1. QUESTION

A data analyst of a luxury retail company needs to transfer 5TB of log data from an Amazon S3 bucket into the Hadoop Distributed File System (HDFS) of the Amazon EMR Cluster. Upon review, the bucket has more than a thousand files of varying sizes.

What is the most efficient solution to achieve this?

a) Write an AWS Lambda function that transfers the files from the bucket into an Amazon EC2 instance. Afterward, create a cronjob in the instance that transfers the files to the HDFS.

b) Use `S3DistCP` tool to transfer the data from the S3 bucket with the existing Amazon EMR Cluster.

c) Use AWS Import/Export to mail a portable storage device to AWS and transfer data directly using Amazon's internal network.

d) Use `DistCP` tool to copy the Amazon S3 files into the HDFS. Configure the `-m` option to 100 and `-filelimit` option to 1000.

2. QUESTION

A startup is using a data catalog to store, annotate, and share metadata in the cloud. The metadata tables in the catalog determine the structure of data in various data stores including its other attributes. The data stores being used are Amazon S3, Amazon RDS, Amazon Redshift, and Amazon DynamoDB. A Data Analyst needs to create a solution that will populate the data catalog on a scheduled basis.

Which of the following can be done to achieve this requirement with the least amount of effort?

a) Create a data catalog using Amazon RDS and schedule the AWS Glue crawler to update the tables in the catalog.

b) Set up an Apache Hive metastore in Amazon EMR and configure the AWS Glue crawler to connect to the data store.

c) Set up an AWS Glue crawler schedule to populate the data catalog.

d) Create a DynamoDB table and use a Lambda function to process the records in the DynamoDB stream.

3. QUESTION

A company is using Amazon Redshift to run complex analytic queries against petabytes of structured data. The Data Analyst received a report that the development team takes 2 hours to analyze the data, whereas the operations team only takes a few minutes. The Data Analyst must resolve the performance issue of the development team's queries.

Which of the following options can satisfy the given requirement?

a) Set up a read replica in Amazon Redshift and run the development team queries on the read replica.
b) Reboot the cluster in Amazon Redshift and run the queries again in the development team.
c) Set up workload management in Amazon Redshift and assign a designated user group for each team.
d) Suspend the long queries of the development team and resume the queries afterward.

4. QUESTION

A data engineer is tasked with creating a data lake for a company’s business data hosted on Amazon S3. These data are saved in the Apache Parquet format and will be used for complex analytic jobs. The company expects that users will have granular access to the data sets.

How can the data engineer meet these requirements in the shortest amount of time and in the MOST cost-effective manner?

a) Using AWS Lake Formation, register the S3 bucket name where the data is stored. For data security, assign granular permissions to different users.
b) Import data from Amazon S3 using an AWS Lake Formation blueprint. Control metadata access with AWS Glue Data Catalog resource policies and table-level access with S3 bucket policies.
c) Run an AWS Glue crawler to create a data catalog for existing data and register its Amazon S3 path location on Lake Formation. For data security, assign granular permissions to different users.
d) Import data from Amazon S3 using an AWS Lake Formation blueprint. Then, provide users granular Lake Formation permissions accordingly.

5. QUESTION

A company has invested in upgrading its data analytics capabilities and improving data processing with its data warehouse, which includes AWS Glue, Amazon Redshift, and Amazon S3. The Data Analyst wants to modify the current ETL workflow and automate another Glue job after the AWS Glue crawler schedule completes. The manager seeks to minimize maintenance and development time.

How can the Data Analyst achieve these requirements?

a) Write an AWS Lambda function that runs the AWS Glue ETL job. Create a rule in Amazon CloudWatch Events that triggers the Lambda process after the crawler completes.

b) Create an Amazon EC2 instance and store the ETL job file inside. Create a rule in Amazon CloudWatch Events that triggers the ETL job after the crawler completes.

c) Create an AWS Glue trigger to directly start a job after the crawler completes.

d) Add an AWS Glue workflow. Create a workflow trigger that runs the ETL job. Create another workflow trigger that prompts the first trigger after the crawler job completes.

6. QUESTION

A Data Analyst is investigating an ETL performance issue, which occurred after a huge amount of data was loaded into a table residing in an Amazon Redshift Cluster.
The `COPY` command ran at expected time duration, but the regular `VACUUM` job took 3 hours longer to complete than usual. The analyst later discovered that a schema change was made to the table since its last run. Furthermore, no other user was logged in or ran another `VACUUM` process in the cluster.

Which of the following are most likely causing the latency? (Select TWO)

a) The `VACUUM` operation is run too frequently.

b) The table has ten more additional columns than the previous run.

c) The `VACUUM` operation was run with `BOOST` option.

d) The source data was not loaded in sort key order.

e) After the load, the table has a very low percentage of unsorted data.

7. QUESTION

A Data Analyst is running a data profiler using Amazon EMR. The results are stored in AWS Glue Data Catalog and an S3 bucket. The Data Analyst uses Amazon Athena and Amazon QuickSight for analysis and data visualization. The Data Catalog is updated to include a new data profiler which stores metrics to a separate Amazon S3 bucket. A new Amazon Athena table is created to reference the new S3 bucket. The Data Analyst used the Athena table as a new data source on Amazon QuickSight, however, the import into SPICE (Super-fast, Parallel, In-memory Calculation Engine) failed.

How should the Data Analyst resolve the issue?

a) Configure the permissions for the AWS Glue Data Catalog from the AWS Glue console.
b) Configure the permissions for the S3 bucket from the Amazon Athena and Amazon QuickSight console.
c) Configure the permissions for the new S3 bucket from the Amazon QuickSight console.
d) Configure the permissions for the new S3 bucket from the Amazon S3 console.

8. QUESTION

A data processing pipeline built with Amazon Kinesis Data Streams and Amazon EMR is ingesting and processing a large amount of small data from several sensor nodes. The data is collected through the data stream and delivered to an Amazon S3 bucket. The accumulated data is queried and processed once a day by a scheduled PySpark Job running on an Amazon EMR cluster.

The data analytics team responsible for the pipeline wants to redesign the current architecture to reduce the data processing cost and improve its efficiency. As much as possible, the solution must enable the team to leverage its technical expertise in PySpark.

Which design meets the criteria of the proposed solution?

a) Deploy a Kinesis Data Firehose delivery stream to collect and convert sensor data to Apache Parquet format. Deliver the transformed data into an Amazon S3 bucket. Process the data from the bucket using a PySpark Job running on an Amazon EMR cluster.
b) Replace the Amazon EMR with AWS Glue. Program an AWS Glue ETL script in Python to merge the small sensor data into larger files and convert them to Apache Parquet format.

c) Create an AWS Lambda function and choose Python as the runtime environment. Configure the Kinesis data stream to trigger the Lambda function to process sensor data.

d) Develop a Python application that uses the Kinesis Client Library (KCL) to consume and process data from the Kinesis data stream.

9. QUESTION

A multinational corporation is using Amazon Athena to analyze the data sets stored in Amazon S3. The Data Analyst needs to implement a solution that will control the maximum amount of data scanned in the S3 bucket and ensure that if the query exceeded the limit, all the succeeding queries will be canceled.

Which of the following approach can be used to fulfill this requirement?

a) Integrate API Gateway with Amazon Athena. Configure an account-level throttling to control the queries in the S3 bucket.

b) Create an IAM policy that will throttle the data limits in the primary workgroup.

c) Set up a workload management (WLM) assignment rule in the primary workgroup.

d) Set data limits in the per query data usage control.

10. QUESTION

A company has started to invest in data analytics and wants to build its data warehousing capabilities from 10 Amazon S3 buckets containing at least a hundred files for each one. The data analytics team decided to use AWS Glue to organize the data lake and build its new Glue Data Catalog. However, they noticed that the crawler has been running for too long and wants to improve its performance.

Which action will help speed up the performance of the crawler job?

a) Run multiple crawler jobs instead of just one.

b) Configure an incremental crawl.

c) Break the large files into multiple smaller files.

d) Compress the S3 files to make the file sizes smaller.

11. QUESTION

A developer wants to build a Kinesis data analytics application that determines stock trend direction and strength. The trade orders to be analyzed will be acquired from a real-time stream source and will include basic information such as ticker symbol, amount, order type, and timestamp. To enrich the application, the developer has created a CSV file that maps ticker symbols to company name, sector, and stock price. This mapping file will be updated at a 1-minute interval.

Which approach should the developer take to join the input stream with the mapping file while minimizing overhead?

a) Upload the CSV file to an S3 bucket and set it as a reference data source for the application. Save a SQL query joining the streaming data with the reference data. Run the `UpdateApplications` API every minute to refresh the reference source.

b) Create a scheduled Lambda function that converts and imports CSV files as JSON objects to a DynamoDB table. Use the table as a reference data source for the application and save a SQL query joining the streaming data with the reference table.

c) Use an S3 bucket as the storage for the CSV file and schedule an AWS Glue crawler every minute. Configure the application to use the Glue Data Catalog as a reference data source. Save a SQL query joining the streaming data with the reference data.

d) Deliver the streaming data to Amazon S3. Upload the mapping file to a separate S3 bucket and tag it as a reference data source. Use Amazon Athena to run SQL joins. Call the `UpdateApplications` API every minute to refresh the reference source.

12. QUESTION

A company has a cross-platform application running across a fleet of Amazon EC2 instances. The company wants to offer better customer service by building a centralized logging system that will collect application logs into a service that provides a near-real-time search engine. The system is designed to quickly detect application issues for a faster mean time to recovery. The company does not want to worry about management and maintenance operations, such as hardware provisioning and software patching.

How can the company efficiently set up a data logging system within AWS?

a) Stream the data logs to Amazon CloudWatch Logs using the CloudWatch agent. Use CloudWatch Logs Subscription Filters to direct the logs to a Kinesis Data Firehose delivery stream and send the output to Amazon DynamoDB.

b) Stream the data logs to Amazon CloudWatch Logs using the CloudWatch agent. Use CloudWatch Logs Subscription Filters to direct the logs to a Kinesis Data Firehose delivery stream and send the output to Amazon Elasticsearch Service.

c) Stream the data logs to Amazon CloudWatch Logs using the CloudWatch agent. Use CloudWatch Logs Subscription Filters to direct the logs to a Kinesis data stream and send the output to Amazon Elasticsearch Service.

d) Stream the data logs to Amazon CloudWatch Logs using the CloudWatch agent. Use CloudWatch Logs Subscription Filters to direct the logs to a Kinesis Data Firehose delivery stream and send the output to Splunk.

13. QUESTION

A company has recently hired a Data Analyst to uncover any untapped value from the records they collected over the past years. The Data Analyst has been instructed to use Amazon Redshift to analyze the historical records. However, the analyst is unsure of which distribution style should be used.

Which of the following is NOT a best practice when choosing the best distribution style?

a) Use a DISTSTYLE ALL distribution for tables that are not frequently updated.

b) Designate a common column for the fact table and the dimension table.

c) Select a DISTKEY with high cardinality.

d) Select the smallest dimension based on the filtered dataset's size

14. QUESTION

A data analyst is planning to build an application for ingesting live market data. The analyst wants to create a trend indicator that computes moving averages over a 1-minute timeframe in near real-time.

Which of the following options can satisfy the given requirement?

a) Send the market data to Amazon Kinesis Data Firehose and run a sliding window analysis using Kinesis Data Analytics for SQL Applications to calculate the moving averages.

b) Send the market data to Amazon Kinesis Data Firehose and save the results to Amazon S3. Load the data to Amazon Redshift and use the AVG window function to compute moving averages.

c) Send the market data to an Amazon SNS topic and configure Amazon SQS to subscribe to the topic. Use Amazon EC2 spot instances to create cost-effective queue workers to compute moving averages.

d) Send the market data to Amazon Kinesis Data Firehose and run a stagger window analysis using Kinesis Data Analytics for SQL Applications to calculate the moving averages.

15. QUESTION

A recruitment agency scans millions of physical application forms as JPEG files and stores them in an Amazon S3 bucket. The forms contain different information, such as the applicant's full name, home address, phone number, job position, and relevant skill sets.

The agency uses Amazon Textract to extract the metadata values from the scanned forms. The agency wants to build a solution that will enable its data analysts to drive insights by analyzing and screening applications based on the extracted text information. The JPEG files should also be downloadable. The agency prioritizes query performance over cost reduction.

Which method satisfies these requirements?

a) Create an Apache Parquet file to store the metadata from the scanned forms and the path of the image file in Amazon S3. Use AWS Glue Data Catalog to index the information from the Parquet file. Run ad-hoc queries using Amazon Athena.

b) Use Amazon Elasticsearch Service to index the metadata from the scanned forms and the path of the image file in Amazon S3. Search and visualize information using Kibana.

c) Attach object tags to each image to add the metadata. Use S3 Select to query the files based on the extracted information.

d) Use an Amazon Redshift table to store the metadata from the scanned forms and the path of the image file in Amazon S3. Submit custom queries by running SQL commands.

16. QUESTION

A food delivery service startup has thousands of riders that serve hundreds of thousands of customers every day. The number of users is expected to increase due to the effect of the pandemic. As a response, the company’s Data Analyst has decided to move the existing data to Amazon Redshift with the following schema:

1. A trips fact table that contains details about completed deliveries.
2. A riders dimension table for rider profiles.
3. A customer fact table for customer profiles.

The Data Analyst wants to evaluate profitability by analyzing the delivery date and time as well as the destination of each trip. The rider's data almost don't change while the customer's data changes frequently.

How should the Data Analyst design the table to achieve optimal query performance?

a) Designate a DISTSTYLE KEY (destination) distribution for the Trips table and sort by delivery time. Use DISTSTYLE ALL for the Riders table. Use DISTSTYLE EVEN for the Customers table.

b) Designate a DISTSTYLE EVEN distribution for the Trips table and sort by delivery time. Use DISTSTYLE ALL for the Riders table. Use DISTSTYLE EVEN for the Customers table.

c) Designate a DISTSTYLE EVEN distribution for the Riders table and sort by delivery time. Use DISTSTYLE ALL for both fact tables.

d) Designate a DISTSTYLE KEY (destination) distribution for the Trips table and sort by delivery time. Use a DISTSTYLE ALL distribution for the Riders and Customers tables.

17. QUESTION

A media company has several Amazon S3 buckets for storing customer data. The company's security compliance requires all buckets to be encrypted with auditable access trails. The company's data engineer plans to use an Amazon EMR cluster with EMR File System (EMRFS) to process and transform the data.

Which configuration will allow the cluster to access the encrypted data?

a) Export the CMK from the AWS KMS Console. Create a copy of the CMK and store it on the master node. Configure the cluster to use the encryption key.

b) Set the default encryption mode of the cluster's security configuration to use SSE-S3.

c) Create an IAM role for each customer. Add an `ALLOW` statement to grant permission to the role to use the CMK in the Key Policy.

d) Modify the cluster's security configuration by delegating the appropriate CMKs for each bucket under the per bucket encryption overrides.

18. QUESTION

A company needs to load streaming data directly into a data store to analyze the price movements of various financial products. Occasionally, the data will be modified using SQL for custom processing. A Data Analyst has been instructed to create a solution that can aggregate data, run complex analytic queries, and publish the results to an interactive dashboard.

Which of the following is the MOST suitable solution that the Data Analyst should implement for this scenario?

a) Use Amazon Kinesis Data Streams to create a delivery stream and set up an Amazon Redshift data warehouse as its destination. Create an interactive dashboard using Amazon QuickSight and select Amazon Redshift as the data source.

b) Use Amazon Kinesis Data Streams to create a delivery stream and an Amazon S3 bucket as its destination. Create an interactive dashboard using Amazon QuickSight and select Amazon S3 as the data source.

c) Use Amazon Kinesis Data Firehose to create a delivery stream and a custom Amazon S3 bucket integrated with Amazon Athena as its destination. Create an interactive dashboard using Amazon QuickSight and select Amazon Athena as the data source.

d) Use Amazon Kinesis Data Firehose to create a delivery stream and set an up Amazon Redshift data warehouse as its destination. Create an interactive dashboard using Amazon QuickSight and select Amazon Redshift as the data source.

19. QUESTION

A Data Engineer working for an advertising agency has to perform an advertisement split testing (A/B testing) for a customer. She needs to collate data based on user feedback and social media reactions. The collected data will be processed and analyzed to identify which ad is more effective. For future analysis, the Data Engineer must catalog the data on a data storage as key-value pairs that require immediate access. She should also have the ability to read, write, and manage petabytes of data using a SQL-like interface. A solution with low operational overhead is preferred.

Which method meets these requirements?

a) Stream and process data with Amazon Kinesis Data Firehose. Save the data to Amazon S3 Standard-IA and use Amazon Athena for analysis.

b) Automate the data transformation with AWS Data Pipeline and use Amazon Kinesis Data Analytics for analysis. Save the data to an Amazon EBS Cold HDD (sc1) volume.

c) Automate the data transformation with AWS Data Pipeline and use Amazon Redshift Spectrum for analysis. Save the data to Amazon S3 Glacier.

d) Analyze data with Apache Hive on Amazon EMR. Save the data to an Amazon DynamoDB table.

20. QUESTION

A large enterprise plans to query data that resides in multiple AWS accounts from a central data lake. Each business unit has a separate account that uses an Amazon S3 bucket to store data unique to its business language. Each account also uses a data catalog using AWS Glue Data Catalog.

The administrator was tasked to enforce role-based access controls for the data lake. Junior Data Analysts from each unit should only have read access to their data. Senior Data Analysts, on the other hand, is allowed to have access in all business units, but for specific columns only.

Which solution will minimize operational overhead and reduce overall costs while meeting the required access patterns?

a) Use AWS Organizations to centrally manage all AWS accounts. Use AWS Glue to migrate all the data from the various S3 buckets in every account to the central data lake account. Grant fine-grained permissions to each user with the corresponding access to specific tables and columns using IAM roles.

b) Maintain the current account structure, create a secondary central data lake, and catalog data across multiple accounts to the new central data lake using AWS Glue Data Catalog. Grant cross-account access for the AWS Glue in the central account to crawl data from the S3 buckets in various accounts to populate the catalog table. Grant fine-grained access controls in the Data Catalog and Amazon S3 to allow the Senior Data Analysts to query specific tables and columns.

c) Build a data lake storage in individual AWS accounts. Catalog data across multiple accounts to the central data lake account using AWS Lake Formation. Update the S3 bucket policy in each account to grant access to the AWS Lake Formation service-linked role. Use Lake Formation permissions to grant fine-grained access controls for the Senior Data Analysts to query specific tables and columns.

d) Build a data lake storage in individual AWS accounts. Create a central S3 bucket in the data lake account and use an AWS Lake Formation Blueprint to ingest

data from the different S3 buckets into the central S3 bucket. Change the S3 bucket policy in each account to grant access to the AWS Lake Formation service-linked role. Use Lake Formation permissions to grant fine-grained access controls for the Junior and Senior Data Analysts to query specific tables and columns.

21. QUESTION

A group of data scientists is conducting analytical research on the current and past criminal activities in a particular city. Thousands of records are collected and dumped into a private Amazon S3 data lake. The group wants to analyze historical logs dating back 10 years to identify activity patterns and find out where the highest crime hour of the day occurs. The logs contain information such as date, district, address, and the NCIC (National Crime Information Center) code which describes the nature of the offense.

Which of the following methods will optimally improve query performance?

a) Use Apache ORC. Partition by date and sort by NCIC code.
b) Use compressed .csv partitioned by date and sorted by NCIC code.
c) Use Apache Parquet. Partition by NCIC code and sort by date.
d) Use compressed nested JSON partitioned by NCIC code and sorted by date.

22. QUESTION

A company seeks to create a new smart home product line and add it to its existing portfolio. These sensors are configured to support MQ Telemetry Transport (MQTT) protocol and they want to use AWS to integrate these IoT devices into its current AWS infrastructure and services. The Data Analyst wants to process the data for time-series analytics, store them, and create a visual reporting dashboard.

Which managed solution will meet these requirements in the most operationally-efficient manner?

a) Enable an Amazon Kinesis Data Firehose delivery stream to collect the MQTT messages from the devices and store them in an S3 bucket for storage. Use Amazon QuickSight to create the dashboards.
b) Enable an Amazon Kinesis Data Stream to collect the MQTT messages from the devices. Write an AWS Lambda function that sends data from the stream into

AWS IoT Analytics. Use IoT Analytics to process data, store them, and create reports.

c) Use AWS IoT Core to collect the MQTT messages from the devices. Set up the AWS IoT Analytics channel, pipeline, and data store to collect data from IoT Core, process data, store them, and create reports.

d) Enable an Amazon Kinesis Data Stream to collect the MQTT messages from the devices. Write an AWS Lambda function that sends data from the stream into Amazon Kinesis Data Analytics. Use Kinesis Data Analytics to process data, store them, and create reports.

23. QUESTION

A company is developing a data analytics application that will be used by several clients. The application will collect, process, and analyze clickstream data from various websites in real-time.

Which of the following is the most suitable service to use for the application?

a) AWS Glue

b) Amazon Redshift Spectrum

c) Amazon Kinesis

d) Amazon EMR with Compute Optimized Instances

24. QUESTION

A media company uses Amazon Kinesis Data Streams to ingest massive volumes of real-time data every day. Lately, the application manager noticed that this process has slowed down significantly. Upon investigation, a Data Analyst discovered that Kinesis is throttling the write requests, and the write performance is significantly reduced. The application manager wants a quick fix without performing significant changes to the architecture.

Which actions should the Data Analyst do to resolve this issue quickly? (Select TWO.)

a) Disable Enhanced Kinesis stream monitoring.

b) Reduce throttling by increasing the retention period of the data stream.

c) Use random partition keys and adjust accordingly to distribute the hash key space evenly across shards.

d) Use an error retry and exponential backoff mechanism in the consumer logic.

e) Use the `UpdateShardCount` API in Amazon Kinesis to increase the number of shards in the data stream.

25. QUESTION

A large financial institution recently launched a new feature for its online customers. The management tasked the Data Analyst to create a dashboard that will visualize customer transactions made through its online platform. The transactional data will be streamed to Amazon Kinesis Data Firehose with a buffer interval of 60 seconds. The dashboard will display the near-real-time status of the transactions so the analysis of the Kinesis Firehose stream is time-sensitive.

Which of the following should the Data Analyst implement to meet the visualization requirements?

a) Deliver the streaming data of Kinesis Data Firehose to Amazon OpenSearch (Amazon ElasticSearch). Create an OpenSearch dashboard that will display the required analyses and visualizations.

b) Deliver the streaming data of Kinesis Data Firehose to an Amazon S3 bucket. Set the S3 bucket as a source for Amazon SageMaker Jupyter notebook. Run the required analyses and generate visualizations from this notebook.

c) Deliver the streaming data of Kinesis Data Firehose to an Amazon S3 bucket. Create an AWS Glue Catalog from this data and use Amazon Athena to analyze it. Use Amazon Neptune to generate the graphs and visualizations.

d) Deliver the streaming data of Kinesis Data Firehose to Amazon Redshift. Connect the cluster to Amazon QuickSight with SPICE. Use QuickSight to analyze and generate the required visualizations.

26. QUESTION

A company is planning to migrate a legacy Hadoop cluster running on-premises to AWS. The cluster must use the latest Amazon EMR release and include its custom scripts and workflows during the migration. The Data Analyst must reuse the existing Java application code on-premises for data processing in the new EMR cluster.

Which of the following is the most suitable solution to meet the requirement?

a) Submit a `PIG` step in the EMR cluster and compile the Java program using the version of the cluster.

b) Submit a `CUSTOM_JAR` step in the EMR cluster and compile the Java program using the version of the cluster.

c) Submit a `STREAMING` step in the EMR cluster and compile the Java program using the version of the cluster.

d) Add a `spark-submit` script in the EMR cluster and compile the Java program using the version of the cluster.

27. QUESTION

In an effort to increase ethnic diversity, a popular university has decided to assemble a data analytics team to analyze the geographic distribution of their students. The team is responsible for identifying data relationships and extracting useful patterns that will help the university achieve its goal. Furthermore, the team should create data visualizations using Amazon QuickSight to allow various stakeholders to view historical trends. The access to the dashboard must be authenticated via Microsoft Active Directory. The data must be encrypted in transit and at rest.

How should Amazon QuickSight be configured to meet the requirements?

a) Create a dashboard using the Amazon QuickSight Standard edition. Set up an identity federation using SAML 2.0 and use the default encryption settings.

b) Create a dashboard using the Amazon QuickSight Enterprise edition. Set up an identity federation using SAML 2.0 and use the default encryption settings.

c) Create a dashboard using the Amazon QuickSight Standard edition and use Active Directory (AD) connector for authenticating access. Import a key material to create a Customer Managed Key (CMK) in AWS KMS and configure Amazon QuickSight to use that key.

d) Create a dashboard using the Amazon QuickSight Enterprise edition and use Active Directory (AD) connector for authenticating access. Import a key material to create a Customer Managed Key (CMK) in AWS KMS and configure Amazon QuickSight to use that key.

28. QUESTION

A company has a mobile application with millions of users around the world. The company plans to monitor and collect user activity logs in near-real-time using Amazon Kinesis Data Streams. An application hosted in Amazon EC2 instance uses the Amazon Kinesis Data Streams API with the AWS SDK for Java to consume and process data

before sending it to an Amazon Elasticsearch cluster for analysis. There have been several occasions where errors in the application result in data loss on the Amazon ES cluster.

Which configuration must be done to recover missing data in case of application error?

a) Configure an AWS Lambda function that will read the missing data from the data stream and send it to Amazon Elasticsearch. Manually invoke the Lambda function when an application error occurs.
b) Modify the Java application code to use idempotent processing by pointing the shard iterator to the shard position before the application error occurred.
c) Configure the mobile application to send logs to the Amazon Elasticsearch cluster directly from the mobile device of each user.
d) Use an Amazon EMR cluster that runs an Apache Spark Streaming application to read the missing data from the data stream and send it to Amazon Elasticsearch.

29. QUESTION

A company is running an iterative data processing on an Amazon EMR cluster. Each day, the workflow begins by loading log files into an Amazon S3 bucket. The EMR cluster processes them in 20 batch jobs, which takes each job about 30 minutes to complete. The company wants to further reduce EMR cost.

Which configuration should be done to meet these requirements?

a) Use transient Amazon EMR clusters. Shut down the cluster when the log processing is done.
b) Create a long-running EMR cluster that uses instance fleets.
c) Trigger a Lambda function to process the batch jobs.
d) Use persistent Amazon EMR clusters. Shut down the cluster when the log processing is done.

30. QUESTION

A Data Analyst plans to use Amazon Redshift as a data warehouse to store millions of access logs from multiple web servers. The data analyst will run a query over the logs on a per-day basis. The schema has the following tables:

- A dimension table that includes information about the URL, server, and customer’s IP address.
- A fact table that includes information about packet size, timestamp, protocol, and customer’s IP address.

The Data Analyst has to merge the two tables and analyze the result by the customer’s IP address and sort by the corresponding timestamp to understand the customer's behavior.

What should the Data Analyst do to achieve optimal query performance?

a) Use a DISTSTYLE KEY distribution for the dimension and fact table. Designate the customer’s IP address as the DISTKEY for both the dimension and fact table.

b) Use a DISTSTYLE KEY distribution for the dimension and fact table. Designate the customer’s IP address and timestamp as the DISTKEY for the dimension table and fact table respectively.

c) Use a DISTSTYLE EVEN distribution for the dimension table and DISTSTYLE KEY for the fact table. Designate the customer’s IP address as the DISTKEY for the fact table.

d) Use a DISTSTYLE KEY distribution for the dimension table and DISTSTYLE EVEN for the fact table. Designate the customer’s IP address as the DISTKEY for the dimension table.

31. QUESTION

A retail company is preparing sales data from its Amazon S3 data lake for analysts to work on. To generate reports, analysts must have access to data that is loaded into the data lake on a daily basis, as well as terabytes of data archived for the past 12 months.

Which combination of solutions should the company implement? (Select THREE.)

a) Scan and identify the schema of daily incoming data by running an AWS Glue crawler.

b) Instantiate an Amazon EMR cluster to perform data transformations on archived data.

c) Instantiate an Amazon SageMaker notebook instance to perform data transformations on archived data.

d) Run data transformations on daily incoming data using AWS Glue workflows with AWS Glue jobs.

e) Scan and identify the schema of daily incoming data using Amazon Athena.

f) Run data transformations on daily incoming data using an Amazon Redshift cluster.

32. QUESTION

A digital marketing firm has been managing social media activity for a client. The posts are collected in an Amazon Kinesis Data Stream and its shards are partitioned based on username. Posts from each user must be validated in the same order they were received before transferring them into an Amazon Elasticsearch cluster.

Lately, the Data Analyst observed that the posts are slow to show in the Elasticsearch service and would frequently take more than 30 minutes to appear during peak hours.

What should the Data Analyst do to reduce the latency issues?

a) Instead of a standard data stream iterator, use an HTTP/2 stream consumer instead.
b) Use multiple AWS Lambda functions to process the Kinesis data stream using the `Parallelization Factor` feature.
c) Use Amazon Kinesis Data Firehose to read and validate the social media posts before transferring to the Elasticsearch cluster.
d) Reshard the stream to increase the number of shards and change the partition key to social media post views instead.

33. QUESTION

A company has hundreds of web applications hosted in a fleet of EC2 instances. The company requires a cost-effective near real-time server log analysis solution without having to manage any infrastructure. The solution has the following requirements:

- Collect and transform log files into JSON format.
- Can handle delivery failures.
- Can analyze and visualize log data.

Which approach is the most suitable and has the least operational overhead?

a) Create a Kinesis Data Streams stream to ingest the log data and use a Lambda Function for format conversion. Deliver the formatted log files into an S3 bucket. Analyze the log files using Amazon Athena and store the results in a separate

bucket. Use Amazon QuickSight to visualize the logs. Send the failed deliveries to an Amazon S3 bucket

b) Create a Kinesis Data Firehose to ingest the log data and use a Lambda Function for format conversion. Send the formatted log files into Amazon Kinesis Data Analytics for log analysis and store the results into an S3 bucket. Use Amazon QuickSight to visualize the logs. Send the failed deliveries to an Amazon S3 bucket.

c) Create a Kinesis Data Streams stream to ingest the log data and use a Lambda Function for format conversion. Deliver the formatted log files into Amazon Elasticsearch Service for log analysis and visualization. Send the failed deliveries to an Amazon S3 bucket.

d) Create a Kinesis Data Firehose to ingest the log data and use a Lambda Function for format conversion. Send the formatted log files into Amazon Elasticsearch Service for log analysis and visualization. Send the failed deliveries to an Amazon S3 bucket.

34. QUESTION

A company launched a streaming application that reads hundreds of shards from Amazon Kinesis Data Streams then directly stores the results to an Amazon S3 bucket every 15 seconds. The data is then analyzed by the data analytics team using Amazon Athena. The team noticed that the query performance in Athena degrades overtime.

How can the data analytics team improve the performance of Amazon Athena in the most cost-effective way?

a) Replicate the data to three different S3 buckets to achieve parallel processing. Configure Amazon Athena to query multiple buckets.

b) Increase the CPU and memory capacity of the streaming application.

c) Scale the number of shards in the data stream.

d) Optimize the file sizes by merging smaller files into larger objects in Amazon S3.

35. QUESTION

A smart home security company wants to add more features to its current system to enhance security and improve customer satisfaction. The company uses sensors that send nested JSON files asynchronously into a Kinesis data stream by utilizing the Kinesis Producer Library (KPL) in Java. Upon inspection, it was found that a faulty sensor tends to push recorded data to the cloud at irregular intervals. The company has

to design a near-real-time analytics solution to get data from the most updated and healthy sensors.

What solution will allow the company to meet these requirements?

a) Deactivate the buffering on the sensor side by setting the value of the `RecordMaxBufferedTime` configuration parameter of the KPL to “0“. Create a dedicated Kinesis Data Firehose delivery stream for each data stream and enable data transformation by configuring an AWS Lambda Function to flatten the JSON file. Load the processed data to an Amazon S3 bucket. Use an Amazon Redshift cluster to read the data from Amazon S3.

b) Modify the sensors code by utilizing the `PutRecord` or `PutRecords` of the Kinesis Data Streams API from the AWS SDK for Java. Process data from the stream using a Streaming ETL Job in AWS Glue. Create an AWS Lambda to push the processed data into an Amazon Elasticsearch Service cluster.

c) Deactivate the buffering on the sensor side by setting the value of the `RecordMaxBufferedTime` configuration parameter of the KPL to “-1“. Direct the data to Amazon Kinesis Data Analytics for data enrichment using a custom anomaly detection SQL script. Send the enriched data to a fleet of Kinesis data streams with data transformation enabled to flatten the JSON file. Use an Amazon Redshift cluster with dense storage as the destination of data coming from the Kinesis Data Firehose delivery stream.

d) Modify the sensors code by utilizing the `PutRecord` or `PutRecords` of the Kinesis Data Streams API from the AWS SDK for Java. Create an Amazon Kinesis Data Analytics application for data enrichment using a custom anomaly detection SQL script. Send the enriched data to an Amazon Kinesis Data Firehose delivery stream and enable data transformation by configuring an AWS Lambda Function to flatten the JSON file. Use Amazon Elasticsearch Service as the destination of data coming from the Kinesis Data Firehose delivery stream.

36. QUESTION

A Data Analyst needs to dive into the company's diversity demographics and visually show the management team the nationality composition of the employees. Due to confidentiality, the data is located in a restricted file currently stored in Amazon S3 with a robust `Deny` bucket policy.

Which steps should the Analyst take to achieve the requirement? (Select TWO.)

a) Configure the bucket's policy by adding the Amazon QuickSight service role as an exception in the `Deny` policy.

b) Configure QuickSight's security permissions and allow access to Amazon S3.
c) Use Amazon QuickSight and use a Pie Chart to display the data.
d) Use Amazon QuickSight and use a Funnel Chart to display the data.
e) Use Amazon QuickSight and use a Word Cloud to display the data.

37. QUESTION

A company is using Amazon EMR to perform data transformation workloads. The Data Analyst is instructed to conduct an unplanned security audit. Upon checking, the Data Analyst noticed that the EMR cluster's root volumes are not encrypted.

Which of the following options would encrypt the EMR cluster's root volumes?

a) Recreate the EMR cluster using the default configuration.
b) Enable in-transit encryption and recreate the cluster.
c) Enable at-rest encryption for local disks and recreate the cluster.
d) Enable at-rest encryption for EMRFS data in Amazon S3 and recreate the cluster.

38. QUESTION

A Data Analyst needs to delegate access to AWS analytics services. Both the accounting and marketing teams will use the business intelligence tool that run Presto queries on the Amazon EMR cluster. The Data Analyst must grant the Marketing team access to the advertisement table only.

Which of the following should the Data Analyst do to grant the required permissions to each team?

a) Create IAM roles for accounting and marketing users. Use AWS Glue resource policies to grant access to the corresponding table in the AWS Glue Data Catalog.
b) Create IAM roles for accounting and marketing users. Use AWS Glue identity-based policies to grant access for all AWS Glue operations.
c) Create IAM roles for accounting and marketing users. Use AWS Glue identity-based policies to grant access to the AWS Glue Data Catalog resources.

d) Create IAM roles for accounting and marketing users. Use AWS Glue resource policies to grant access for all AWS Glue operations.

39. QUESTION

An energy company is constructing two weather stations to collect and record temperature data from a solar farm.

- Weather station A has 20 sensors with each unique ID
- Weather station B has 10 sensors with each unique ID

Onsite engineers strategically determined the placement and orientation of the weather stations. The company plans to use Amazon Kinesis Data Streams to collect data from each sensor.

A single Kinesis data stream with two shards was created based on the total data throughput gathered from the initial testing. The partition keys were created based on the two station names. A bottleneck on data coming from Station A has been spotted during the dry-run. However, there were no problems with Station B. Upon checking, it was inferred that the currently allocated throughput for the Kinesis Data Streams is still greater than the total stream throughput of the sensor data.

Which solution will resolve the bottleneck issue without increasing the overall cost?

a) Provision a different Kinesis data stream with two shards to stream sensor data coming from Station A.

b) Assign the sensor ID as the partition key instead of the station name.

c) Increase the level of parallelism for greater throughput by increasing the number of shards in Amazon Kinesis Data Streams.

d) Decrease the number of sensors in Station A from 20 to 10 sensors.

40. QUESTION

A company is using an Application Load Balancer (ALB) to distribute the incoming traffic across EC2 instances. The ALB logs are stored in the S3 bucket for further analysis. The log data will be joined with the other tables by a proprietary Business Intelligence (BI) tool that fetches data from an Amazon S3 data lake. A Data Analyst has been instructed to use a JDBC driver to integrate other AWS services with the BI tool.

How can you achieve this requirement with the LEAST effort?

a) Create a persistent cluster in Amazon EMR every night and load the new logs files from Amazon S3 into HDFS. Use Amazon Athena to run queries in HDFS and to connect to the BI tool.

b) Set up a Lambda function to transform and move the new log file to Amazon Redshift. Use Amazon Athena to run queries in Redshift and to connect to the BI tool.

c) Create a transient cluster in Amazon EMR every night and load the new logs files in Amazon Redshift. Use Amazon Athena to run queries in Redshift and to connect to the BI tool.

d) Set up a Lambda function to transform and move the new log file to another S3 bucket. Use Amazon Athena to run queries in the S3 bucket and to connect to the BI tool.

41. QUESTION

A Data Analytics team in Hong Kong uses several data sources, including a 2-node Amazon Redshift cluster in `ap-southeast-1` region, for monthly reporting dashboards.

One of the analysts needs to use new confidential data from the Redshift cluster for reporting and has decided to use Amazon QuickSight Enterprise Edition. Unfortunately, the analyst is unable to find the data through the console.

What is most likely causing the issue?

a) The Redshift cluster is in a private subnet. Amazon QuickSight can only access data in a public subnet.

b) The Redshift cluster is in another VPC. The analyst needs to configure a security group for QuickSight console that allows traffic to and from the cluster.

c) The QuickSight Console used by the analyst is in a different region, while the Redshift cluster is in a VPC. Amazon QuickSight can only access the cluster data in the same region.

d) The data is located in a table that cannot be accessed by the analyst's Redshift cluster user credentials.

42. QUESTION

A company is currently storing customer records on a local data center. Most of the records are infrequently accessed by data engineers performing analysis and data visualizations. The company plans to migrate its data to AWS Cloud for long-term

storage and to leverage the analytics capabilities of Amazon EMR, Amazon Athena, and Amazon QuickSight. The company wants a solution that automates and accelerates the copying of large amounts of data to an AWS storage service. Furthermore, the solution should provide automatic data integrity checks and encryption of data-in-transit and at-rest.

a) Deploy an AWS DataSync agent on an Amazon EC2 instance and replicate the data to an S3 Standard-IA bucket.

b) Deploy an AWS DataSync agent on-premises and use Amazon S3 Transfer Acceleration to migrate the data to an S3 Standard-IA bucket.

c) Deploy an AWS DataSync agent on-premises instance and replicate the data to an S3 Standard-IA bucket.

d) Deploy an AWS DataSync agent on-premises and use AWS Transfer for SSH File Transfer Protocol (SFTP) to migrate the data to an S3 Standard-IA bucket.

43. QUESTION

A company has a Java application hosted on-premises that processes Extract, Load, Transform (ETL) jobs to an Amazon EMR cluster. The company requires its Security Operations (SecOps) team to enable root device volume encryption on all nodes in the EMR cluster. The solution must reduce overhead for the system administrators without modifying the application's code. The SecOps team should also use AWS CloudFormation in creating AWS resources to comply with the company standards.

Which is the MOST suitable solution for the scenario?

a) Provision an Amazon EC2 instance with encrypted root device volumes. Connect to the instance and install Apache Hadoop. Specify the instance in the CloudFormation template.

b) In the CloudFormation template, define an EMR cluster that uses a custom AMI with encrypted root device volume under the `CustomAmiId` property.

c) In the CloudFormation template, define a custom bootstrap action under the `BootstrapActionConfig` property of the EMR cluster to enable Transport Layer Security (TLS).

d) In the CloudFormation template, define a custom bootstrap action under the `BootstrapActionConfig` property of the EMR cluster to encrypt the root device volume of the master node.

44. QUESTION

A company is using Amazon Kinesis Data Streams to capture real-time messages from over 150 websites. To handle the traffic spikes, the data stream has been provisioned with 40 shards to achieve maximum data throughput. An Amazon Kinesis Client Library (KCL) application hosted in an Auto Scaling group of EC2 instances consumes the stream, analyzes the data, and store the results in a DynamoDB table. The average CPU utilization across all servers is 20% including peak times. The DynamoDB table has a provisioned write capacity unit set to 5.

The manager received a report that the application increased its latency during peak times. Upon initial investigation, there are no `ProvisionedThroughputExceededException` errors found on the KCL logs and the CPU Utilization of the instances didn't exceed its limit. The Data Analyst is instructed to implement a solution that will resolve the latency problem.

Which of the following is the best approach to solve this issue?

a) Increase the write throughput of the DynamoDB table.
b) Scale up the KCL application by using a higher EC2 instance type to increase network performance.
c) Increase the number of shards in the Kinesis Data Stream.
d) Update the Auto Scaling group to increase the minimum number of running EC2 instances.

45. QUESTION

An AgriTech startup has managed to set up thousands of harvesting devices with sensor data stored in thousands of small comma-separated value files in an Amazon S3 bucket. The Data Analyst team intends to transfer data in these files into a 6-node Amazon Redshift cluster with 4 node slices each.

How can the team optimize the loading speed of data into the Redshift cluster while minimizing the costs of queries?

a) Transform the data from CSV files to 24 large files in Apache Avro format using an Amazon EMR cluster. Run the `COPY` command to load the file into the Redshift cluster. To minimize costs, query the files with Amazon Athena from the Amazon S3 bucket.
b) Transform the data from CSV files to a single large file in Apache Parquet format using AWS Glue. Run the `COPY` command to load the file into the Redshift cluster. To minimize costs, query the files with Amazon Athena from the Amazon S3 bucket.

c) Transform the data from CSV files to a single large file in Apache Optimized Row Columnar (ORC) format using AWS Glue. Run the `COPY` command to load the file into the Redshift cluster. To minimize costs, query the files with Amazon Athena from the Amazon S3 bucket.

d) Transform the data from CSV files to 24 large files in Apache Parquet format using AWS Glue. Run the `COPY` command to load the file into the Redshift cluster. To minimize costs, query the files with Amazon Athena from the Amazon S3 bucket.

46. QUESTION

A company has multiple data analytics teams that run their own Amazon EMR cluster. The teams have their own metadata for running different SQL queries using Hive. A centralized metadata layer must be created that exposes S3 objects as tables that can be used by all teams.

What should be done to fulfill this requirement?

a) Use Amazon EMR Notebooks.

b) Alter table recover partitions.

c) Enable EMRFS consistent view.

d) Configure an external metastore for Hive.

47. QUESTION

A startup is using Apache HBase in Amazon EMR with a single master node to process its mission-critical workloads. The stored data in its Hadoop Distributed File System (HDFS) is over 8 TB. The Data Analyst has been instructed to provide a solution that provides the highest level of availability to the EMR cluster.

Which is the MOST cost-effective solution that the Data Analyst should implement to fulfill this requirement?

a) Use EMR File System (EMRFS) to store the data and enable the EMRFS consistent view feature. Run the two separate primary EMR clusters in each Availability Zone. Point the two clusters to the same S3 bucket and `hbase.rootdir` location.

b) Launch Spot instances for the core and task nodes. Create a new EMR HBase cluster and configure it in multiple master nodes.

c) Use EMR File System (EMRFS) to store the data and enable the EMRFS consistent view feature. Create a new EMR HBase cluster and configure it to use multiple master nodes.

d) Use EMR File System (EMRFS) to store the data and enable the EMRFS consistent view feature. Launch two separate EMR clusters in two different Availability Zones. Launch a primary cluster with multiple master nodes and create a secondary read replica cluster in a different Availability Zone. Point the two clusters to the same S3 bucket and `hbase.rootdir` location.

48. QUESTION

A company hosts its web application in an Auto Scaling group of Amazon EC2 instances. The data analytics team needs to create a solution that will collect and analyze the logs from all of the EC2 instances running in production. The solution must be highly accessible and allows the viewing of the new log information in near real-time.

Which of the following is the most suitable solution to meet the requirement?

a) Use Amazon CloudWatch Logs subscriptions to process log data in real-time. Send the data to Amazon Kinesis Data Streams, which will deliver the data to Amazon Elasticsearch Service and Amazon QuickSight.

b) Enable the detailed monitoring feature on all the EC2 instances. Use CloudWatch to collect metrics and logs. Analyze the data using Amazon Kinesis Data Analytics.

c) Install the Amazon Kinesis Producer Library agent in the EC2 instances. Use the agent to collect and send the data to Amazon Kinesis Data Streams, which will deliver the data to Amazon Elasticsearch Service and Amazon QuickSight.

d) Install the Amazon Kinesis Producer Library agent in the EC2 instances. Use the agent to collect and send the logs to a data stream. Use the data stream as a source for Amazon Kinesis Data Firehose, which will deliver the log data to Amazon Elasticsearch Service and Kibana.

49. QUESTION

A financial institution is using Amazon S3 to store data in nested JSON format collected from multiple sources. The Data Analyst must create an automated solution that will flatten the nested JSON data into a structured format and analyze the data stored in both Amazon S3 and Amazon Redshift cluster.

Which of the following options would be the most cost-effective solution?

a) Load the data in the Amazon Redshift cluster using the Amazon Redshift `COPY` command. Analyze the data with Amazon Athena.

b) Filter the data in Amazon S3 using the AWS Glue `Filter` class. Create external tables using Amazon Redshift Spectrum and join with the internal tables.

c) Use Amazon S3 Select to retrieve the nested JSON. Load the data into a table using the Amazon Redshift `COPY` command.

d) Transform the data in Amazon S3 using the AWS Glue `Relationalize` PySpark Transforms class. Create external tables using Amazon Redshift Spectrum and join with the other existing tables.

50. QUESTION

A company runs a Reserved Amazon EC2 instance to process ETL jobs before sending the results into an Amazon Redshift cluster. Because of scaling issues, the company eventually replaced the EC2 instance with AWS Glue and Amazon S3. Since the architecture has changed, the Data Analyst must also make necessary changes in the workflow. Part of the new process is to save the Redshift query results to an external storage for occasional analysis.

Which of the following methods is the most cost-efficient solution for the new process?

a) Move the Redshift query results to an Amazon S3 bucket using the `UNLOAD` command.

b) Move the Redshift query results to an external table in Amazon Redshift Spectrum using the `UNLOAD` command.

c) Move the Redshift query results to an Amazon S3 bucket using the `COPY` command.

d) Move the Redshift query results to an external table in Amazon Redshift Spectrum using the `COPY` command.

51. QUESTION

A company wants to implement a continuous monitoring system for the advertisement videos on its social media application. The system will be designed to detect sentiment changes in user feeds and track all video playback issues. The company will collect and analyze data to react to the user sentiment in less than 30 seconds. The transmitted data is in JSON format with a consistent, well-defined schema.

Which collection and processing methods should the company do to meet these requirements?

a) Ingest the data to Amazon Managed Streaming for Kafka (Amazon MSK), and choose a Kinesis Data Analytics (KDA) application as the destination to process, detect, and react to a sentiment change. Directly store the raw output data in a DynamoDB table from the KDA application.

b) Ingest the incoming data into Amazon Kinesis Data Firehose, and deliver the data to an Amazon Kinesis Data Analytics (KDA) application to process, detect, and react to a sentiment change. Configure the KDA application to directly send the raw JSON data to an Amazon S3 bucket.

c) Ingest the incoming data into Amazon Kinesis Data Streams and deliver the data into an S3 bucket. Enable event notifications to trigger a Lambda function that will process, detect, and react to a sentiment change. Store the raw data in a DynamoDB table.

d) Ingest the incoming data into Amazon Kinesis Data Streams, and choose an Amazon Kinesis Data Analytics (KDA) application as the destination to process, detect, and react to a sentiment change. Configure a Kinesis Data Firehose delivery stream as an output of the KDA application to store the raw JSON data in an Amazon S3 bucket.

52. QUESTION

A team of scientists on the South Pole has been gathering climate data for the past two years. The terabytes of data were imported to Amazon S3 and the team will analyze it using Amazon Athena. Each member will run their own analysis so they want to have a limit on the data each member queries. Research funds are limited so the team needs to track all ad-hoc queries and how much data had been processed using Athena.

Which of the following should the scientists implement to achieve these requirements?

a) Create an Athena Workgroup for each member. Assign each member with their own S3 bucket and associated bucket policy. Export query-related metrics for all bucket queries in CloudWatch.

b) Create an S3 bucket for each member and assign specific Athena policies for each member to query their own buckets. Track query-related metrics for all workgroup queries in CloudWatch.

c) Create an S3 bucket for each member and assign bucket policies that grant appropriate permissions to individual IAM users. Export query-related metrics for all bucket queries in CloudWatch.

d) Create an Athena Workgroup for each member and enable the `Publish query metrics to AWS CloudWatch` option. Track query-related metrics for all workgroup queries in CloudWatch.

53. QUESTION

A sports technology company plans to build the latest kneepads version that can collect data from athletes wearing them. The product owner is looking to develop them with wearable medical sensors to ingest near-real-time data securely at scale and store it in durable storage. Furthermore, it should only collect non-confidential information from the streaming data and exclude those classified as sensitive data.

Which solution achieves these requirements with the least operational overhead?

a) Using Amazon Kinesis Data Streams, ingest the streaming data, and use Amazon S3 for durable storage. Write an AWS Lambda function that removes sensitive data. Schedule a separate job that invokes the Lambda function once the data is stored in Amazon S3.

b) Using Amazon Kinesis Data Firehose, ingest the streaming data, and use Amazon S3 for durable storage. Write an AWS Lambda function that removes sensitive data. During the creation of the Kinesis Data Firehose delivery stream, enable record transformation and use the Lambda function.

c) Using Amazon Kinesis Data Streams, ingest the streaming data, and use an Amazon EC2 instance for durable storage. Write an Amazon Kinesis Data Analytics application that removes sensitive data.

d) Using Amazon Kinesis Data Firehose, ingest the streaming data, and use Amazon S3 for durable storage. Write an AWS Lambda function that removes sensitive data. Schedule a separate job that invokes the Lambda function once the data is stored in Amazon S3.

54. QUESTION

A company needs to upgrade an Amazon Redshift cluster to support the new features of its data warehouse application. There will be several changes to the current database such as user permission updates and table schema modifications. Before running the upgrade scripts, the Data Analyst must create point-in-time backups to restore the service to its previous state if problems arise.

Which of the following options could help fulfill this task?

a) Create a manual snapshot of the cluster.
b) Use AWS Lake Formation to automatically take the snapshot of the Amazon Redshift cluster and store the data in Amazon S3.
c) Unload the results of Amazon Redshift to Amazon S3.
d) Restore the service using the automated snapshot.

55. QUESTION

A group of medical researchers is using computer simulations in studying the growth of cancer cells. The simulations generate millions of data points that are partitioned and stored in Amazon S3 One Zone-Infrequent Access (S3 One Zone-IA). The analytics processing for the data is performed on Amazon EMR clusters using EMRFS with consistent view enabled. The researchers noticed that the overall performance of the cluster can't keep up with the increasing number of concurrent queries and analytics jobs running on it. It has been determined that the EMR task nodes are taking longer to list objects in Amazon S3.

Which of the following actions will most likely increase the performance of the cluster in reading Amazon S3 objects?

a) Rebalance the EMRs cluster to span multiple Availability Zones to better spread the load of reading data points from the Amazon S3 bucket.
b) Increase the number of EMR task nodes to better spread the load of reading data points from the Amazon S3 bucket.
c) When storing the data points on the S3 bucket, add a sequential date-based naming as a prefix for the object filename.
d) Increase Amazon S3 read capacity by using a lifecycle policy to change the S3 storage class for the data points from S3 One Zone-IA to S3 Standard.

56. QUESTION

A scientist in a research institution runs an Apache Hive script to batch process agricultural data stored on an Amazon S3 bucket. The script needs to run at 4:00 PM every day after all new data is saved to the S3 bucket. The output of the script is saved in another Amazon S3 bucket. Running the script on a three-node cluster on the building's data center takes about 1-2 hours to complete. The scientist wants to move this batch process to the cloud and run it on a regular basis efficiently.

Which of the following is the most cost-effective solution for scheduling and executing the batch process?

a) Schedule an Amazon CloudWatch Events rule to invoke a Lambda function to run at 4:00 PM daily. Configure your AWS Lambda function to provision an Amazon EMR cluster with Hue (Hadoop User Experience), Apache Hive, and Apache Oozie. Set the termination protection flag to FALSE and use Spot Instances for the core nodes of the cluster. Configure an Oozie workflow in the cluster to invoke the Hive script at bootup.

b) Create an AWS Step Function workflow to schedule running a Lambda function daily at 4:00 PM. Have the Lambda function load the Hive runtime and copy the Hive script. Add a step to call the `RunJobFlow` API to provision an EMR cluster and bootstrap the Hive script.

c) Schedule an Amazon CloudWatch Events rule to invoke a Lambda function to run at 4:00 PM daily. Configure your AWS Lambda function to provision an Amazon EMR cluster with a Hive execution step. On the `RunJobFlow` API, set the `KeepJobFlowAliveWhenNoSteps` to FALSE and disable the termination protection flag.

d) Configure an AWS Glue job to run at 4:00 PM using a time-based schedule. Configure the job to include the Hive script to perform the batch operation at the specified time.

57. QUESTION

A company has an application that sends 80 GB of compressed data from an on-premises server to an S3 bucket every day. The Data Analyst loads this data into an Amazon Redshift cluster by executing the `COPY` command. The cluster is configured with a compute node that has four slices.

Which approach should be made to achieve the optimal performance for the `COPY` command?

a) Split the file into 4 parts, with each part having a size of 20 GB.

b) Split the file into 2 parts, with each part having a size of 40 GB.

c) Split the file into 80 parts, with each part having a size of 1 GB.

d) Split the file into 160 parts with different file sizes. Ensure that each part has a size of less than 1 GB.

58. QUESTION

A car company runs computer simulations to test its new electric engine. Each simulation outputs data to an Amazon S3 bucket and the company uses Apache Hive on Amazon EMR for ad-hoc queries. Several data engineers are complaining that the cluster is slow and sluggish when they simultaneously submit their queries. Almost 90% of all queries are submitted within 1 hour after the completion of each simulation. The Hadoop Distributed File System (HDFS) utilization never exceeds 10% usage for each run.

Which of the following options will help solve the performance issues of the EMR cluster?

a) Set up instance group configurations for both the core and task nodes of the Amazon EMR cluster. Configure an automatic scaling policy to scale-out or scale-in the instance groups depending on the threshold value for the Amazon CloudWatch `YARNMemoryAvailablePercentage` metric.

b) Set up instance fleet configurations for both the core and task nodes. Configure an automatic scaling policy to scale-out or scale-in the instance groups depending on the threshold value for the Amazon CloudWatch `CapacityRemainingGB` metric.

c) Set up instance fleet configurations for both the core and task nodes. Configure an automatic scaling policy to scale-out or scale-in the instance groups depending on the threshold value for the Amazon CloudWatch `YARNMemoryAvailablePercentage` metric.

d) Set up instance group configurations for both the core and task nodes. Configure an automatic scaling policy to scale-out or scale-in the instance groups depending on the threshold value for the Amazon CloudWatch `CapacityRemainingGB` metric.

59. QUESTION

A company has a clickstream analytics solution using Amazon Elasticsearch Service. The solution ingests 2 TB of data from Amazon Kinesis Data Firehose and stores the latest data collected within 24 hours in an Amazon ES cluster. The cluster is running on a single index that has 12 data nodes and 3 dedicated master nodes. The cluster is configured with 3,000 shards and each node has 3 TB of EBS storage attached. The Data Analyst noticed that the query performance of Elasticsearch is sluggish, and some intermittent errors are produced by the Kinesis Data Firehose when it tries to write to the index. Upon further investigation, there were occasional `JVMMemoryPressure` errors found in Amazon ES logs.

What should be done to improve the performance of the Amazon Elasticsearch Service cluster?

a) Improve the cluster performance by increasing the number of master nodes of Amazon Elasticsearch.

b) Improve the cluster performance by decreasing the number of shards of the Amazon Elasticsearch index.

c) Improve the cluster performance by increasing the number of shards of the Amazon Elasticsearch index.

d) Improve the cluster performance by decreasing the number of data nodes of Amazon Elasticsearch.

60. QUESTION

A particle physics laboratory is generating up to 1 TB of data per day as physicists generate simulations for their experiments. The raw data is converted into large .csv files and stored in an Amazon S3 bucket with folders partitioned by date. At the end of each business day, the data is loaded into Amazon Redshift data warehouse to run analysis and detect patterns on the experiments. However, it takes a lot of time whenever data is loaded from the S3 bucket to Redshift.

Which of the following actions will help improve the data loading times?

a) Vacuum the table in Amazon Redshift after loading each .csv file in an unsorted key order to improve the loading time.

b) Stream the large .csv files in parallel to Amazon Kinesis Data Firehose and ingest into Amazon Redshift.

c) Store the .csv files in Amazon S3 but split the large .csv files into smaller chunks. Use the `COPY` command to load the files into Amazon Redshift.

d) Store the .csv files in Amazon S3 in compressed format then issue the `INSERT` command to load the files into Amazon Redshift.

61. QUESTION

An Apache Hive running on Amazon EMR cluster is used and managed by three groups of data scientists. The cluster uses EMRFS to read and write files to three separate S3 buckets — respectively owned by each group. The cluster is accessed from on-premises server through an Active Directory (AD) and authenticated using Kerberos. As part of security requirements, the access to S3 bucket must be limited to the members of the group who owned the bucket.

Which steps should be followed to achieve their goal?

a) Create a service role for the EMR cluster that grants no access to Amazon S3. Create three IAM roles for each group and edit their permissions to ensure that only the group that owned the S3 bucket can access it. Add the additional IAM roles to the cluster's EMR role for the EC2 trust policy. Configure EMRFS security configuration for the three IAM roles that will be assumed by the groups from the Active Directory.

b) Create a service role for the EMR cluster that grants full access to Amazon S3. Create three IAM roles for each group and edit their permissions to ensure that only the group that owned the S3 bucket can access it. Add the additional IAM roles to the cluster's EMR role for the EC2 trust policy. Configure EMRFS security configuration for the three IAM roles that will be assumed by the groups from the Active Directory.

c) Create a service role for the EMR cluster that grants full access to Amazon S3. Create three IAM roles for each group and edit their permissions to ensure that only the group that owned the S3 bucket can access it. Add the service role for the EMR cluster EC2 instances to the trust policies of the three IAM roles. Configure EMRFS security configuration for the three IAM roles that will be assumed by the groups from the Active Directory.

d) Create a service role for the EMR cluster that grants no access to Amazon S3. Create three IAM roles for each group and edit their permissions to ensure that only the group that owned the S3 bucket can access it. Add the service role for the EMR cluster EC2 instances to the trust policies of the three IAM roles. Configure EMRFS security configuration for the three IAM roles that will be assumed by the groups from the Active Directory.

62. QUESTION

A government office is storing state-wide census data to an encrypted Amazon S3 bucket. The census data contains personally identifiable information (PII) and the data analysts were tasked to create population reports from it. The analysts can launch Amazon EMR clusters and load the census data as long as they comply with strict security requirements. The data must not be publicly accessible throughout the process and any created EMR cluster must not be exposed to the public Internet.

Which of the following actions will help meet the EMR compliance requirements with the least amount of effort?

a) Associate a security configuration when creating EMR clusters. Add the rule to allow only internal communication on the security configuration.

b) Ensure that no inbound traffic from IPv4 `0.0.0.0/0` or IPv6 `::/0` is allowed by scheduling regular checks to the attached security group of the EMR clusters.

c) From the Amazon EMR Console, enable the 'block public access' setting to prevent any user from creating a cluster that is accessible from the public Internet.

d) Create Network ACL rules to block public Internet access to all EMR clusters.

63. QUESTION

A streaming platform company uses an Amazon EMR Cluster (v4.0) with Apache Spark for analytics. The data analyst team needs to concatenate several files in Apache Parquet format together and decides to use the `S3DistCP` tool. However, upon completion of the process, they discovered that the generated Parquet files could not be read properly through the application. Moreover, one of the analysts discovered error messages inside the Parquet files.

What should the data analyst team do to fix this?

a) Ensure `--s3ServerSideEncryption` option is enabled during the `s3-dist-cp` job.

b) Use PySpark to concatenate the Parquet files instead.

c) The Parquet files need to be processed incrementally during the `s3-dist-cp` job.

d) Ensure the `--groupBy` and `--targetSize` arguments are set accordingly based on the source files in the `s3-dist-cp` job.

64. QUESTION

A company is working on an analytics platform that ingests a vast amount of CSV-formatted data from multiple sources and stores them in an Amazon S3 Standard bucket. The S3 bucket is expected to store around 25 GB of raw data every day. The company runs Amazon Athena aggregate functions to gain a summarized view of data that dates back 6 months ago, after which the data becomes infrequently accessed. The company's policy requires raw data to be archived 2 years after creation. An average query scans around 200 MB of data with less than a minute response time. A data engineer is required to optimize the cost of running the platform.

Which set of steps should the data engineer do?

a) Compress, partition, and transform the raw data into a row-based data format using an AWS Glue ETL job. Then, query the processed data using Amazon Athena. Create a lifecycle policy that will transfer the processed data in the S3 Standard-IA storage class 6 months after object creation. Create another

lifecycle policy that will archive the raw data into Amazon S3 Glacier based on the last date that the object was accessed.

b) Compress, partition, and transform the raw data into a columnar data format using an AWS Glue ETL job. Then, query the processed data using Amazon Athena. Create a lifecycle policy that will transfer the processed data in the S3 Standard-IA storage class 6 months after object creation. Create another lifecycle policy that will archive the raw data into Amazon S3 Glacier based on the last date that the object was accessed.

c) Compress, partition, and transform the raw data into a columnar data format using an AWS Glue ETL job. Then, query the processed data using Amazon Athena. Create a lifecycle policy that will transfer the processed data in the S3 Standard-IA storage class 6 months after object creation. Create another lifecycle policy that will archive the raw data into Amazon S3 Glacier after 2 years.

d) Compress, partition, and transform the raw data into a row-based data format using an AWS Glue ETL job. Then, query the processed data using Amazon Athena. Create a lifecycle policy that will transfer the processed data in the S3 Glacier storage class 6 months after object creation. Enable expedited retrieval. Create another lifecycle policy that will archive the raw data into Amazon S3 Glacier Deep Archive after 2 years.

65. QUESTION

A digital marketing company uses Amazon DynamoDB and highly-available Amazon EC2 instances for one of its solutions. Its application logs are pushed to Amazon CloudWatch logs. The team of data analysts wants to enrich these logs with data from DynamoDB in near-real-time and use the output for further study.

Which among these steps will enable collection and enrichment based on the requirements stated above?

a) Install Amazon Kinesis Agent on the EC2 instance. Configure the application to write the logs in a local filesystem and configure 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 and enrich it with data from the DynamoDB table. Store the enriched output stream in an Amazon S3 bucket using Amazon Kinesis Data Firehose.

b) Export the EC2 application logs to Amazon S3 on an hourly basis using AWS CLI. Use AWS Glue crawlers to catalog the logs. Configure an AWS Glue connection to the DynamoDB table and an AWS Glue ETL job to enrich the data. Store the enriched data in an Amazon S3 bucket.

c) Write an AWS Lambda function that will export the EC2 application logs to Amazon S3 on an hourly basis. 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 an Amazon S3 bucket.

d) Write an AWS Lambda function that will enrich the logs with the DynamoDB data. Create an Amazon Kinesis Data Firehose delivery stream, configure it to subscribe to Amazon CloudWatch Logs, and set an Amazon S3 bucket as its destination. Create a CloudWatch Logs subscription that sends log events to your delivery stream.