to migrate the Flink application. The company must ensure that data that resides in databases within the VPC does not traverse the internet. The application must be able to process all the data that comes from the company's AWS solution, on-premises resources, and the public internet.
Which solution will meet these requirements with the LEAST operational overhead?

A. Implement Flink on Amazon EC2 within the company's VPC. Create Amazon Managed Streaming for Apache Kafka (Amazon MSK) clusters in the VPC to collect data that comes from applications and databases within the VPC. Use Amazon Kinesis Data Streams to collect data that comes from the public internet. Configure Flink to have sources from Kinesis Data Streams Amazon MSK, and any on-premises Kafka clusters by using AWS Client VPN or AWS Direct Connect.

B. Implement Flink on Amazon EC2 within the company's VPC. Use Amazon Kinesis Data Streams to collect data that comes from applications and databases within the VPC and the public internet. Configure Flink to have sources from Kinesis Data Streams and any on-premises Kafka clusters by using AWS Client VPN or AWS Direct Connect.

C. Create an Amazon Kinesis Data Analytics application by uploading the compiled Flink .jar file. Use Amazon Kinesis Data Streams to collect data that comes from applications and databases within the VPC and the public internet. Configure the Kinesis Data Analytics application to have sources from Kinesis Data Streams and any on-premises Kafka clusters by using AWS Client VPN or AWS Direct Connect.

D. Create an Amazon Kinesis Data Analytics application by uploading the compiled Flink .jar file. Create Amazon Managed Streaming for Apache Kafka (Amazon MSK) clusters in the company's VPC to collect data that comes from applications and databases within the VPC. Use Amazon Kinesis Data Streams to collect data that comes from the public internet. Configure the Kinesis Data Analytics application to have sources from Kinesis Data Streams, Amazon MSK, and any on-premises Kafka clusters by using AWS Client VPN or AWS Direct Connect.

Question #143

Topic 1

A technology company has an application with millions of active users every day. The company queries daily usage data with Amazon Athena to understand how users interact with the application. The data includes the date and time, the location ID, and the services used. The company wants to use Athena to run queries to analyze the data with the lowest latency possible.
Which solution meets these requirements?

A. Store the data in Apache Avro format with the date and time as the partition, with the data sorted by the location ID.

B. Store the data in Apache Parquet format with the date and time as the partition, with the data sorted by the location ID.

C. Store the data in Apache ORC format with the location ID as the partition, with the data sorted by the date and time.

D. Store the data in .csv format with the location ID as the partition, with the data sorted by the date and time.

Question #144 *Topic 1*

A real estate company maintains data about all properties listed in a market. The company receives data about new property listings from vendors who upload the data daily as compressed files into Amazon S3. The company's leadership team wants to see the most up-to-date listings as soon as the data is uploaded to
Amazon S3. The data analytics team must automate and orchestrate the data processing workflow of the listings to feed a dashboard. The team also must provide the ability to perform one-time queries and analytical reporting in a scalable manner.
Which solution meets these requirements MOST cost-effectively?

A. Use Amazon EMR for processing incoming data. Use AWS Step Functions for workflow orchestration. Use Apache Hive for one-time queries and analytical reporting. Bulk ingest the data in Amazon OpenSearch Service (Amazon Elasticsearch Service). Use OpenSearch Dashboards (Kibana) on Amazon OpenSearch Service (Amazon Elasticsearch Service) for the dashboard.

B. Use Amazon EMR for processing incoming data. Use AWS Step Functions for workflow orchestration. Use Amazon Athena for one-time queries and analytical reporting. Use Amazon QuickSight for the dashboard.

C. Use AWS Glue for processing incoming data. Use AWS Step Functions for workflow orchestration. Use Amazon Redshift Spectrum for one-time queries and analytical reporting. Use OpenSearch Dashboards (Kibana) on Amazon OpenSearch Service (Amazon Elasticsearch Service) for the dashboard.

D. Use AWS Glue for processing incoming data. Use AWS Lambda and S3 Event Notifications for workflow orchestration. Use Amazon Athena for one-time queries and analytical reporting. Use Amazon QuickSight for the dashboard.

Question #145 *Topic 1*

A marketing company collects data from third-party providers and uses transient Amazon EMR clusters to process this data. The company wants to host an
Apache Hive metastore that is persistent, reliable, and can be accessed by EMR clusters and multiple AWS services and accounts simultaneously.
The metastore must also be available at all times.
Which solution meets these requirements with the LEAST operational overhead?

A. Use AWS Glue Data Catalog as the metastore

B. Use an external Amazon EC2 instance running MySQL as the metastore

C. Use Amazon RDS for MySQL as the metastore

D. Use Amazon S3 as the metastore

Question #146 *Topic 1*

A data engineer is using AWS Glue ETL jobs to process data at frequent intervals. The processed data is then copied into Amazon S3. The ETL jobs run every 15 minutes. The AWS Glue Data Catalog partitions need to be updated automatically after the completion of each job.
Which solution will meet these requirements MOST cost-effectively?

A. Use the AWS Glue Data Catalog to manage the data catalog. Define an AWS Glue workflow for the ETL process. Define a trigger within the workflow that can start the crawler when an ETL job run is complete.

B. Use the AWS Glue Data Catalog to manage the data catalog. Use AWS Glue Studio to manage ETL jobs. Use the AWS Glue Studio feature that supports updates to the AWS Glue Data Catalog during job runs.

C. Use an Apache Hive metastore to manage the data catalog. Update the AWS Glue ETL code to include the enableUpdateCatalog and partitionKeys arguments.

D. Use the AWS Glue Data Catalog to manage the data catalog. Update the AWS Glue ETL code to include the enableUpdateCatalog and partitionKeys arguments.

Question #147 *Topic 1*

A reseller that has thousands of AWS accounts receives AWS Cost and Usage Reports in an Amazon S3 bucket. The reports are delivered to the S3 bucket in the following format:

<example-report-prefix>/<example-report-name>/yyyymmdd-yyyymmdd/<example-report-name>.parquet

An AWS Glue crawler crawls the S3 bucket and populates an AWS Glue Data Catalog with a table. Business analysts use Amazon Athena to query the table and create monthly summary reports for the AWS accounts. The business analysts are experiencing slow queries because of the accumulation of reports from the last

5 years. The business analysts want the operations team to make changes to improve query performance.

Which action should the operations team take to meet these requirements?

A. Change the file format to .csv.zip

B. Partition the data by date and account ID

C. Partition the data by month and account ID

D. Partition the data by account ID, year, and month

Question #148 *Topic 1*

A company using Amazon QuickSight Enterprise edition has thousands of dashboards, analyses, and datasets. The company struggles to manage and assign permissions for granting users access to various items within QuickSight. The company wants to make it easier to implement sharing and permissions management.

Which solution should the company implement to simplify permissions management?

A. Use QuickSight folders to organize dashboards, analyses, and datasets. Assign individual users permissions to these folders.

B. Use QuickSight folders to organize dashboards, analyses, and datasets. Assign group permissions by using these folders.

C. Use AWS IAM resource-based policies to assign group permissions to QuickSight items.

D. Use QuickSight user management APIs to provision group permissions based on dashboard naming conventions.

Question #149 *Topic 1*

A company has 10-15 |¢|’ 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 option for data format and compression meets these requirements?

A. CSV compressed with zip

B. JSON compressed with bzip2

C. Apache Parquet compressed with Snappy

D. Apache Avro compressed with LZO

Question #150 *Topic 1*

A company uses Amazon Redshift to store its data. The reporting team runs ad-hoc queries to generate reports from the Amazon Redshift database. The reporting team recently started to experience inconsistencies in report generation. Ad-hoc queries used to generate reports that would typically take minutes to run can take hours to run. A data analytics specialist debugging the issue finds that ad-hoc queries are stuck in the queue behind long-running queries.
How should the data analytics specialist resolve the issue?

A. Create partitions in the tables queried in ad-hoc queries.

B. Configure automatic workload management (WLM) from the Amazon Redshift console.

C. Create Amazon Simple Queue Service (Amazon SQS) queues with different priorities. Assign queries to a queue based on priority.

D. Run the VACUUM command for all tables in the database.

Question #151 *Topic 1*

A company provides an incentive to users who are physically active. The company wants to determine how active the users are by using an application on their mobile devices to track the number of steps they take each day. The company needs to ingest and perform near-real-time analytics on live data. The processed data must be stored and must remain available for 1 year for analytics purposes.
Which solution will meet these requirements with the LEAST operational overhead?

A. Use Amazon Cognito to write the data from the application to Amazon DynamoDB. Use an AWS Step Functions workflow to create a transient Amazon EMR cluster every hour and process the new data from DynamoDB. Output the processed data to Amazon Redshift for analytics. Archive the data from Amazon Redshift after 1 year.

B. Ingest the data into Amazon DynamoDB by using an Amazon API Gateway API as a DynamoDB proxy. Use an AWS Step Functions workflow to create a transient Amazon EMR cluster every hour and process the new data from DynamoDB. Output the processed data to Amazon Redshift to run analytics calculations. Archive the data from Amazon Redshift after 1 year.

C. Ingest the data into Amazon Kinesis Data Streams by using an Amazon API Gateway API as a Kinesis proxy. Run Amazon Kinesis Data Analytics on the stream data. Output the processed data into Amazon S3 by using Amazon Kinesis Data Firehose. Use Amazon Athena to run analytics calculations. Use S3 Lifecycle rules to transition objects to S3 Glacier after 1 year.

D. Write the data from the application into Amazon S3 by using Amazon Kinesis Data Firehose. Use Amazon Athena to run the analytics on the data in Amazon S3. Use S3 Lifecycle rules to transition objects to S3 Glacier after 1 year.

Question #152 *Topic 1*

A company needs to implement a near-real-time messaging system for hotel inventory. The messages are collected from 1,000 data sources and contain hotel inventory data. The data is then processed and distributed to 20 HTTP endpoint destinations. The range of data size for messages is 2-500 KB.

The messages must be delivered to each destination in order. The performance of a single destination HTTP endpoint should not impact the performance of the delivery for other destinations.

Which solution meets these requirements with the LOWEST latency from message ingestion to delivery?

A. Create an Amazon Kinesis data stream, and ingest the data for each source into the stream. Create 30 AWS Lambda functions to read these messages and send the messages to each destination endpoint.

B. Create an Amazon Kinesis data stream, and ingest the data for each source into the stream. Create a single enhanced fan-out AWS Lambda function to read these messages and send the messages to each destination endpoint. Register the function as an enhanced fan-out consumer.

C. Create an Amazon Kinesis Data Firehose delivery stream, and ingest the data for each source into the stream. Configure Kinesis Data Firehose to deliver the data to an Amazon S3 bucket. Invoke an AWS Lambda function with an S3 event notification to read these messages and send the messages to each destination endpoint.

D. Create an Amazon Kinesis data stream, and ingest the data for each source into the stream. Create 20 enhanced fan-out AWS Lambda functions to read these messages and send the messages to each destination endpoint. Register the 20 functions as enhanced fan-out consumers.

Question #153 *Topic 1*

A retail company stores order invoices in an Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster Indices on the cluster are created monthly.

Once a new month begins, no new writes are made to any of the indices from the previous months. The company has been expanding the storage on the Amazon

OpenSearch Service (Amazon Elasticsearch Service) cluster to avoid running out of space, but the company wants to reduce costs. Most searches on the cluster are on the most recent 3 months of data, while the audit team requires infrequent access to older data to generate periodic reports. The most recent 3 months of data must be quickly available for queries, but the audit team can tolerate slower queries if the solution saves on cluster costs

Which of the following is the MOST operationally efficient solution to meet these requirements?

A. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to store the indices in Amazon S3 Glacier. When the audit team requires the archived data, restore the archived indices back to the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster.

B. Archive indices that are older than 3 months by taking manual snapshots and storing the snapshots in Amazon S3. When the audit team requires the archived data, restore the archived indices back to the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster.

C. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to migrate the indices to Amazon OpenSearch Service (Amazon Elasticsearch Service) UltraWarm storage.

D. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to migrate the indices to Amazon OpenSearch Service (Amazon Elasticsearch Service) UltraWarm storage. When the audit team requires the older data, migrate the indices in UltraWarm storage back to hot storage.

Question #154 *Topic 1*

A financial services company is building a data lake solution on Amazon S3. The company plans to use analytics offerings from AWS to meet user needs for one- time querying and business intelligence reports. A portion of the columns will contain personally identifiable information (PII) Only authorized users should be able to see plaintext PII data.
What is the MOST operationally efficient solution that meets these requirements?

A. Define a bucket policy for each S3 bucket of the data lake to allow access to users who have authorization to see PII data. Catalog the data by using AWS Glue. Create two IAM roles. Attach a permissions policy with access to PII columns to one role. Attach a policy without these permissions to the other role.

B. Register the S3 locations with AWS Lake Formation. Create two IAM roles. Use Lake Formation data permissions to grant Select permissions to all of the columns for one role. Grant Select permissions to only columns that contain non-PII data for the other role.

C. Register the S3 locations with AWS Lake Formation. Create an AWS Glue job to create an ETL workflow that removes the PII columns from the data and creates a separate copy of the data in another data lake S3 bucket. Register the new S3 locations with Lake Formation. Grant users the permissions to each data lake data based on whether the users are authorized to see PII data.

D. Register the S3 locations with AWS Lake Formation. Create two IAM roles. Attach a permissions policy with access to PII columns to one role. Attach a policy without these permissions to the other role. For each downstream analytics service, use its native security functionality and the IAM roles to secure the PII data.

Question #155 *Topic 1*

A gaming company is building a serverless data lake. The company is ingesting streaming data into Amazon Kinesis Data Streams and is writing the data to
Amazon S3 through Amazon Kinesis Data Firehose. The company is using 10 MB as the S3 buffer size and is using 90 seconds as the buffer interval. The company runs an AWS Glue ETL job to merge and transform the data to a different format before writing the data back to Amazon S3.
Recently, the company has experienced substantial growth in its data volume. The AWS Glue ETL jobs are frequently showing an OutOfMemoryError error.
Which solutions will resolve this issue without incurring additional costs? (Choose two.)

A. Place the small files into one S3 folder. Define one single table for the small S3 files in AWS Glue Data Catalog. Rerun the AWS Glue ETL jobs against this AWS Glue table.

B. Create an AWS Lambda function to merge small S3 files and invoke them periodically. Run the AWS Glue ETL jobs after successful completion of the Lambda function.

C. Run the S3DistCp utility in Amazon EMR to merge a large number of small S3 files before running the AWS Glue ETL jobs.

D. Use the groupFiles setting in the AWS Glue ETL job to merge small S3 files and rerun AWS Glue ETL jobs.

E. Update the Kinesis Data Firehose S3 buffer size to 128 MB. Update the buffer interval to 900 seconds.

Question #156 *Topic 1*

A healthcare company ingests patient data from multiple data sources and stores it in an Amazon S3 staging bucket. An AWS Glue ETL job transforms the data, which is written to an S3-based data lake to be queried using Amazon Athena. The company wants to match patient records even when the records do not have a common unique identifier.
Which solution meets this requirement?

A. Use Amazon Macie pattern matching as part of the ETLjob

B. Train and use the AWS Glue PySpark filter class in the ETLjob

C. Partition tables and use the ETL job to partition the data on patient name

D. Train and use the AWS Glue FindMatches ML transform in the ETLjob

Question #157 *Topic 1*

A social media company is using business intelligence tools to analyze its data for forecasting. The company is using Apache Kafka to ingest the low-velocity data in near-real time. The company wants to build dynamic dashboards with machine learning (ML) insights to forecast key business trends. The dashboards must provide hourly updates from data in Amazon S3. Various teams at the company want to view the dashboards by using Amazon QuickSight with ML insights. The solution also must correct the scalability problems that the company experiences when it uses its current architecture to ingest data.
Which solution will MOST cost-effectively meet these requirements?

A. Replace Kafka with Amazon Managed Streaming for Apache Kafka. Ingest the data by using AWS Lambda, and store the data in Amazon S3. Use QuickSight Standard edition to refresh the data in SPICE from Amazon S3 hourly and create a dynamic dashboard with forecasting and ML insights.

B. Replace Kafka with an Amazon Kinesis data stream. Use an Amazon Kinesis Data Firehose delivery stream to consume the data and store the data in Amazon S3. Use QuickSight Enterprise edition to refresh the data in SPICE from Amazon S3 hourly and create a dynamic dashboard with forecasting and ML insights.

C. Configure the Kafka-Kinesis-Connector to publish the data to an Amazon Kinesis Data Firehose delivery stream that is configured to store the data in Amazon S3. Use QuickSight Enterprise edition to refresh the data in SPICE from Amazon S3 hourly and create a dynamic dashboard with forecasting and ML insights.

D. Configure the Kafka-Kinesis-Connector to publish the data to an Amazon Kinesis Data Firehose delivery stream that is configured to store the data in Amazon S3. Configure an AWS Glue crawler to crawl the data. Use an Amazon Athena data source with QuickSight Standard edition to refresh the data in SPICE hourly and create a dynamic dashboard with forecasting and ML insights.

Question #158 *Topic 1*

A manufacturing company is storing data from its operational systems in Amazon S3. The company's business analysts need to perform one-time queries of the data in Amazon S3 with Amazon Athena. The company needs to access the Athena network from the on-premises network by using a JDBC connection. The company has created a VPC Security policies mandate that requests to AWS services cannot traverse the Internet.
Which combination of steps should a data analytics specialist take to meet these requirements? (Choose two.)

A. Establish an AWS Direct Connect connection between the on-premises network and the VPC.

B. Configure the JDBC connection to connect to Athena through Amazon API Gateway.

C. Configure the JDBC connection to use a gateway VPC endpoint for Amazon S3.

D. Configure the JDBC connection to use an interface VPC endpoint for Athena.

E. Deploy Athena within a private subnet.

Question #159 *Topic 1*

A company stores revenue data in Amazon Redshift. A data analyst needs to create a dashboard so that the company's sales team can visualize historical revenue and accurately forecast revenue for the upcoming months.
Which solution will MOST cost-effectively meet these requirements?

A. Create an Amazon QuickSight analysis by using the data in Amazon Redshift. Add a custom field in QuickSight that applies a linear regression function to the data. Publish the analysis as a dashboard.

B. Create a JavaScript dashboard by using D3.js charts and the data in Amazon Redshift. Export the data to Amazon SageMaker. Run a Python script to run a regression model to forecast revenue. Import the data back into Amazon Redshift. Add the new forecast information to the dashboard.

C. Create an Amazon QuickSight analysis by using the data in Amazon Redshift. Add a forecasting widget Publish the analysis as a dashboard.

D. Create an Amazon SageMaker model for forecasting. Integrate the model with an Amazon QuickSight dataset. Create a widget for the dataset. Publish the analysis as a dashboard.

Question #160 *Topic 1*

A company is using an AWS Lambda function to run Amazon Athena queries against a cross-account AWS Glue Data Catalog. A query returns the following error:

HIVE_METASTORE_ERROR -
The error message states that the response payload size exceeds the maximum allowed size. The queried table is already partitioned, and the data is stored in an
Amazon S3 bucket in the Apache Hive partition format.
Which solution will resolve this error?

A. Modify the Lambda function to upload the query response payload as an object into the S3 bucket. Include an S3 object presigned URL as the payload in the Lambda function response.

B. Run the MSCK REPAIR TABLE command on the queried table.

C. Create a separate folder in the S3 bucket. Move the data files that need to be queried into that folder. Create an AWS Glue crawler that points to the folder instead of the S3 bucket.

D. Check the schema of the queried table for any characters that Athena does not support. Replace any unsupported characters with characters that Athena supports.

Question #161 *Topic 1*

A machinery company wants to collect data from sensors. A data analytics specialist needs to implement a solution that aggregates the data in near-real time and saves the data to a persistent data store. The data must be stored in nested JSON format and must be queried from the data store with a latency of single-digit milliseconds.
Which solution will meet these requirements?

A. Use Amazon Kinesis Data Streams to receive the data from the sensors. Use Amazon Kinesis Data Analytics to read the stream, aggregate the data, and send the data to an AWS Lambda function. Configure the Lambda function to store the data in Amazon DynamoDB.

B. Use Amazon Kinesis Data Firehose to receive the data from the sensors. Use Amazon Kinesis Data Analytics to aggregate the data. Use an AWS Lambda function to read the data from Kinesis Data Analytics and store the data in Amazon S3.

C. Use Amazon Kinesis Data Firehose to receive the data from the sensors. Use an AWS Lambda function to aggregate the data during capture. Store the data from Kinesis Data Firehose in Amazon DynamoDB.

D. Use Amazon Kinesis Data Firehose to receive the data from the sensors. Use an AWS Lambda function to aggregate the data during capture. Store the data in Amazon S3.

Question #162 *Topic 1*

An ecommerce company ingests a large set of clickstream data in JSON format and stores the data in Amazon S3. Business analysts from multiple product divisions need to use Amazon Athena to analyze the data. The company's analytics team must design a solution to monitor the daily data usage for Athena by each product division. The solution also must produce a warning when a division exceeds its quota.
Which solution will meet these requirements with the LEAST operational overhead?

A. Use a CREATE TABLE AS SELECT (CTAS) statement to create separate tables for each product division. Use AWS Budgets to track Athena usage. Configure a threshold for the budget. Use Amazon Simple Notification Service (Amazon SNS) to send notifications when thresholds are breached.

B. Create an AWS account for each division. Provide cross-account access to an AWS Glue Data Catalog to all the accounts. Set an Amazon CloudWatch alarm to monitor Athena usage. Use Amazon Simple Notification Service (Amazon SNS) to send notifications.

C. Create an Athena workgroup for each division. Configure a data usage control for each workgroup and a time period of 1 day. Configure an action to send notifications to an Amazon Simple Notification Service (Amazon SNS) topic.

D. Create an AWS account for each division. Configure an AWS Glue Data Catalog in each account. Set an Amazon CloudWatch alarm to monitor Athena usage. Use Amazon Simple Notification Service (Amazon SNS) to send notifications.

Question #163 *Topic 1*

A banking company is currently using Amazon Redshift for sensitive data. An audit found that the current cluster is unencrypted. Compliance requires that a database with sensitive data must be encrypted using a hardware security module (HSM) with customer managed keys.
Which modifications are required in the cluster to ensure compliance?

A. Create a new HSM-encrypted Amazon Redshift cluster and migrate the data to the new cluster.

B. Modify the DB parameter group with the appropriate encryption settings and then restart the cluster.

C. Enable HSM encryption in Amazon Redshift using the command line.

D. Modify the Amazon Redshift cluster from the console and enable encryption using the HSM option.

Question #164 *Topic 1*

An advertising company has a data lake that is built on Amazon S3. The company uses AWS Glue Data Catalog to maintain the metadata. The data lake is several years old and its overall size has increased exponentially as additional data sources and metadata are stored in the data lake. The data lake administrator wants to implement a mechanism to simplify permissions management between Amazon S3 and the Data Catalog to keep them in sync.
Which solution will simplify permissions management with minimal development effort?

A. Set AWS Identity and Access Management (IAM) permissions for AWS Glue

B. Use AWS Lake Formation permissions

C. Manage AWS Glue and S3 permissions by using bucket policies

D. Use Amazon Cognito user pools