**AWS Certified Data Analytics** - **Specialty (DAS-C01) Practice Questions**

**Requirement**: Create & Share 20 DAS-C01 practice questions.

**Important Note**: The practice questions should appropriately belong to Data Analytics Specialty in terms of exam objectives & difficulty level.

**Delivery Timeline**: May-17

**Question Response Types**

There are two types of questions:

● Multiple Choice Single Response – **1** correct answer **3** incorrect responses (distractors).

● Multiple Choice Multiple Response – **2** or more correct answers out of **5** or more options.

**Important Note**

● Do write Question Number for quick identification. Q# 1, Q# 2 …. & so on.

● Please mention Domain (based on Data Analytics exam blueprint), Topic & Sub-Topic (If Applicable) with every question.

● Note that we’re expecting standard scenario based questions & NOT straight-forward definition kind of questions.

● The options should not have any obviously incorrect option. We need to word the options so that all of them should appear correct for the students, but a subtle point should mark the correct answer without any ambiguity. So, one answer should be the best choice without any doubt.

● The answer / explanation section should contain explanations on why the answer is correct and others are incorrect. It should also contain the relevant resource link (for details) preferably from AWS documentation.

○ Example

■ Option A is incorrect because..

■ Option B is CORRECT because…

■ Option C is incorrect because..

■ Option D is incorrect because..

● Try to balance the domains based on weightage % defined in the exam blueprint.

● Any AWS service or feature must be approximately 6 months old to figure out in Practice Tests. Put a note in the explanation for any latest service or feature that might be on the borderline of appearing in the real exam.

● **Plagiarism** in any form - Question or in Explanation will be **rejected.** Questions & Explanations should reflect your own professional experience & AWS skills. Author’s who indulge in plagiarism will be **blacklisted** & dropped from our author’s list.

● The ownership of the questions once approved & published on Whizlabs LMS platform, lies solely with Whizlabs Software Pvt. Ltd. You can’t share or publish it elsewhere in any circumstances.

**Sample Format of** **Questions**

----------------------------------------------------------------------------------------------------------------------------

**Question​ ​:​** #

**Main​ ​Topic​ ​:​** < >

**Sub​ ​Topic​ ​:​** [optional]

**Domain:** < >

**Question text**:

<Scenario based. Should be clear in terms of requirements. No ambiguity. No duplicate options. In case of multiple answers, at the end, you should include number of expected answers. e.g. (Select TWO) or (Select THREE) etc. For single answers this is NOT required>

**A)** Option A...

**B)** Option B...

**C)** Option C...

**D)** Option D...

**Answer:** A and C

**Explanation:**

**Option A is CORRECT because...**

**Option B is incorrect because...**

**Option C is CORRECT because...**

**Option D is incorrect because...**

[Insert the explanation in clear and lucid language here.]

**Diagram:** [Optional] [Insert the architectural or conceptual diagram here.]

**Reference:** [Insert the references here - which may include links to AWS Documentation, Blog, re:Invent video, Authority YouTube video].

----------------------------------------------------------------------------------------------------------------------------

**Data Analytics** **Specialty has 5 Domains**

|  |  |  |  |
| --- | --- | --- | --- |
| **S. No.** | **Name of the Domain** | **Weight** | **Estimated No. of Questions**  (Out of 65 as per weightage %) |
| 1 | Collection | 18% | 12 |
| 2 | Storage and Data Management | 22% | 14 |
| 3 | Processing | 24% | 15 |
| 4 | Analysis and Visualization | 18% | 12 |
| 5 | Security | 18% | 12 |

--------------------------------------Question Section Starts-----------------------------------------------------

Question: 1

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine the operational characteristics of a storage solution for analytics**

**Domain:** Collection

**Question text**:

You are a data scientist working for a healthcare services company that produces a Software as a Service (SaaS) offering to its industry affiliates. Your current architecture is built upon an aging infrastructure that is all run on-premises using local physical servers. Your SaaS solution provides data analytics and visualization services for your partner healthcare firms. These firms send you healthcare data for consolidation. Your service ingests the data from these disparate sources, transforms the data, processes it, and finally produces visualizations for consumption by your partners.

Unfortunately, your current solution, depicted in the diagram below, is prone to failures and is not easily scaled. You also have a difficult optimizing the costs of your SaaS offering because you are using sunk costs (the physical servers). You have been tasked with modernizing your analytics solution.

[](https://www.draw.io/?page-id=B8PfgTc1ObyOHXG35zZY&scale=auto#G1dfCJuuEcrWGeKUHvVPOuYyu-bpN3l92R)

Which option provides the most cost effective, performant, and scalable solution?

1. Healthcare data sources -> Kinesis Data Firehose -> S3 -> EMR running Hive -> S3 Data Lake -> Redshift -> Athena
2. Healthcare data sources -> Kinesis Data Firehose -> S3 -> EMR running Spark -> S3 Data Lake -> Redshift -> QuickSight
3. Healthcare data sources -> Kinesis Data Streams -> S3 -> EMR running Hive -> S3 Data Lake -> Redshift -> QuickSight
4. Healthcare data sources -> Kinesis Data Streams -> S3 -> EMR running Spark -> S3 Data Lake -> Redshift -> Athena

**Answer:** B

**Explanation:**

Option A is incorrect. Apache Hive is used to manage Hive clusters, but it wouldn’t give you the capability to transform your data in preparation for integration with Redshift. Also, using Athena to produce visualizations will be less cost effective than using QuickSight because Athena will require more development work on your part.

Option B is correct. Using Kinesis Data Firehose to collect your healthcare data, Spark running on your EMR cluster to transform your data, and Redshift to source your QuickSight visualizations is the most cost effective, scalable, and performant option.

Option C is incorrect. Apache Hive is used to manage Hive clusters, but it wouldn’t give you the capability to transform your data in preparation for integration with Redshift. Also, Kinesis Data Streams will require you to write producer and consumer applications to complete your data collection pipeline. This will be less cost effective than using Kinesis Data Firehose.

Option D is incorrect. Kinesis Data Streams will require you to write producer and consumer applications to complete your data collection pipeline. This will be less cost effective than using Kinesis Data Firehose. Also, using Athena to produce visualizations will be less cost effective than using QuickSight because Athena will require more development work on your part.

**Reference:**

Please see the **Amazon EMR overview page** (<https://aws.amazon.com/emr/>), the Amazon EMR features page titled **Apache Hive on AMazon EMR** (<https://aws.amazon.com/emr/features/hive/>), the article titled **Tutorial: Using Redshift and Amazon QuickSight to deliver business analytics** (<https://www.stitchdata.com/blog/tutorial-using-redshift-and-amazon-quicksight-to-deliver-business-analytics/>)

Question: 2

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Select a collection system that handles the frequency, volume, and source of data**

**Domain:** Collection

**Question text**:

You are a data scientist working for a securities trading firm that receives trading data from multiple market data producer sources. Your task is to consume the data from these producers cost effectively while also maximizing the performance of your data collection system. Your data collection system must deliver the aggregated producer data to your firm’s data lake for analytics application use.

You have built a collection system as follows: a Kinesis Producer Library application that writes records to your Kinesis Data Streams stream shards. You have configured your Kinesis Producer constructor as follows:

RecordMaxBufferedTime = 200

MaxConnections = 2

RequestTimeout = 5000

Region = us-east-1

Your Kinesis Data Streams writes to Kinesis Data Firehose. Kinesis Data Firehose uses a Lambda function to transform your data into the Avro format before writing it to your S3 bucket in your data lake.

You have noticed that your data collection pipeline is not performing as well as you had expected. What may be the cause, and what can you do to improve the situation?

1. Your RecordMaxBufferedTime value is too low, resulting in lower aggregation efficiency, so your pipeline throughput is slow. Dynamically change the RecordMaxBufferedTime to 3000. This allows the KinesisProducer to deliver larger aggregate packages to your Kinesis Data Stream.
2. Your RecordMaxBufferedTime value is too high, resulting in lower aggregation efficiency, so your pipeline throughput is slow. Change the RecordMaxBufferedTime to 50 and restart your KPL application. This allows the KinesisProducer to deliver larger aggregate packages to your Kinesis Data Stream.
3. Your RecordMaxBufferedTime value is too low, resulting in lower aggregation efficiency, so your pipeline throughput is slow. Change the RecordMaxBufferedTime to 3000 and restart your KPL application. This allows the KinesisProducer to deliver larger aggregate packages to your Kinesis Data Stream.
4. Your RequestTimeout value is too high, resulting in lower aggregation efficiency, so your pipeline throughput is slow. Dynamically change the RequestTimeout to 2000. This allows the KinesisProducer to deliver larger aggregate packages to your Kinesis Data Stream.

**Answer:** C

**Explanation:**

Option A is incorrect. Changing your RecordMaxBufferedTime to a higher value will increase your aggregate package size, thereby improving the performance of your pipeline throughput. However, you cannot dynamically change your KinesisProducer once it has been initialized without restarting your KPL application.

Option B is incorrect. Lowering your RecordMaxBufferedTime will actually result in a smaller aggregate package size. This will make your KPL producer to Kinesis Data Streams pipe even slower.

Option C is correct. Changing your RecordMaxBufferedTime to a higher value will increase your aggregate package size, thereby improving the performance of your pipeline throughput. Also, you must restart your KPL application if you want to change any of the KinesisProducerConfiguration values.

Option D is incorrect. Changing the RequestTimeout will not improve the performance of your pipeline throughput. It will allow your KPL application to wait longer before failing when trying to write to your Kinesis Data Streams stream.

**Reference:**

Please see the Amazon Kinesis Data Streams developer guide titled **Developing Producers Using the Amazon Kinesis Producer Library** (<https://docs.aws.amazon.com/streams/latest/dev/developing-producers-with-kpl.html>), the Amazon Kinesis Data Streams developer guide titled **Configuring the Kinesis Producer Library** (<https://docs.aws.amazon.com/streams/latest/dev/kinesis-kpl-config.html>), the Amazon Kinesis Data Streams developer guide titled **Using the KPL with Kinesis Data Firehose** (<https://docs.aws.amazon.com/streams/latest/dev/kpl-with-firehose.html>), and the Amazon Kinesis Data Firehose developer guide titled **Writing to Kinesis Data Firehose Using Kinesis Data Streams** (<https://docs.aws.amazon.com/firehose/latest/dev/writing-with-kinesis-streams.html>)

Question: 3

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Select a collection system that addresses the key properties of data, such as order, format, and compression**

**Domain:** Collection

**Question text**:

You are a data scientist working for a transportation company that specializes in delivering cargo to manufacturing companies. You have been tasked with building a data collection system to gather all of your logistics data into a data lake. This data will be used by analytics applications to perform operations management tasks such as solving the “traveling salesman” problem, where your analytics application needs to find the optimal path for your delivery truck to take to all of its destinations. This optimal path needs to maximize cost efficiency as well as meeting delivery timelines.

You have constructed a Kinesis Data Streams infrastructure with KPL producer applications delivering the transportation data into your Kinesis shards. You are in the process of building your Kinesis Consumer Library application code to consume the streaming data from Kinesis and write the data to your S3 buckets. What happens when your KCL worker code fails in the middle of retrieving a record from your Kinesis stream?

1. Your KCL implementation takes advantage of checkpointing, where KCL stores a cursor on S3 to keep track of records that you have read from a shard. To recover from a failed KCL read, a new KCL worker uses the cursor to restart from the record so you don’t lose the record from the shard.
2. Your KCL implementation takes advantage of checkpointing, where KCL stores a cursor in DynamoDB to keep track of records that you have read from a shard. To recover from a failed KCL read, a new KCL worker uses the cursor to restart from the record so you don’t lose the record from the shard.
3. Your KCL worker will throw an exception from its processRecords call. You need to handle the exception in your KCL application code.
4. Your KCL worker shuts down with shutdown reason ZOMBIE and the KCL throws a ShutdownException.

**Answer:** B

**Explanation:**

Option A is incorrect. Your KCL worker does take advantage of checkpointing, but it persists its checkpoint cursor data to DynamoDB, not S3.

Option B is correct. Your KCL worker takes advantage of checkpointing, persisting its checkpoint cursor data to DynamoDB. The KCL will use the cursor information to restart at the exact record where the previous worker failed.

Option C is incorrect. A processRecords exception results in records being skipped by your consumer code. If your worker fails in the middle of retrieving a record from the Kinesis shard, the KCL uses its cursor stored in DynamoDB to recover from that failed record.

Option D is incorrect. The ZOMBIE reason code is returned by the KCL when a worker has its shards taken by another worker. This scenario describes the situation where a worker fails in the middle of reading a record from your Kinesis shard, which is a different failure scenario.

**Reference:**

Please see the AWS Big Data blog titled **Persist Streaming Data to Amazon S3 using Amazon Kinesis Firehose and AWS Lambda** (<https://aws.amazon.com/blogs/big-data/persist-streaming-data-to-amazon-s3-using-amazon-kinesis-firehose-and-aws-lambda/>), the Amazon Kinesis Data Streams developer guide titled **Troubleshooting Kinesis Data Streams Consumers** (<https://docs.aws.amazon.com/streams/latest/dev/troubleshooting-consumers.html>), and the Amazon Kinesis Data Streams getting started guide titled **Getting started with Amazon Kinesis Data Streams** (<https://aws.amazon.com/kinesis/data-streams/getting-started/>)

Question: 4

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Select a collection system that addresses the key properties of data, such as order, format, and compression**

**Domain:** Collection

**Question text**:

You are a data scientist working for a sports media publishing company. Your company owns a series of sports websites that give sports fans access to data about their sport such as game schedules, player statistics, game results, live streaming game data, etc.

You have been tasked with collecting the sports data from your 100+ websites in real-time, process it, and then store it in your S3 data lake. You also need to store the raw data you receive in a separate S3 prefix so that the raw data can be used by other analytics applications in the future.

Which of the following options is the most cost effective and simplest to implement that addresses your requirements?

1. Recieve sports site data into Kinesis Data Streams. Simultaneously read the Kinesis Data Stream into your EMR cluster running Spark AND your KCL/KCL-C worker application. The EMR Spark instance transforms the raw data and writes it to your S3 bucket processed prefix. The KCL/KCL-C application writes the raw data to the raw prefix on your S3 bucket.
2. Recieve sports site data into Kinesis Data Streams. Simultaneously read the Kinesis Data Stream into your EMR cluster using Kinesis Data Analytics AND your Kinesis Data Firehose stream. The EMR Kinesis Data Analytics instance transforms the raw data and writes it to your S3 bucket processed prefix. The Kinesis Data Firehose leverages a lambda function you write to aggregate the raw data and then write the raw data to the raw prefix on your S3 bucket.
3. Recieve sports site data into Kinesis Data Streams. Simultaneously read from Kinesis Data Streams into your EMR cluster running Spark AND your KPL/KPL-C worker application. The EMR Spark instance transforms the raw data and writes it to your S3 bucket processed prefix. The KPL/KPL-C application writes the raw data to the raw prefix on your S3 bucket.
4. Recieve sports site data into Kinesis Data Streams. Simultaneously read the Kinesis Data Stream into your EMR cluster running Spark AND your Kinesis Data Firehose stream. The EMR Spark instance transforms the raw data and writes it to your S3 bucket processed prefix. The Kinesis Data Firehose leverages a lambda function you write to aggregate the raw data and then write the raw data to the raw prefix on your S3 bucket.

**Answer:** D

**Explanation:**

Option A is incorrect. While this option technically works, using the KCL/KCL-C application code requires more coding effort than using Kinesis Data Firehose and Lambda. Also, with this option you have to provision, maintain, and pay for EC2 servers to run your KCL/KCL-C application.

Option B is incorrect. This option is technically incorrect. You cannot use Kinesis Data Analytics on an EMR cluster to perform data transformation.

Option C is incorrect. While you can use the Kinesis Producer Library (KPL) to produce data to place into your Kinesis Data Stream, it will not work as a consumer as depicted in this option. Also, there is no KPL-C producer library.

Option D is correct. The Kinesis Data Firehose with Lambda function transformation requires less coding work when compared to writing the KCL/KCL-C worker. Also, Kinesis Data Firehose and Lambda are serverless, therefore they require no EC2 provisioning, maintenance, or cost.

**Reference:**

Please see the AWS Big Data blog titled **Persist Streaming Data to Amazon S3 using Amazon Kinesis Firehose and AWS Lambda** (<https://aws.amazon.com/blogs/big-data/persist-streaming-data-to-amazon-s3-using-amazon-kinesis-firehose-and-aws-lambda/>), the Amazon Kinesis Data Streams getting started guide titled **Getting started with Amazon Kinesis Data Streams** (<https://aws.amazon.com/kinesis/data-streams/getting-started/>), and the Amazon EMR release guide titled **Kinesis** (<https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-kinesis.html>)

Question: 5

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine the operational characteristics of a storage solution for analytics**

**Domain:** Storage and Data Management

**Question text**:

You recently started working as a data scientist for a large real estate company. Your real estate brokers need near real-time streaming data on interest rates and loan offerings for their regional markets. They also need near real-time streaming data describing their regional real estate inventory, for example on the market, sold, pending sale, etc.

You have constructed a Kinesis Data Firehose data collection pipeline to gather the data. You now wish to store the data in a DynamoDB database for access via REST APIs by your real estate agents out in the field using their mobile devices. You have implemented the REST APIs using API Gateway.

When you run your first canary deployment of your Lambda function you notice that your Lambda function attempts to process your buffered Kinesis records 3 times and then skips the batch of records. What might be the cause of the problem, and how can you correct the issue?

1. Your Kinesis Firehose buffer interval is set to 60 seconds. This setting is too high, causing your buffer to grow too large for your Lambda invocation and resulting in an invocation limit error. Lower your Kinesis Firehose buffer interval.
2. Your Kinesis Firehose buffer size is set to 1 MB. This setting is too low, causing your Kinesis Firehose to invoke your Lambda function in asynchronous invocation mode. Increase your Kinesis Firehose buffer size.
3. Your Kinesis Firehose buffer size is set to 7 MB. This setting is too high, causing your Lambda function to fail with an invocation limit error. Lower your Kinesis Firehose buffer size.
4. Your Lambda function is returning a response payload of 3 MB. This payload is too large, causing your Lambda function to fail with an invocation limit error. Lower the size of your Lambda function response payload.

**Answer:** C

**Explanation:**

Option A is incorrect. The buffer interval range for Kinesis Data Firehose is 60 to 900 seconds. Therefore, you cannot lower your buffer interval.

Option B is incorrect. Kinesis Firehose invokes your Lambda function in synchronous mode, not asynchronous mode.

Option C is correct. Lambda has an invocation payload limit of 6 MB for synchronous invocations. Kinesis Firehose invokes your Lambda function in synchronous invocation mode. This type of data transformation failure results in three tries before skipping the batch of records. Lowering your Kinesis Firehose buffer size to a value 6 MB or less will solve the issue.

Option D is incorrect. The Lambda function response payload limit is 6 MB. Therefore, a response payload of 3 MB would not cause your Lambda function to fail with an invocation limit error.

**Reference:**

Please see the AWS Database blog titled **Query your AWS database from your serverless application** (<https://aws.amazon.com/blogs/database/query-your-aws-database-from-your-serverless-application/>), the AWS Lambda developer guide titled **Invoking AWS Lambda functions** (<https://docs.aws.amazon.com/lambda/latest/dg/lambda-invocation.html>), the AWS Architecture blog titled **Understanding the Different Ways to Invoke Lambda Functions** (<https://aws.amazon.com/blogs/architecture/understanding-the-different-ways-to-invoke-lambda-functions/>), the Amazon API Gateway developer guide titled **Creating a REST API in Amazon API Gateway** (<https://docs.aws.amazon.com/apigateway/latest/developerguide/how-to-create-api.html>), the Amazon Kinesis Data Firehose developer guide titled **Amazon Kinesis Data Firehose Data Transformation** (<https://docs.aws.amazon.com/firehose/latest/dev/data-transformation.html>), the AWS Lambda developer guide titled **AWS Lambda limits** (<https://docs.aws.amazon.com/lambda/latest/dg/gettingstarted-limits.html>), the Amazon Kinesis Data Firehose developer guide titled **Configure Settings** (<https://docs.aws.amazon.com/firehose/latest/dev/create-configure.html>), and the **Amazon Kinesis Data Firehose FAQs** (<https://aws.amazon.com/kinesis/data-firehose/faqs/>)

Question: 6

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine data access and retrieval patterns**

**Domain:** Storage and Data Management

**Question text**:

You have just received additional venture capital funding for your start-up software firm. With this new funding you are going to try to expand your software service into other regions around the globe. Specifically, you current;ly offer your product in the us-east-1 and us-west-1 AWS regions. You are now going to add the eu-west-1 region to your software distribution.

With this latest version of your software you have decided to use DynamoDB global tables to make your application data available across your distribution as a multi-master, multi-region database. You have transactional operations within your application code base that require atomicity, consistency, isolation, and durability (ACID) capabilities.

You have deployed your DynamoDB global tables in your three regions but you are now noticing that when you issue a TransactWriteItems operation in one region you see partially completed transactions for a period of time in the other two regions. Why is this, and what can you do to remediate the situation?

1. DynamoDB global tables support transactional (ACID) TransactWriteItems operations in only one region, the region where the transaction originated. The changes in the source region will be replicated to the other two regions as the transaction is executed in the source region. Therefore, occasionally the replication regions may not receive the changes until the transaction is committed in the source region. To remediate use autoscaling on your global table.
2. DynamoDB global tables support transactional (ACID) TransactWriteItems operations in only one region, the region where the transaction originated. The changes in the source region will be replicated to the other two regions once transaction is executed in the source region. Therefore, the replication regions will not receive the changes until the transaction is committed in the source region. Since this is the intended design of global tables, there is no remediation for this scenario.
3. DynamoDB global tables support transactional (ACID) TransactWriteItems operations in only one region, the region where the transaction originated. The changes in the source region will be replicated to the other two regions once transaction is executed in the source region. Therefore, the replication regions will not receive the changes until the transaction is committed in the source region. To remediate use on-demand capacity or enable autoscaling on the table.
4. DynamoDB global tables support transactional (ACID) TransactWriteItems operations in only one region, the region where the transaction originated. The changes in the source region will be replicated to the other two regions once transaction is executed in the source region. Therefore, the replication regions will not receive the changes until the transaction is committed in the source region. To remediate associate the AWSServiceRoleForDynamoDBReplication service role to your global table.

**Answer:** B

**Explanation:**

Option A is incorrect. The changes in the source region will be replicated to the other two regions only when the transaction is executed in the source region. Also, autoscaling will have no effect on ACID transaction replication.

Option B is correct. ACID transactions are replicated from the source region to the replica regions only after the source region change is committed. This is the intended design of DynamoDB global changes, primarily to ensure ACID consistency across all global table replicas.

Option C is incorrect. On-demand capacity changes will have no effect on ACID transaction replication.

Option D is incorrect. The AWSServiceRoleForDynamoDBReplication service role is automatically associated with your global tables when you create the tables. The role gives DynamoDB the permission to manage cross-Region replication for you.

**Reference:**

Please see the Amazon DynamoDB developer guide titled **Global Tables: Multi-Region Replication with DynamoDB** (<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/GlobalTables.html>), the Wikipedia page titled **ACID** (<https://en.wikipedia.org/wiki/ACID>), and the Amazon DynamoDB developer guide titled **Using IAM with Global Tables** (<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/V2gt_IAM.html>)

Question: 7

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine data access and retrieval patterns**

**Domain:** Storage and Data Management

**Question text**:

You work as a data scientist for an ocean cruise ship resort company. You have been tasked with building an S3 data lake to store information about customer interaction and satisfaction with the company’s resort offerings. The data will be captured from social media and the firm’s website.

Your data collection system will need to stream the social media and web site comments in real-time to your data store. Your management team wishes to use the data in the data store to perform ad-hoc analysis of the customer feedback in real-time. Which option gives you the most cost efficient and performant solution?

1. Streaming customer data -> Kinesis Data Streams -> Kinesis Firehose -> Redshift -> Athena
2. Streaming customer data -> Kinesis Firehose -> Redshift -> Athena
3. Streaming customer data -> AWS IoT Core -> Kinesis Firehose -> S3 -> Athena
4. Streaming customer data -> AWS IoT Core -> AWS Glue -> S3 -> Athena

**Answer:** D

**Explanation:**

Option A is incorrect. While you could receive the web and social media streamed data into your Kinesis Data Streams, stream it through your Kinesis Firehose, and copy it into Redshift, Athena can not query Redshift. Redshift is an alternative to Athena. You use Athena to query S3 directly.

Option B is incorrect. You could stream your web and social media data to Kinesis Firehose and then copy it into Redshift. However, Athena can not query Redshift. Redshift is an alternative to Athena. You use Athena to query S3 directly.

Option C is incorrect. AWS IoT Core can be used to receive social media and web streamed traffic, but Kinesis Firehose is extraneous in this scenario. You can use IoT Core rule actions to write your streamed data directly to S3.

Option D is correct. You can receive your web and social media streamed data into AWS IoT Core. Then write the messages directly to S3 using the S3 IoT Core rule action. You can use Glue to crawl and catalog your data so that you can easily query it from Athena.

**Reference:**

Please see the AWS IoT developer guide titled **AWS IoT rule actions** (<https://docs.aws.amazon.com/iot/latest/developerguide/iot-rule-actions.html>), the Amazon Athena user guide titled **Integration with AWS Glue** (<https://docs.aws.amazon.com/athena/latest/ug/glue-athena.html>), and the AWS Big Data blog titled **Analyzing Data in S3 using Amazon Athena** (<https://aws.amazon.com/blogs/big-data/analyzing-data-in-s3-using-amazon-athena/>)

Question: 8

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Define a data lifecycle based on usage patterns and business requirements**

**Domain:** Storage and Data Management

**Question text**:

You work as a data scientist for a financial services firm that trades commodities on the futures markets in the United States. Specifically, your traders trade the S&P 500, Nasdaq-100, Yen, and Bitcoin equity index products on the Chicago Mercantile Exchange (CME). In order to have the real-time information needed to make informed trades, your traders need futures market data streamed in real-time into their data repository. They need to use their machine learning models to perform predictive analytics on their data.

Which option meets your requirements and gives you the most cost efficient solution to your design problem?

1. Streaming futures market data -> Kinesis Data Streams -> Kinesis Firehose -> Redshift -> SageMaker
2. Streaming futures market data -> Kinesis Firehose -> Redshift -> SageMaker
3. Streaming futures market data -> Kinesis Firehose -> S3 -> SageMaker
4. Streaming futures market data -> Kinesis Firehose -> Redshift -> EMR -> SageMaker

**Answer:** C

**Explanation:**

Option A is incorrect. You can stream your data through a pipeline of Kinesis Data Streams to Kinesis Firehose and then copy the data into Redshift. However, you cannot source your SageMaker models directly from Redshift.

Option B is incorrect. You can stream your data through a pipeline using Kinesis Firehose and then copy the data into Redshift. However, you cannot source your SageMaker models directly from Redshift.

Option C is correct. Stream your futures market data using Kinesis Firehose. Firehose writes the data to S3. SageMaker sources its model with the raw data in S3. This is the most efficient option that also meets your requirements.

Option D is incorrect. You can stream your data through a pipeline using Kinesis Firehose and then copy the data into Redshift. You can also extract the data from Redshift into your EMR cluster using python code. Then you can source your SageMaker model from your data stored in the EMR cluster. However, the Redshift component in this option introduces an unnecessary step, which will cost your development time as well as AWS services costs. Therefore, this option is not the most cost effective option.

**Reference:**

Please see the **CME Group website** (<https://www.cmegroup.com/>), the Investopedia Futures and Commodities Trading page titled **Chicago Mercantile Exchange** (<https://www.investopedia.com/terms/c/cme.asp>), the AWS Machine Learning blog titled **Exploring data warehouse tables with machine learning and Amazon SageMaker notebooks** (<https://aws.amazon.com/blogs/machine-learning/exploring-data-warehouse-tables-with-machine-learning-and-amazon-sagemaker-notebooks/>), the AWS Labs GitHub repository titled **working\_with\_redshift\_data** (<https://github.com/awslabs/amazon-sagemaker-examples/blob/master/advanced_functionality/working_with_redshift_data/working_with_redshift_data.ipynb>), and the **Amazon Kinesis Data Firehose overview page** (<https://aws.amazon.com/kinesis/data-firehose/>)

Question: 9

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Select an appropriate data layout, schema, structure, and format**

**Domain:** Storage and Data Management

**Question text**:

You work as a data scientist for a national polling institute. Your institute performs state-wide and national polls in the areas of politics, elections, and general public interest subjects. Your data collection system receives hundreds of thousands of data records through your data streaming pipeline. You have chosen DynamoDB as the data store for several of these data structures.

As you create your DynamoDB table for your political election polling data store you need to select a partition key and a sort key, since you wish to use a composite key to improve performance and DynamoDB capacity management. You have several choices for your political election polling partition/sort key combination. Your researchers need to produce several visualizations of the data to understand the distribution of votes by age, nationality, political party affiliation, selected candidate, etc. An example would be to visualize votes collected for a particular candidate by age group and by voter nationality.

Which option will give you the best performance for your political election polling table?

1. Partition key: registered voter political party, Sort key: selected candidate name
2. Composite Partition key: selected candidate political party+selected candidate name, Sort key: voter age
3. Partition key: registered voter id, Sort key: selected candidate name
4. Composite Partition key: registered voter id+voter age, Sort key: voter age

**Answer:** C

**Explanation:**

Option A is incorrect. The choice of partition key does not have high cardinality; there will be many records in your streaming data with the same voter registered political party from a limited pool of registered political parties.

Option B is incorrect. The choice of the composite partition key does not have high cardinality; there will be many records in your streaming data with the same selected candidate political party+selected candidate name from a limited pool of selected candidate political parties and selected candidate names.

Option C is correct. The choice of registered voter id for your primary key gives you high cardinality; every registered voter will have a unique voter id. Therefore, there is no need for a composite partition key.

Option D is incorrect. The choice of partition key has high cardinality, however, the composite key is not needed. A partition key of the voter id has very high cardinality (each id is unique) so using a composite key in this case over complicates the schema.

**Reference:**

Please see the Statewide Database titled **2018 General Election Precinct Data** (<https://statewidedatabase.org/d10/g18.html>), the Amazon DynamoDB developer guide titled **Best Practices for Designing and Using Partition Keys Effectively** (<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/bp-partition-key-design.html>), the Amazon DynamoDB developer guide titled **Partitions and Data Distribution** (<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/HowItWorks.Partitions.html>), and the AWS Database Blog titled **Choosing the Right DynamoDB Partition Key** (<https://aws.amazon.com/blogs/database/choosing-the-right-dynamodb-partition-key/>)

Question: 10

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine appropriate data processing solution requirements**

**Domain:** Processing

**Question text**:

You work as a data scientist for a data analytics company that specializes in supplying data sets to industry partners for use in their machine learning models. Your company’s data sets are used by your partners as seed data for their own corporate data stores, allowing your partners to leverage a much larger sample of data for their models.

One of your partners needs you to transform industry data that is sourced in the JSON format so that the data can be used by their machine learning model in the CSV format. You have chosen AWS Glue as your transformation tool. One particular requirement is that your ETL script needs to convert a composite JSON format of, for example {“id”: 1435678, “product name”: “product A”, “product cost”: 54.23}, to values in your CSV file of int, string, and double.

Which option leverages AWS Glue to perform the required JSON transformation in the most cost effective optimal manner?

1. Write a PySpark function that uses the RDD API to reformat the JSON into the required elements
2. Write a PySpark function that uses Spark SQL to reformat the JSON into the required elements
3. Use the Glue built-in transform Relationalize to reformat the JSON into the required elements
4. Use the Glue built-in transform Unbox to reformat the JSON into the required elements

**Answer:** D

**Explanation:**

Option A is incorrect. Writing a PySpark function when you can use a built-in Glue transform is not the most cost effective option.

Option B is incorrect. Writing a PySpark function when you can use a built-in Glue transform is not the most cost effective option.

Option C is incorrect. The Glue Relationalize built-in transform would convert your JSON into rows and columns, but it wouldn’t reformat your composite JSON field to separate fields.

Option D is correct. The Glue Unbox built-in transform reformats string fields, like your composite JSON field, into distinct fields that represent the types of the composites.

**Reference:**

Please see the AWS Glue developer guide titled **Authoring Jobs in AWS Glue** (<https://docs.aws.amazon.com/glue/latest/dg/author-job.html>), the AWS Glue developer guide titled **Built-In Transforms** (<https://docs.aws.amazon.com/glue/latest/dg/built-in-transforms.html>), and the Spark SQL guide titled **Spark SQL, DataFrames and Datasets Guide** (<http://spark.apache.org/docs/latest/sql-programming-guide.html#json-datasets>)

Question: 11

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Automate and operationalize a data processing solution**

**Domain:** Processing

**Question text**:

You work as a data scientist working on a data analytics project to construct a data warehouse for your organization’s analytics applications and decision support. This data warehouse needs to be loaded with the petabytes of data from your company’s operational data stores. To achieve this goal, you need to load petabytes of data from your EMR cluster to your newly created Redshift cluster.

Which of the following steps are NOT required when moving your data from your EMR cluster to your Redshift cluster? (SELECT TWO)

1. Configure IAM permissions
2. Get the Redshift cluster public key and node IP addresses
3. Add the Redshift cluster public key to each of your EMR EC2 host's authorized keys file
4. Create a manifest file
5. Copy the manifest file to an EMR Hadoop Distributed File System (HDFS)
6. Configure the EMR hosts to accept all of the Redshift cluster's IP addresses
7. Use the Redshift COPY command to load the data

**Answers:** D, E

**Explanation:**

Option A is incorrect. You need to configure the IAM permissions needed to allow for running the COPY command.

Option B is incorrect. You need the Redshift cluster public key to enable the Redshift cluster nodes to SSH to the EMR cluster hosts. You need the IP addresses of the Redshift cluster nodes to change the EMR host’s security groups to allow access from your Redshift cluster.

Option C is incorrect. You need to add the Redshift cluster public key to the EMR cluster host's authorized keys file to allow the Redshift cluster to connect via SSH.

Option D is correct. A manifest file is needed when you are loading data from a remote host, not from an EMR cluster in your AWS account.

Option E is correct. A manifest file is needed when you are loading data from a remote host, not from an EMR cluster in your AWS account.

Option F is incorrect. You need to update the EMR cluster instance's security groups to have ingress rules to allow access from the Redshift IP addresses.

Option G is incorrect. You run the COPY command to load the data into your Redshift table.

**Reference:**

Please see the Amazon Redshift developer guide titled **Loading Data** (<https://docs.aws.amazon.com/redshift/latest/dg/t_Loading_data.html>), the Amazon Redshift developer guide titled **Using a COPY command to load data** (<https://docs.aws.amazon.com/redshift/latest/dg/t_Loading_tables_with_the_COPY_command.html>), the Amazon Redshift developer guide titled **Loading data from Amazon EMR** (<https://docs.aws.amazon.com/redshift/latest/dg/loading-data-from-emr.html>), and the Amazon Redshift developer guide titled **Loading data from remote hosts** (<https://docs.aws.amazon.com/redshift/latest/dg/loading-data-from-remote-hosts.html>)

Question: 12

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Design a solution for transforming and preparing data for analysis**

**Domain:** Processing

**Question text**:

You work as a data analytics specialist for a retail clothing chain. Your retail stores generate sales and inventory data that is streamed in real-time into your data collection system. Your data scientists use this data to run predictive analytics applications using machine learning models based on the DeepAR SageMaker built-in algorithm. You also load your data into a Redshift cluster for use by your data scientists for analytics applications and business intelligence Key Performance Indicators (KPIs).

You have loaded your initial retail store data into your Redshift cluster using the COPY command to copy the data from your S3 data lake. For the last several weeks you have been streaming the retail store data into your Redshift tables. You are now noticing that your queries from your Redshift tables are taking longer to execute. Which of the following options will correct your performance problem?

1. Run the ANALYZE then VACUUM commands on your Redshift cluster once an hour, every hour to clean up deleted rows, sort new data inserted in the last hour, and update metadata.
2. Run the ANALYZE then VACUUM commands on your Redshift cluster once a day, every day to clean up deleted rows, sort new data inserted in the last hour, and update metadata.
3. Run the VACUUM then ANALYZE commands on your Redshift cluster once a day, every day to clean up deleted rows, sort new data inserted in the last hour, and update metadata.
4. Run the VACUUM then ANALYZE commands on your Redshift cluster once a week, every week to clean up deleted rows, sort new data inserted in the last hour, and update metadata.

**Answer:** C

**Explanation:**

Option A is incorrect. The best practice is to run the VACUUM command first, then run the ANALYZE command. Also, running the VACUUM and ANALYZE commands every hour is probably too often and will impact the overall performance of your Redshift cluster.

Option B is incorrect. The best practice is to run the VACUUM command first, then run the ANALYZE command.

Option C is correct. The best practice is to run the VACUUM command first, then run the ANALYZE command. Also, running the VACUUM and ANALYZE commands once per day is the best practice as far as frequency of restructuring your Redshift cluster tables without impacting overall performance by running the commands too frequently.

Option D is incorrect. The best practice is to run the VACUUM command first, then run the ANALYZE command. However, only running the VACUUM and ANALYZE commands once per week is probably not frequent enough. Running the VACUUM and ANALYZE commands once per day is the best practice as far as frequency of restructuring your Redshift cluster tables without impacting overall performance by running the commands too frequently.

**Reference:**

Please see the Amazon Redshift developer guide titled **Loading Data** (<https://docs.aws.amazon.com/redshift/latest/dg/t_Loading_data.html>), the Amazon SageMaker developer guide titled **DeepAR Forecasting Algorithm** (<https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html>), the Investopedia page titled **Key Performance Indicators (KPIs)** (<https://www.investopedia.com/terms/k/kpi.asp>), the Tech Blog titled **AWS Redshift best practices, tips and tricks - part 2** (<https://source.coveo.com/2017/09/15/redshift-best-practices2/>), the Amazon Redshift developer guide titled **Vacuuming tables** (<https://docs.aws.amazon.com/redshift/latest/dg/t_Reclaiming_storage_space202.html>), and the Amazon Redshift developer guide titled **VACUUM** (<https://docs.aws.amazon.com/redshift/latest/dg/r_VACUUM_command.html>)

Question: 13

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Design a solution for transforming and preparing data for analysis**

**Domain:** Processing

**Question text**:

You work as a data analytics specialist for a social media software company. Your product generates data that your company can use in predictive analytics applications that leverage machine learning. These applications use Natural Language Processing (NLP) and click prediction techniques for use in targeted advertising on your social media app.

You need to build an EMR cluster to store and process this streaming data to prepare it for use in your machine learning analytics applications. Based on your streaming data activity volume you estimate that your cluster will need to have more than 50 nodes.

Based on your streaming data volume and your machine learning based use cases, which types of EC2 instances should you use for your master node and core/task nodes?

1. Master node: m4.xlarge. Core/Task: Cluster Compute instance type
2. Master node: m5.xlarge. Core/Task: High CPU instance type
3. Master node: m5.xlarge. Core/Task: High Memory instance type
4. Master node: m4.xlarge. Core/Task: High CPU instance type

**Answer:** A

**Explanation:**

Option A is correct. The best practice is to use an m4.xlarge instance type for your master node if your cluster will have more than 50 nodes. Also, for NLP and machine learning applications the Cluster Compute instance type is recommended.

Option B is incorrect. The best practice is to use an m4.xlarge instance type for your master node if your cluster will have more than 50 nodes. The m5.xlarge instance type is recommended for your master node if your cluster will have fewer than 50 nodes. Also, the High CPU instance type is recommended for computation-intensive clusters, but your cluster will be used for NLP and machine learning applications for which the Cluster Compute instance type is recommended.

Option C is incorrect. The best practice is to use an m4.xlarge instance type for your master node if your cluster will have more than 50 nodes. The m5.xlarge instance type is recommended for your master node if your cluster will have fewer than 50 nodes. Also, the High Memory instance type is recommended for clusters running database and memory-caching applications, but your cluster will be used for NLP and machine learning applications for which the Cluster Compute instance type is recommended.

Option D is incorrect. The best practice is to use an m4.xlarge instance type for your master node if your cluster will have more than 50 nodes. However, the High CPU instance type is recommended for computation-intensive clusters, but your cluster will be used for NLP and machine learning applications for which the Cluster Compute instance type is recommended.

**Reference:**

Please see the Amazon EMR management guide titled **Cluster Configuration Guidelines and Best Practices** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html>), the Amazon EMR management guide titled **What Is Amazon EMR?** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-what-is-emr.html>), and the Amazon SageMaker developer guide titled **Use Amazon SageMaker built-in algorithms** (<https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html>)

Question: 14

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Automate and operationalize a data processing solution**

**Domain:** Processing

**Question text**:

You work as a data analytics specialist for a television network that has started using data analytics for its sports broadcasts. You receive sports streaming data into your data collection system and store it in your EMR cluster for use in real-time analytics during the broadcast of sporting events. The analytics are overlaid onto the live sports action to give detailed insight into the action. The analytics are also broadcast out via your website for consumption by your millions of users worldwide.

Based on the schedule of sporting events and the popularity of some events, such as the Fédération Internationale de Football Association (FIFA) football world cup, you need to be able to scale your EMR cluster EC2 instances in or out depending on the particular demand for analytics for the given event. Your goal is to provide adequate performance for the given workload while also maintaining the most cost effective environment over time.

Which type of scaling plan should you use for your EMR cluster?

1. Automatically modify your EC2 instance type and resize your core instance group and task instance groups by leveraging automatic scaling to add or remove EC2 instances. Do this by defining rules that Auto Scaling uses based on a CloudWatch metric you specify.
2. Define your EC2 instance type during the initial configuration of your instance groups. Then automatically resize your core instance group and task instance groups by leveraging automatic scaling to add or remove EC2 instances. Do this by defining rules that Auto Scaling uses based on a CloudWatch metric you specify.
3. Reconfigure and resize your instance groups simultaneously as your needs change based on rules you specify that Auto Scaling uses based on a CloudWatch metric you specify.
4. Define your EC2 instance type during the initial configuration of your instance groups. Then automatically resize your master instance group, core instance group, and task instance groups by leveraging automatic scaling to add or remove EC2 instances. Do this by defining rules that Auto Scaling uses based on a CloudWatch metric you specify.

**Answer:** B

**Explanation:**

Option A is incorrect. You cannot automatically modify the EC2 instance type used by your EMR cluster. You can only define the EC2 instance type during the initial creation of your instance group.

Option B is correct. You can only define the EC2 instance type during the initial creation of your instance group. Leveraging automatic scaling to add or remove EC2 instances to your core instance group and task instance groups based on the changes in a CloudWatch metric is the best practice for maintaining the most cost effective and performant EMR cluster.

Option C is incorrect. You cannot reconfigure and resize your instance groups simultaneously. You can perform one or the other independently, but not simultaneously.

Option D is incorrect. You cannot automatically resize your master instance group.

**Reference:**

Please see the Amazon EMR management guide titled **Cluster Configuration Guidelines and Best Practices** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html>), the Amazon EMR management guide titled **What Is Amazon EMR?** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-what-is-emr.html>), and the Amazon EMR management guide titled **Scaling Cluster Resources** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-scale-on-demand.html>)

Question: 15

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine the operational characteristics of an analysis and visualization solution**

**Domain:** Analysis and Visualization

**Question text**:

You work as a data analytics specialist for a gaming software company that produces mobile phone games. You have built a data warehouse in Redshift to house gaming data for analytics apps to help understand the usage patterns of your users. You also have built a streaming data collection system for game player real-time interaction. You have been asked to create analytics ad-hoc visualizations that join data from your Redshift data warehouse with your real-time streaming data stored in S3.

What combination of AWS services and tasks implement your required visualization solution in the most cost effective manner?

1. Use Athena to do a join query of the data in Redshift and the data in S3.
2. Use Redshift Spectrum to create an internal schema for your data stored on S3 using the CREATE INTERNAL SCHEMA command. Use Spectrum SELECT statement to do a join query of the data in Redshift and the data in S3.
3. Use Redshift Spectrum to create an external schema for your data stored on S3 using the CREATE EXTERNAL SCHEMA command. Use Spectrum SELECT statement to do a join query of the data in Redshift and the data in S3.
4. Use Redshift Spectrum to create an external schema and an external table for your data stored on S3 using the CREATE EXTERNAL SCHEMA and CREATE EXTERNAL TABLE commands. Use Spectrum SELECT statement to do a join query of the data in Redshift and the data in S3.

**Answer:** D

**Explanation:**

Option A is incorrect. You should use Redshift Spectrum if you need to join data across S3 and Redshift. In order to use Redshift Spectrum to query your S3 data you first need to create an external schema and an external table. Athena cannot join across S3 and Redshift.

Option B is incorrect. The name of the schema used with Redshift Spectrum is external schema, not internal schema.

Option C is incorrect. You need to create both an external schema and an external table in order to join data across S3 and Redshift.

Option D is correct. You should use Redshift Spectrum if you need to join data across S3 and Redshift. In order to use Redshift Spectrum to query your S3 data you first need to create an external schema and an external table.

**Reference:**

Please see the Amazon Redshift Database developer guide titled **Getting started with Amazon Redshift Spectrum** (<https://docs.aws.amazon.com/redshift/latest/dg/c-getting-started-using-spectrum.html>), the Amazon Redshift Database developer guide titled **Creating external schemas for Amazon Redshift Spectrum** (<https://docs.aws.amazon.com/redshift/latest/dg/c-spectrum-external-schemas.html>), the Amazon Redshift Database developer guide titled **Step 3: Create an external schema and an external table** (<https://docs.aws.amazon.com/redshift/latest/dg/c-getting-started-using-spectrum-create-external-table.html>), the Amazon Redshift Database developer guide titled **Step 4: Query your data in Amazon S3** (<https://docs.aws.amazon.com/redshift/latest/dg/c-getting-started-using-spectrum-query-s3-data.html>), the Amazon Redshift Database developer guide titled **CREATE EXTERNAL SCHEMA** (<https://docs.aws.amazon.com/redshift/latest/dg/r_CREATE_EXTERNAL_SCHEMA.html>), and the **Amazon Athena FAQs page** (<https://aws.amazon.com/athena/faqs/>)

Question: 16

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Select the appropriate data analysis solution for a given scenario**

**Domain:** Analysis and Visualization

**Question text**:

You work as a data scientist for a management consulting company. The management team of your company’s business process improvement practice needs real-time visualizations of Key Performance Improvement (KPI) outliers for their clients. You have a large historical data set and you also have real-time streaming data from your current engagements.

Which option gives you the most cost effective solution to your data analysis visualization problem?

1. Train a model based on the SageMaker built-in Random Cut Forest algorithm to detect the outliers in your clients KPI data.
2. Train a model based on the SageMaker built-in Principal Component Analysis algorithm to detect the outliers in your clients KPI data.
3. Use an anomaly detection insight in QuickSight to detect the outliers in your clients KPI data.
4. Use the QuickSight sheet capability to load your KPI data and then visualize it in your QuickSight Storyboard.

**Answer:** C

**Explanation:**

Option A is incorrect. You could build a Random Cut Forest based model, train it using your historical data, and then gather inferences using your streaming data, but this solution will be much more expensive than using the ML Insight feature of QuickSight.

Option B is incorrect. The Principal Component Analysis algorithm is used to reduce the complexity of your feature set in your machine learning data set. It would not be a good fit for anomaly detection of KPI data. Also, this solution will be much more expensive than using the ML Insight feature of QuickSight.

Option C is correct. The anomaly detection insight of QuickSight allows you to continually analyze your KPI data to find anomalies. You can then visualize your insight data using the insight widget in QuickSight. This option is far more cost effective than building a SageMaker machine learning model.

Option D is incorrect. The QuickSight sheets capability allows you to create sets of visualizations and show them on the same page. It doesn’t assist with anomaly detection.

**Reference:**

Please see the Amazon QuickSight user guide titled **Working with ML Insights** (<https://docs.aws.amazon.com/quicksight/latest/user/making-data-driven-decisions-with-ml-in-quicksight.html>), the Amazon QuickSight user guide titled **Detecting Outliers with ML-Powered Anomaly Detection** (<https://docs.aws.amazon.com/quicksight/latest/user/anomaly-detection.html>), the Amazon QuickSight user guide titled **Concepts for Anomaly or Outlier Detection** (<https://docs.aws.amazon.com/quicksight/latest/user/anomaly-detection-outliers-and-key-drivers.html>), the Amazon QuickSight user guide titled **Setting Up ML-Powered Anomaly Detection for Outlier Analysis** (<https://docs.aws.amazon.com/quicksight/latest/user/anomaly-detection-using.html>), the Amazon QuickSight user guide titled **Working with Multiple Sheets in an Amazon QuickSight Analysis** (<https://docs.aws.amazon.com/quicksight/latest/user/working-with-multiple-sheets.html>) and the Amazon SageMaker developer guide titled **Use Amazon SageMaker built-in algorithms** (<https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html>)

Question: 17

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Select the appropriate data analysis solution for a given scenario**

**Domain:** Analysis and Visualization

**Question text**:

You work as a data scientist for an online lodging company that leverages available space in hosts’ homes and apartments in many regional areas around the world. Your team ingests streamed data from many sources to populate your company’s data warehouse on Redshift. Your management team uses your data warehouse to build enterprise reporting and business intelligence interactive analytics applications.

You also need to provide the capability to perform ad-hoc queries of your accumulated streamed data to your management team. Which option gives you the most cost effective and performant solution for your ad-hoc query needs?

1. Use the Kinesis Consumer Library to write a Kinesis Stream consumer application to provide ad-hoc capabilities.
2. Stream the data through a Lambda function that stores the data in DynamoDB. Perform ad-hoc queries on the DynamoDB tables.
3. Stream the data to S3. Perform ad-hoc queries on the data stored in S3.
4. Stream your data through Kinesis Analytics. Perform ad-hoc queries on the data using Kinesis Analytics.

**Answer:** C

**Explanation:**

Option A is incorrect. This option is technically feasible, however it would involve a considerable amount of coding compared to using Athena. Therefore, this option is not the most cost effective.

Option B is incorrect. This option is technically feasible, however it would involve coding a Lambda function, which will require more development effort compared to using Athena. Therefore, this option is not the most cost effective.

Option C is correct. Staging your streaming data on S3 allows you to take advantage of Athena’s ad-hoc query capability without writing any code. This is the most cost effective and performant option.

Option D is incorrect. Since you are attempting to provide ad-hoc query capability on accumulated data, staging your data on S3 and using Athena is a better choice than using Kinesis Analytics. Kinesis Analytics queries your streaming data as it moves through your data collection pipeline.

**Reference:**

Please see the **Amazon Kinesis Data Analytics overview page** (<https://aws.amazon.com/kinesis/data-analytics/>**),** the AWS Big Data blog titled **Create real-time clickstream sessions and run analytics with Amazon Kinesis Data Analytics, AWS Glue, and Amazon Athena** (<https://aws.amazon.com/blogs/big-data/create-real-time-clickstream-sessions-and-run-analytics-with-amazon-kinesis-data-analytics-aws-glue-and-amazon-athena/>), the Projects on AWS page titled **Build a serverless Real-Time Data Processing App: Aggregate data** (<https://aws.amazon.com/getting-started/projects/build-serverless-real-time-data-processing-app-lambda-kinesis-s3-dynamodb-cognito-athena/2/>), the Projects on AWS page titled **Build a serverless Real-Time Data Processing App: Process streaming data** (<https://aws.amazon.com/getting-started/projects/build-serverless-real-time-data-processing-app-lambda-kinesis-s3-dynamodb-cognito-athena/3/>), the Projects on AWS page titled **Build a serverless Real-Time Data Processing App: Store & Query Data** (<https://aws.amazon.com/getting-started/projects/build-serverless-real-time-data-processing-app-lambda-kinesis-s3-dynamodb-cognito-athena/4/>), and the Overview of Amazon Web Services page titled **Analytics** (<https://docs.aws.amazon.com/whitepapers/latest/aws-overview/analytics.html#amazon-kinesis-analytics>)

Question: 18

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Select the appropriate data visualization solution for a given scenario**

**Domain:** Analysis and Visualization

**Question text**:

You work as a data scientist who produces visualizations for real-time analytics of your company’s Internet of Things (IoT) devices out in the field. Your company manufactures appliances and in-home fixtures (such as refrigerators, faucets, shower heads, etc.) that have sensors built into them. These sensors stream IoT MQTT messages back to your AWS account.

Which set of AWS services allow you to build visualizations of your IoT data in the most expeditious, cost effective manner?

1. IoT Core -> Kinesis Firehose -> S3 <- Athena <- Quicksight
2. Kinesis Data Streams -> Kinesis Firehose -> S3 <- Athena <- Quicksight
3. IoT Core -> Kinesis Streams -> S3 <- Athena <- Quicksight
4. IoT Core -> Kinesis Firehose -> S3 <- Athena

**Answer:** A

**Explanation:**

Option A is correct. You use IoT Core to receive the MQTT messages from your IoT sensors. Kinesis Firehose loads the sensor data directly to S3. Athena is used to run queries against your streamed data in S3 in real-time. QuickSight is used to build sophisticated real-time visualizations. This option uses completely serverless (managed by AWS) services. It requires very little build-out work on your behalf, making it the most expeditious and cost effective option.

Option B is incorrect. To use Kinesis Data Streams to ingest your IoT MQTT messages you would have to write a Kinesis Producer Library (KPL) application and provision an EC2 instance on which to run the KPL application. This option will take much longer to construct and will be much more expensive to run.

Option C is incorrect. To use Kinesis Data Streams to process your IoT MQTT messages from your IoT Core service you would have to write a Kinesis Producer Library (KPL) application and provision an EC2 instance on which to run the KPL application. This option will take much longer to construct and will be much more expensive to run.

Option D is incorrect. This option lacks a visualization service, such as QuickSight. With only Athena, you will have query results but you will have to create a visualization component to show the results of your Athena queries. This option will take much longer to construct and will be much more expensive to run.

**Reference:**

Please see the Amazon What’s New page titled **Amazon QuickSight adds support for Amazon Athena** (<https://aws.amazon.com/about-aws/whats-new/2016/12/amazon-quicksight-adds-support-for-amazon-athena/>), and the AWS Big Data blog titled **Derive Insights from IoT in Minutes using AWS IoT, Amazon Kinesis Firehose, Amazon Athena, and Amazon QuickSight** (<https://aws.amazon.com/blogs/big-data/derive-insights-from-iot-in-minutes-using-aws-iot-amazon-kinesis-firehose-amazon-athena-and-amazon-quicksight/>)

Question: 19

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Select appropriate authentication and authorization mechanisms**

**Domain:** Security

**Question text**:

You work as a data scientist for a large global bank. Your team has been assigned the task of building a data warehouse in Redshift. Your data sources are in S3 buckets in your company's AWS account, with the account number of 0045-2345-2329. The data in your S3 buckets is encrypted using a customer managed key with a key-d of 6sdghd7r-dslf-6dgd-6fdt-ye73judhy7df.

How will you grant permissions to your Glue crawler to access your encrypted S3 buckets?

1. Assign an IAM policy to your Glue crawler with the following permissions:  
   {

"Version":"2012-10-17",

"Statement":[

{

"Effect":"Allow",

"Action":[

"kms:Encrypt"

],

"Resource":[

"arn:aws:kms:\*:0045-2345-2329:key/6sdghd7r-dslf-6dgd-6fdt-ye73judhy7df"

]

}

]

}

1. Assign an IAM role to your Glue crawler with the following policy attached:  
   {

"Version":"2012-10-17",

"Statement":[

{

"Effect":"Allow",

"Action":[

"kms:Decrypt"

],

"Resource":[

"arn:aws:kms:\*:0045-2345-2329:key/6sdghd7r-dslf-6dgd-6fdt-ye73judhy7df"

]

}

]

}

1. Assign an IAM role to your Glue crawler with the following policy attached:  
   {

"Version":"2012-10-17",

"Statement":[

{

"Effect":"Allow",

"Action":[

"kms:Decrypt"

],

"Resource":[

"arn:aws:kms:\*:004523452329:key/6sdghd7r-dslf-6dgd-6fdt-ye73judhy7df"

]

}

]

}

1. Assign an IAM role to your Glue crawler with the following policy attached:  
   {

"Version":"2012-10-17",

"Statement":[

{

"Effect":"Allow",

"Action":[

"kms:Decrypt",

"kms:Encrypt"

],

"Resource":[

"arn:aws:kms:\*:004523452329:key/6sdghd7r-dslf-6dgd-6fdt-ye73judhy7df"

]

}

]

}

**Answer:** C

**Explanation:**

Option A is incorrect. You need to assign an IAM role with an attached policy that allows for decryption of the S3 bucket objects. Also, the account ID in the Resource section of the policy has hyphens in it, for the policy definition you need to enter the account ID without the hyphens.

Option B is incorrect. The account ID in the Resource section of the policy has hyphens in it, for the policy definition you need to enter the account ID without the hyphens.

Option C is correct. The policy attached to the Role assigned to the Glue crawler has the correct allowed action of Decrypt. Also, the Resource has the account ID correctly entered without hyphens.

Option D is incorrect. The policy attached to the Role assigned to the Glue crawler has an extraneous allowed action of Encrypt, since your crawler is only reading the S3 objects, you should use the least privilege approach of only granting Decrypt in the allowed Actions.

**Reference:**

Please see the AWS Glue developer guide titled **Working with Crawlers on the AWS Glue Console** (<https://docs.aws.amazon.com/glue/latest/dg/console-crawlers.html>), the AWS Glue developer guide titled **Step 2: Create an IAM Role for AWS Glue** (<https://docs.aws.amazon.com/glue/latest/dg/create-an-iam-role.html>), and the Amazon Simple Storage Service developer guide titled **Protecting Data Using Server-Side Encryption with CMKs Stored in AWS Key Management Service (SSE-KMS)** (<https://docs.aws.amazon.com/AmazonS3/latest/dev/UsingKMSEncryption.html>)

Question: 20

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Apply data protection and encryption techniques**

**Domain:** Security

**Question text**:

You work as a data scientist for a securities custodian bank that is responsible for holding the asset information for a large number of brokerage and mutual fund firms. Part of your job is to maintain the EMR clusters that are used to build your firm’s data lake. The data lake holds aggregated information from your brokerage and mutual fund client firms. This information is highly regulated and confidential. Therefore, you must configure your EMR clusters to use encryption at rest. You have selected to use AWS Key Management Service for the key provider for your EMR volume encryption. You have set up your EMR security configuration with local disk encryption enabled.

Which option is the best practice when configuring EMR device volumes using the EMR security configuration?

1. LUKS encryption
2. Secure Hadoop RPC encryption
3. Hadoop in Secure Mode encryption
4. EBS encryption

**Answer:** D

**Explanation:**

Option A is incorrect. LUKS encryption is applied only to attached storage volumes,it does not encrypt the root device volume. AWS recommends EBS encryption over LUKS encryption.

Option B is incorrect. Secure Hadoop RPC encryption is used for in-transit encryption, not at-rest encryption.

Option C is incorrect. Hadoop in Secure Mode encryption is used for in-transit encryption, not at-rest encryption.

Option D is correct. EBS encryption is applied to the attached storage volumes as well as to the root volume. AWS recommends using EBS encryption in your EMR security configuration definition as a best practice.

**Reference:**

Please see the Amazon EMR management guide titled **Encryption Options** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-data-encryption-options.html>), the Amazon EMR management guide titled **Create a Security Configuration** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-create-security-configuration.html>), and the Amazon Elastic Compute Cloud API reference titled **DescribeVolumes** (<https://docs.aws.amazon.com/AWSEC2/latest/APIReference/API_DescribeVolumes.html>)

Question: 21

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Apply data protection and encryption techniques**

**Domain:** Security

**Question text**:

You work as a data scientist for a government agency where your main task is to gather information about your citizens and build a data lake and data warehouse using your citizen information. You ingest your citizen information from many sources and collect it in your data lake using EMR.

Which option depicts the best measures you can take to secure your EMR clusters and the other AWS services used in your data collection/analytics pipeline are as isolated from outside influence as possible?

1. Launch your EMR clusters in a private subnet in your VPC with no public IP and no Internet Gateway access. Place your EMR master, core, and task EC2 instances in an EMR-managed security group. Create a private VPC Endpoint in your private subnet so that your EMR clusters can access your S3 data lake. Create a NAT Gateway in your public subnet to allow your EMR clusters access to KMS and Redshift to access your S3 buckets.
2. Launch your EMR clusters in a private subnet in your VPC with no public IP and no Internet Gateway access. Place your EMR master instance in an EMR-managed security group, place your core, and task EC2 instances in a second EMR-managed security group. Create a private VPC Endpoint in your private subnet so that your EMR clusters can access your S3 data lake. Create a NAT Gateway in your public subnet which allows your EMR clusters access to KMS and allows Redshift to access your S3 buckets.
3. Launch your EMR clusters in a private subnet in your VPC with no public IP and no Internet Gateway access. Place your EMR master instance in an EMR-managed security group, place your core, and task EC2 instances in a second EMR-managed security group. Create a VPC Endpoint in your public subnet so that your EMR clusters can access your S3 data lake. Create a NAT Gateway in your public subnet which allows your EMR clusters access to KMS and allows Redshift to access your S3 buckets.
4. Launch your EMR clusters in a private subnet in your VPC with no public IP and no Internet Gateway access. Place your EMR master instance in an EMR-managed security group, place your core, and task EC2 instances in a second EMR-managed security group. Create a NAT Gateway in your public subnet which allows your EMR clusters access to KMS and your S3 buckets, and allows Redshift to access your S3 buckets.

**Answer:** B

**Explanation:**

Option A is incorrect. The EMR master instance must be placed in a different EMR-managed security group than your core and task instances.

Option B is correct. Placing your EMR clusters in your private subnet is the first level of isolation. Placing your EMR master instance in an EMR-managed security group and your core/task instances in another EMR-managed security group provides the next level of isolation. Using a VPC endpoint and a NAT Gateway to control access to other AWS services provides a third level of isolation.

Option C is incorrect. For the best protection and isolation of your EMR clusters, the VPC Endpoint needs to be placed in your private subnet.

Option D is incorrect. NAT Gateway is not the method to use to access your S3 buckets from your private subnet. The preferred S3 access method is a VPC Endpoint.

**Reference:**

Please see the Amazon EMR management guide titled **Working With Amazon EMR-Managed Security Groups** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-man-sec-groups.html>), the Amazon EMR management guide titled **Specifying Amazon EMR-Managed and Additional Security Groups** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-sg-specify.html>), the Amazon Virtual Private Cloud user guide titled **Endpoints for Amazon S3** (<https://docs.aws.amazon.com/vpc/latest/userguide/vpc-endpoints-s3.html>), and the Amazon Redshift Cluster management guide titled **Amazon Redshift enhanced VPC routing** (<https://docs.aws.amazon.com/redshift/latest/mgmt/enhanced-vpc-routing.html>)

Question: 22

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Apply data governance and compliance controls**

**Domain:** Security

**Question text**:

You work as a data scientist for a healthcare corporation where you are required by Health Insurance Portability and Accountability Act (HIPAA) regulations to record all changes to your data stores for auditing purposes. You have created a data collection pipeline using Kinesis Data Streams, Kinesis Data Firehose and S3 to build your corporate data lake. You have also established AWS Config to record the configuration changes for your AWS resources.

Your AWS Config rules for your S3 buckets in your data lake should send you notifications whenever an S3 bucket is created, modified, or deleted. However, you are not receiving these notifications when your S3 resources change. What might be the cause of this problem?

1. The IAM role you assigned to your S3 buckets does not include the AWSConfigRulesExecutionRole managed policy that allows AWS Config to record changes to your S3 buckets.
2. The IAM role you assigned to AWS Config does not include the AWSConfigRulesExecutionRole managed policy to allow AWS Config to record changes to your S3 buckets.
3. The IAM role you assigned to your S3 buckets does not include the AWSConfigRole managed policy that allows AWS Config to record changes to your S3 buckets.
4. The IAM role you assigned to AWS Config does not include the AWSConfigRole managed policy to allow AWS Config to record changes to your S3 buckets.

**Answer:** D

**Explanation:**

Option A is incorrect. The AWSConfigRulesExecutionRole managed policy is used to allow Lambda to access the AWS Config API and configuration snapshots recorded for your S3 buckets. You are not using the Lambda capability in this scenario. Also, you need to assign a managed policy (the AWSConfigRole policy) to AWS Config, not your S3 buckets, to allow AWS Config to record and notify on changes to your S3 buckets.

Option B is incorrect. The AWSConfigRulesExecutionRole managed policy is used to allow Lambda to access the AWS Config API and configuration snapshots recorded for your S3 buckets. You are not using the Lambda capability in this scenario.

Option C is incorrect. You need to assign a managed policy (the AWSConfigRole policy) to AWS Config, not your S3 buckets, to allow AWS Config to record and notify on changes to your S3 buckets.

Option D is correct. The AWSConfigRole managed policy, associated with the role assigned to AWS Config, allows AWS Config to record and notify on changes to your S3 buckets.

**Reference:**

Please see the **AWS Config FAQs** (<https://aws.amazon.com/config/faq/>), the AWS Config developer guide titled **AWS Config Supported Resource Types and Resource Relationships** (<https://docs.aws.amazon.com/config/latest/developerguide/resource-config-reference.html>), the AWS CloudFormation user guide titled **AWS Resource and Property Types Reference** (<https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/aws-template-resource-type-ref.html>), the AWS Config developer guide titled **Troubleshooting for recording S3 buckets** (<https://docs.aws.amazon.com/config/latest/developerguide/iamrole-permissions.html#troubleshooting-recording-s3-bucket-policy>), and the **HHS.gov HIPPA page** (<https://www.hhs.gov/hipaa/index.html>)

Question: 23

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine the operational characteristics of the collection system**

**Domain:** Data Collection

**Question text**:

You work as a data analytics specialist for a web retail company with vast warehouses across the globe. All of the products sold on your company’s retail web store are distributed from these warehouses to the end customer. You have been asked to produce data analytics applications that allow your management team to understand movement of product through your warehouse system. You have built a data collection system consisting of a Kinesis Data stream fed by data producers written using the Kinesis Producer Library (KPL) and consumed by Kinesis applications written using the Kinesis Client Library (KCL). The consumers use the Kinesis Connector Library to write your records to S3.

Because of the very large number of records produced by your KPL applications, you have decided to use KPL aggregation. Also, your KCL record processing code relies on unique identifiers for the processing of your KPL user records. What attribute of your streamed records can you use as your unique identifier for your KPL user records after de-aggregating the Kinesis Data Stream record?

1. Use the aggregated Kinesis Data Stream record sequence number
2. Use the KPL user record sequence number
3. Use the KPL user record subsequence number
4. Use the KPL user record hashCode

**Answer:** B

**Explanation:**

Option A is incorrect. The aggregated Kinesis Data Stream record contains an aggregated group of your user records, which represent your product data. Therefore, the aggregated record sequence number does not represent the data on which you wish to perform analytics.

Option B is correct. You can use the KPL user record sequence number as your unique identifier as long as you use the KPL Record or UserRecord class hashCode and equals operations when comparing your user records.

Option C is incorrect. The KCL subsequence number is used primarily for checkpointing.

Option D is incorrect. The KCL hashCode method of the Record and UserRecord KCL classes is used for comparison of user records, not as an identifier.

**Reference:**

Please see the Amazon Kinesis Data Streams developer guide titled **Considerations When Using KPL Aggregation** (<https://docs.aws.amazon.com/streams/latest/dev/kinesis-producer-adv-aggregation.html>), the Amazon Kinesis Data Streams developer guide titled **Consumer De-aggregation** (<https://docs.aws.amazon.com/streams/latest/dev/kinesis-kpl-consumer-deaggregation.html>), and the Amazon Kinesis Data Streams getting started page titled **Getting started with Amazon Kinesis Data Streams** (<https://aws.amazon.com/kinesis/data-streams/getting-started/>)

Question: 24

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine the operational characteristics of the collection system**

**Domain:** Data Collection

**Question text**:

You work as a data analytics specialist for an international airline. Your company gathers sensor data from all of the airlines in your company’s fleet. These sensors emit data describing the flight patterns of their planes, the health of the airplane components, and fuel usage patterns during flight.

Your team is responsible for gathering this sensor data into a data warehouse for use in analytics apps used by your data scientists to gain insights into the efficiency of their airline fleet. Which of the following options is the simplest and most cost effective method of getting the airline data into your data warehouse?

1. Sensor data -> Kinesis Data Firehose -> Redshift
2. Sensor data -> Kinesis Data Streams -> KCL/KCL-C -> Redshift
3. Sensor data -> Kinesis Data Firehose -> S3 -> Redshift
4. Sensor data -> Kinesis Data Streams -> S3 -> Redshift

**Answer:** C

**Explanation:**

Option A is incorrect. To load your data directly into Redshift from Kinesis Data Firehose you need to use an S3 bucket as an intermediary staging location from which Kinesis issues the Redshift COPY command to load the data into your Redshift tables. See the following screen shot from the AWS console:



Option B is incorrect. While this solution will work, it is less cost efficient (you would have to write KCL/KCL-C application code) and not as simple as using Kinesis Firehose to write to S3 from which Kinesis then issues the Redshift COPY command to load your data into your Redshift tables.

Option C is correct. The most cost efficient and simplest solution is to use Kinesis Firehose to write to S3 from which Kinesis then issues the Redshift COPY command to load your data into your Redshift tables.

Option D is incorrect. In order to be able to write your Kinesis Data Stream data to S3 you would have to write a Kinesis client application, which will be less cost efficient and not as simple as using Kinesis Firehose to write to S3 from which Kinesis then issues the Redshift COPY command to load your data into your Redshift tables.

**Reference:**

Please see the Amazon Kinesis Data Streams getting started page titled **Getting started with Amazon Kinesis Data Streams** (<https://aws.amazon.com/kinesis/data-streams/getting-started/>), the **Amazon Redshift overview page** (<https://aws.amazon.com/redshift/>), the GitHub repo titled **Amazon Kinesis Connector Library** (<https://github.com/awslabs/amazon-kinesis-connectors>), and the Amazon Kinesis Data Firehose developer guide titled **Amazon Kinesis Data Firehose Data Delivery** (<https://docs.aws.amazon.com/firehose/latest/dev/basic-deliver.html>)

Question: 25

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Select a collection system that addresses the key properties of data, such as order, format, and compression**

**Domain:** Data Collection

**Question text**:

You work as a data analytics specialist for a gaming software company. You are building a streaming data collection pipeline to capture real-time game statistics to use in recommendation engine analytics, such as in-app purchase and online friend match suggestion.

You are streaming your game data through a Kinesis Data stream with the Kinesis Producer Library (KPL) aggregating your user records. Your management team has stressed that real-time recommendations need to be as fast as possible. Which option gives you the best consumer application deaggregation method to retrieve your user records?

1. Use the Kinesis API GetRecords using the deaggregate static operation from Record.java
2. Use the Kinesis Consumer Library (KCL) getData operation from Record.java
3. Use the Kinesis API GetRecords using the deaggregate static operation from UserRecord.java
4. Use the Kinesis Consumer Library (KCL) getData operation from UserRecord.java

**Answer:** C

**Explanation:**

Option A is incorrect. You should use the Kinesis Data Streams API to retrieve and deaggregate to get your user records because speed of processing is an important requirement for your solution. However, the GetRecords method is a static operation in UserData.java, not Record.java

Option B is incorrect. You should use the Kinesis Data Streams API to retrieve and deaggregate to get your user records because speed of processing is an important requirement for your solution. Therefore, you should not use the KCL since the KCL carries a performance overhead compared to the use of the Kinesis Data Streams API.

Option C is correct. You should use the Kinesis Data Streams API to retrieve and deaggregate to get your user records because speed of processing is an important requirement for your solution. The GetRecords method is a static operation in UserData.java.

Option D is incorrect. You should use the Kinesis Data Streams API to retrieve and deaggregate to get your user records because speed of processing is an important requirement for your solution. Therefore, you should not use the KCL since the KCL carries a performance overhead compared to the use of the Kinesis Data Streams API.

**Reference:**

Please see the Amazon Kinesis Data Streams developer guide titled **Considerations When Using KPL Aggregation** (<https://docs.aws.amazon.com/streams/latest/dev/kinesis-producer-adv-aggregation.html>), the Amazon Kinesis Data Streams developer guide titled **Consumer De-aggregation** (<https://docs.aws.amazon.com/streams/latest/dev/kinesis-kpl-consumer-deaggregation.html>), and the Amazon Kinesis Data Streams developer guide titled **Developing a Kinesis Client Library Consumer in Java** [**https://docs.amazonaws.cn/en\_us/streams/latest/dev/kinesis-record-processor-implementation-app-java.html**](https://docs.amazonaws.cn/en_us/streams/latest/dev/kinesis-record-processor-implementation-app-java.html))

Question: 26

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine data access and retrieval patterns**

**Domain:** Storage and Data Management

**Question text**:

You work as a data analytics specialist for the Program Management Office (PMO) for one of your company’s biggest new business programs in this year’s budget. You are responsible for gathering streaming data about customer activity on your company’s various application platforms: web, mobile, IoT sensors such as those on Apple Watch, etc. This data is being used in analytics applications to understand customer interaction with your company’s software product, which will be used to guide the program development and implementation plan.

You have built a data collection system that is based on Kinesis Data Streams feeding Kinesis Data Firehose which then loads your data into your Redshift data warehouse. During testing of your data collection system you have noticed that your data is not being delivered into your Redshift cluster. Which of the following should you NOT check when trying to determine the problem?

1. Check the STL\_CONNECTION\_LOG table on your Redshift cluster to see if Kinesis Data Firehose can successfully connect to your Redshift cluster
2. Check the STL\_LOAD\_ERRORS Redshift table to find information on the failure to COPY
3. Ensure the HEC token you’re using is valid
4. Check that the IAM role that is specified in your Kinesis Data Firehose delivery stream can access the S3 bucket from which your Redshift cluster copies data

**Answer:** C

**Explanation:**

Option A is incorrect. The STL\_CONNECTION\_LOG shows information about your Firehose connectivity to your Redshift cluster.

Option B is incorrect. The STL\_LOAD\_ERRORS table shows information about Redshift COPY failures.

Option C is correct. The HEC token is used when attempting to deliver data to Splunk, not Redshift.

Option D is incorrect. The IAM role specified in your Kinesis Data Firehose delivery stream allows Firehose to write your data to S3 before Redshift can COPY the data.

**Reference:**

Please see the Amazon Kinesis Data Firehose developer guide titled **Troubleshooting Amazon Kinesis Data Firehose** (<https://docs.aws.amazon.com/firehose/latest/dev/troubleshooting.html>), and the Amazon Kinesis Data Firehose developer guide titled **Sending Data to an Amazon Kinesis Data Firehose Delivery Stream** (<https://docs.aws.amazon.com/firehose/latest/dev/basic-write.html>)

Question: 27

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine the operational characteristics of a storage solution for analytics**

**Domain:** Storage and Data Management

**Question text**:

You work as a data analytics specialist for a Software as a Service (SaaS) provider that provides software to the insurance industry. Your software allows small to medium sized insurance agencies to manage their client base and their insurance premium data.

You have built a data collection system that uses Kinesis Data Streams to feed a Kinesis Data Firehose stream. You want to configure the Firehose stream to leverage Lambda to transform your data and then write your data to your Elasticsearch cluster so that you can provide a cached data search capability for your SaaS offering.

When you conduct your first tests you find that your streaming data is not being delivered to your Elasticsearch domain. What may be the root of the problem?

1. Check the SucceedProcessing metric data in CloudWatch
2. Check the KMSKeyNotFound metric data in CloudWatch
3. Check the STL\_LOAD\_ERRORS table
4. Check that your Firehose stream can access your VPC

**Answer:** A

**Explanation:**

Option A is correct. The SucceedProcessing metric data in CloudWatch tells you how many of records were successfully processed over a period of time when using Lambda for transformation. You are using Lambda to transform your data prior to writing it to Elasticsearch.

Option B is incorrect. The scenario does not mention using encryption so this option should be eliminated.

Option C is incorrect. The STL\_LOAD\_ERRORS table is a Redshift construct. You are not using Redshift in this scenario.

Option D is incorrect. VPC access comes into the picture when writing to Splunk, not Elasticsearch.

**Reference:**

Please see the Amazon Kinesis Data Firehose developer guide titled **Troubleshooting Amazon Kinesis Data Firehose** (<https://docs.aws.amazon.com/firehose/latest/dev/troubleshooting.html>), and the Amazon Kinesis Data Firehose developer guide titled **Monitoring Kinesis Data Firehose Using CloudWatch Metrics** (<https://docs.aws.amazon.com/firehose/latest/dev/monitoring-with-cloudwatch-metrics.html>), the Kinesis Data Firehose developer guide titled **Sending Data to an Amazon Kinesis Data Firehose Delivery Stream** (<https://docs.aws.amazon.com/firehose/latest/dev/basic-write.html>), and the **Amazon Elasticsearch Service overview page** (<https://aws.amazon.com/elasticsearch-service/>)

Question: 28

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine an appropriate system for cataloging data and managing metadata**

**Domain:** Storage and Data Management

**Question text**:

You work as a data analytics specialist for the accounting department of a large multinational bank where you need to manage the company’s AWS account expenditures in an automated way. Your AWS accounts have many VPCs through which your applications and services send their IP traffic. Your task is to use this VPC flow information to help lend insight into your AWS monthly spend by account.

Your data collection solution looks like this: VPC Flow logs are sent to CloudWatch. The CloudWatch data is streamed through Kinesis Data Firehose which leverages Lambda to decompress the log data and finally sends the decompressed log data to Splunk for analytics and monitoring.

How can you most efficiently use metadata in your existing stream records to categorize the stream records by account owner, VPC, and service used?

1. Have your Lambda to tag your log data using the Kinesis Data Streams API AddTagsToStream call
2. Assign tags by using the CreateDeliveryStream Kinesis Data Firehose API operation when you create a new delivery stream
3. Assign tags by using the CreateDeliveryStream Kinesis Data Firehose API operation on your existing delivery stream
4. Assign tags by using the TagDeliveryStream Kinesis Data Firehose API operation on your existing delivery streams

**Answer:** D

**Explanation:**

Option A is incorrect. The AddTagsToStream API call is a Kinesis Data Streams API call. You are using Kinesis Data Firehose.

Option B is incorrect. The scenario states that you need to add metadata to your existing delivery streams. This option describes how to tag your streams when you create them.

Option C is incorrect. The CreateDeliveryStream API call is used when you create a new delivery stream, not when you wish to modify an existing stream by adding tags to it.

Option D is correct. The TagDeliveryStream Kinesis Data Firehose API operation allows you to apply tags to an existing delivery stream.

**Reference:**

Please see the Amazon Kinesis Data Firehose developer guide titled **Tagging Your Delivery Streams in Amazon Kinesis Data Firehose** (<https://docs.aws.amazon.com/firehose/latest/dev/firehose-tagging.html>), the Amazon Kinesis Data Firehose API reference titled **CreateDeliveryStream** (<https://docs.aws.amazon.com/firehose/latest/APIReference/API_CreateDeliveryStream.html>), the Amazon Kinesis Data Streams developer guide titled **Tagging Your Streams in Amazon Kinesis Data Streams** (<https://docs.aws.amazon.com/streams/latest/dev/tagging.html>), the Amazon Kinesis Data Streams Service API reference titled **AddTagsToStream** (<https://docs.aws.amazon.com/kinesis/latest/APIReference/API_AddTagsToStream.html>), and the Kinesis Data Firehose developer guide titled **Tutorial: Sending VPC Flow Logs to Splunk Using Amazon Kinesis Data Firehose** (<https://docs.aws.amazon.com/firehose/latest/dev/vpc-splunk-tutorial.html>)

Question: 29

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine the operational characteristics of a storage solution for analytics**

**Domain:** Storage and Data Management

**Question text**:

You work as a data analytics specialist for a music streaming service. Your team has been assigned the task of capturing the music selection activity of your millions of users and storing the data in a data lake for use in analytics applications.

You have chosen to populate your data lake via a data collection system that uses Kinesis Data Streams to capture the data records from a producer application. You have the requirement to keep the data records in sequence order so that your analytics applications can infer when users take a sequence of actions, such as selecting a song to play and then either skipping the song or marking the song as a favorite.

As you start to test your data collection pipeline you notice that some of your data records arrive out of sequence. Which option can help you correct this problem?

1. Change your KMS key policy action to sequential
2. Use the PutRecords Kinesis API call to write your records to your Kinesis stream
3. Set the SequenceNumberForOrdering in your PutRecords Kinesis API call when you write your records to your Kinesis stream
4. Use the PutRecord API call to write your records to your Kinesis stream

**Answer:** D

**Explanation:**

Option A is incorrect. There is no KMS key policy sequential action

Option B is incorrect. With the PutRecords API call, a failed record is skipped and all subsequent records are processed. Therefore, the PutRecords API call does not guarantee data record ordering.

Option C is incorrect. The SequenceNumberForOrdering parameter is only used in the PutRecord API call, not the PutRecords API call

Option D is correct. With the PutRecords API call, a failed record is skipped and all subsequent records are processed. Therefore, the PutRecords API call does not guarantee data record ordering. When you write your records to the same shard, the PutRecord API call will guarantee data record ordering.

**Reference:**

Please see the Amazon Kinesis Data Streams developer guide titled **Troubleshooting Amazon Kinesis Data Streams Producers** (<https://docs.aws.amazon.com/streams/latest/dev/troubleshooting-producers.html>), the Amazon Kinesis Data Streams Service API reference titled **PutRecords** (<https://docs.aws.amazon.com/kinesis/latest/APIReference/API_PutRecords.html>), the Amazon Kinesis Data Streams developer guide titled **Developing Producers Using the Amazon Kinesis Data Streams API with the AWS SDK for Java** (<https://docs.aws.amazon.com/streams/latest/dev/developing-producers-with-sdk.html#kinesis-using-sdk-java-add-data-to-stream>), and the AWS Key Management Service developer guide titled **Using key policies in AWS KMS** (<https://docs.aws.amazon.com/kms/latest/developerguide/key-policies.html>)

Question: 30

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Select an appropriate data layout, schema, structure, and format**

**Domain:** Storage and Data Management

**Question text**:

You work as a data analytics specialist for a polling analytics firm that is building a data warehouse to hold its polling data for an upcoming parliamentary election. The data will need to be loaded into your data warehouse in compressed format to allow for the best performance when querying the warehouse.

You have decided to use the Redshift automatic compression feature to accomplish your data compression in your warehouse. Which of the following options is NOT one of the operations performed by automatic compression when loading your data into your Redshift tables using the COPY command?

1. Automatic compression loads an initial sample of rows from the input file
2. Automatic compression chooses compression options for each column
3. Automatic compression copies the sample rows to the table
4. Automatic compression recreates the table with the chosen compression encodings
5. Automatic compression loads and compresses the entire input file using the new encodings

**Answer:** C

**Explanation:**

Option A is incorrect. The initial sample of rows load is the first step automatic compression performs in the compression process.

Option B is incorrect. As the second step in the compression process, automatic compression chooses the compression option for each column.

Option C is correct. As the third step in the compression process, automatic compression removes the sample rows from the table, it does not copy the sample rows to the table.

Option D is incorrect. As the fourth step in the compression process, automatic compression recreates the table with the chosen compression encodings.

Option E is incorrect. As the last step in the compression process, automatic compression loads and compresses all of the input file contents using the chosen encodings.

**Reference:**

Please see the Amazon Redshift Database developer guide titled **Choosing a column compression type** (<https://docs.aws.amazon.com/redshift/latest/dg/t_Compressing_data_on_disk.html>), and the Amazon Redshift Database developer guide titled **Loading tables with automatic compression** (<https://docs.aws.amazon.com/redshift/latest/dg/c_Loading_tables_auto_compress.html>)

Question: 31

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Select an appropriate data layout, schema, structure, and format**

**Domain:** Storage and Data Management

**Question text**:

You work as a data analytics specialist for a food delivery mobile application service. Your service matches restaurants offering delivery service with customers in the regional area of the given restaurants in your system. Your management team needs to gain insight into fulfillment, delivery route efficiency, customer satisfaction, and other key metrics of their service. To meet this end your team is building a data warehouse that will store the business data and allow you to produce key metric analytics views and dashboards to your management team.

You have loaded all of your data into your Redshift data warehouse and have started to create the business intelligence analytics views for your management team. However, your queries are not performing as well as they should. The response time for producing analytics insights is slow. You have decided to leverage column compression on your Redshift tables to improve query performance. How would you apply compression to your customer table in Redshift?

1. update table customer(

customer\_id int encode raw,

customer\_name char(20) encode az64);

1. create table customer(

customer\_id int encode raw,

customer\_name char(20) encode bytedict);

1. create table customer(

customer\_id int encode raw,

customer\_name char(20) encode mostly8);

1. create schema customer(

customer\_id int encode raw,

customer\_name char(20) encode bytedict);

**Answer:** B

**Explanation:**

Option A is incorrect. You cannot change (update) the compression encoding on a table after it has been created. You need to create a new table with your desired compression encoding and then load your existing data into the new table. Also, the AZ64 encoding does not support CHAR or VARCHAR.

Option B is correct. After you run the ANALYZE COMPRESSION command on your existing table you can use the results to select the compression encodings you’ll use when you create a new table to populate with your existing data. You need to create a new table with your desired compression encodings and then load your existing data into the new table.

Option C is incorrect. The MOSTLY8 encoding does not support CHAR or VARCHAR.

Option D is incorrect. To create a new table with your desired encodings, you use the CREATE TABLE command, not the CREATE SCHEMA command.

**Reference:**

Please see the Amazon Redshift Database developer guide titled **Choosing a column compression type** (<https://docs.aws.amazon.com/redshift/latest/dg/t_Compressing_data_on_disk.html>), and the Amazon Redshift Database developer guide titled **Compression encodings** (<https://docs.aws.amazon.com/redshift/latest/dg/c_Compression_encodings.html>), the Amazon Redshift Database developer guide titled **Loading tables with automatic compression** (<https://docs.aws.amazon.com/redshift/latest/dg/c_Loading_tables_auto_compress.html>), and the Amazon Redshift Database developer guide titled **SQL commands** (<https://docs.aws.amazon.com/redshift/latest/dg/c_SQL_commands.html>)

Question: 32

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Define a data lifecycle based on usage patterns and business requirements**

**Domain:** Storage and Data Management

**Question text**:

You work as a data analytics specialist for a food processing corporation. Your company processes plant and animal products for distribution across the globe. You need to maintain a data warehouse to store information about your food products such as their production date, their shelf life, and their destination. This information must be backed up and stored for auditing purposes.

You have chosen Redshift as your data warehouse technology. You are now configuring your snapshot schedule for your primary tables. Which of the following options defines your snapshot schedule to occur every day of the week starting at 12:30 AM on a 1 hour increment?

1. cron(30 12/1 ALL)
2. create-snapshot-schedule(30 12/1 \*)
3. create-snapshot-schedule(30 1/12 \*)
4. cron(30 0/1 \*)

**Answer:** D

**Explanation:**

Option A is incorrect. The ‘ALL’ keyword in the command is incorrect. To schedule your snapshot to run every day you should have an asterix in place of the word ALL. Also, this example will run your snapshots on a 1 hour interval starting at 12:30 PM. You want to start at 12:30 AM.

Option B is incorrect. The create-snapshot-schedule command still requires a cron command within it such as create-snapshot-schedule --schedule-identifier “theschedule" --schedule-definition "cron(30 12 \*)”. Also, this example will run your snapshots on a 1 hour interval starting at 12:30 PM. You want to start at 12:30 AM.

Option C is incorrect. The create-snapshot-schedule command still requires a cron command within it such as create-snapshot-schedule --schedule-identifier “theschedule" --schedule-definition "cron(30 12 \*)”. Also, this example will run your snapshots on a 12 hour interval starting at 1:30 AM. You want to start at 12:30 AM with a 1 hour interval.

Option D is correct. This command will run your snapshot every day starting at 12:30 AM on a one hour increment.

**Reference:**

Please see the Amazon Redshift Database developer guide titled **Data warehouse system architecture** (<https://docs.aws.amazon.com/redshift/latest/dg/c_high_level_system_architecture.html>), and the Amazon Redshift Cluster management guide titled **Amazon Redshift snapshots** (<https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-snapshots.html>)

Question: 33

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine appropriate data processing solution requirements**

**Domain:** Processing

**Question text**:

You work as a data analytics specialist for a vineyard that produces wine for the european market. Your company classifies their wine in the usual appellations such as Rhône, Loire, Alsace, Bordeaux, etc. You need to create a data warehouse to store all of the data about your company’s wines by appellation. This data warehouse will be used for various data analytics applications to allow your management team to gain insight into their customer’s tastes and preferences. These analytics applications will be used to drive product development for the future.

You are designing your Redshift data warehouse cluster and you need to construct your basic queries for extracting your data for your analytics applications. Because of the large amount of data stored in your warehouse you want to use distributed queries across your warehouse compute nodes to improve performance (see the diagram below).

[](https://www.draw.io/?page-id=8T0nsWxOn3itxknRsEfa&scale=auto#G1_YGlINztwZn7VlPmss9mKjLBiU_yI6eG)

Which option represents an example query you would run on your compute nodes?

1. select \* from PG\_EXTERNAL\_SCHEMA where schemaname = current\_schema() limit 10;
2. select current\_schema(), appellation from vineyards;
3. select slice, col, tbl from STV\_BLOCKLIST where tombstone > 17;
4. select appellation\_id from PG\_STATISTIC\_INDICATOR where stairows > 100;

**Answer:** C

**Explanation:**

Option A is incorrect. This query references a system catalog table, PG\_EXTERNAL\_SCHEMA. Therefore, this query runs exclusively on the leader node.

Option B is incorrect. This query references a table that is located on your compute nodes (vineyards), but the current\_schema method only runs on your leader node. This query will result in an error.

Option C is correct. This query references an STV table, which is a system table that is located on your compute nodes.

Option D is incorrect. This query references a system catalog table, PG\_STATISTIC\_INDICATOR. Therefore, this query runs exclusively on the leader node.

**Reference:**

Please see the Amazon Redshift Database developer guide titled **Data warehouse system architecture** (<https://docs.aws.amazon.com/redshift/latest/dg/c_high_level_system_architecture.html>), and the Amazon Redshift Database developer guide titled **SQL functions supported on the leader node** (<https://docs.aws.amazon.com/redshift/latest/dg/c_sql-functions-leader-node.html>), the Amazon Redshift Database developer guide titled **Leader node–only functions** (<https://docs.aws.amazon.com/redshift/latest/dg/c_SQL_functions_leader_node_only.html>), the Amazon Redshift Database developer guide titled **Compute node–only functions** (<https://docs.aws.amazon.com/redshift/latest/dg/c_SQL_functions_compute_node_only.html>), and the Amazon Redshift Database developer guide titled **System catalog tables** (<https://docs.aws.amazon.com/redshift/latest/dg/c_intro_catalog_views.html>)

Question: 34

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Automate and operationalize a data processing solution**

**Domain:** Processing

**Question text**:

You work as a data analytics specialist for a transportation company that is using their EMR cluster to ingest and provide query capabilities for their massive amounts of shipping routes and load information.

You need your metastore for your EMR cluster to be persistent. Based on your requirements, which option gives you the best metastore option with the least amount of effort?

1. Save your metastore information in your Hive MySQL database on the master node file system.
2. Store your metastore information in an external metastore in DynamoDB tables.
3. Save your metastore information in an external metastore in Aurora.
4. Save your metastore information in an external metastore in the Glue data catalog.

**Answer:** D

**Explanation:**

Option A is incorrect. The master node terminates when you terminate your cluster, therefore your metastore will be lost (not persistent) since it resides on the master node’s file system.

Option B is incorrect. The two options for external (persistent) storage of your Hive metastore are Glue data catalog and RDS/Aurora. Therefore, DynamoDB is an invalid option.

Option C is incorrect. While using an Aurora database as your external metastore for your Hive metastore information, this option is not as simple to implement as the Glue data catalog option.

Option D is correct. Using the Glue data catalog as your external Hive metastore is the simplest option of the two valid options listed.

**Reference:**

Please see the **Amazon EMR overview page** (<https://aws.amazon.com/emr/>), the Amazon EMR release guide titled **Configuring an External Metastore for Hive** (<https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-metastore-external-hive.html>), the Amazon EMR release guide titled **Using the AWS Glue Data Catalog as the Metastore for Hive** (<https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hive-metastore-glue.html>), the Amazon EMR release guide titled **Using an External MySQL Database or Amazon Aurora** (<https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hive-metastore-external.html>)

Question: 35

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine the operational characteristics of an analysis and visualization solution**

**Domain:** Analysis and Visualization

**Question text**:

You work as a data analytics specialist for a real estate company that owns high rise buildings in large cities around the world. You are creating a data visualization of your firm’s properties to show property preparedness for restructuring (ventilation, seating capacity, contact tracing, etc.) based on proposed changes in building occupancy regulations.

You have a data collection pipeline that loads building data from your various data sources (data warehouses, relational databases, etc.) into your data lake on S3. You are creating these visualizations in QuickSight. Several of your data files arrive on your S3 buckets in the gzip compressed format, while others are in the XLSX and TSV formats. How can you prepare these source data files for use as data sources in your QuickSight visualizations?

1. Import the compressed files, XLSX, and TSV files into QuickSight as data sources as-is
2. Unzip the compressed files, then import the uncompressed files, the XLSX, and TSV files into QuickSight as data sources
3. Unzip the compressed files, convert the XLSX files to CSV, then import all files into QuickSight as data sources
4. Unzip the compressed files, convert the XLSX and TSV files to CSV, then import all files into QuickSight as data sources

**Answer:** A

**Explanation:**

Option A is correct. You can import compressed files in the gzip format as-is into QuickSight as data sources. Also, the XLSX and TSV file format types are supported natively (no need to convert them) by QuickSight as data sources.

Option B is incorrect. You don’t need to unzip your gzip files because QuickSight can load gzip files as-is as data sources.

Option C is incorrect. You don’t need to unzip your gzip files because QuickSight can load gzip files as-is as data sources. Also, you don’t need to convert XLSX formatted files because QuickSight supports this file type natively.

Option D is incorrect. You don’t need to unzip your gzip files because QuickSight can load gzip files as-is as data sources. Also, you don’t need to convert the XLSX and TSV formatted files because QuickSight supports these file types natively.

**Reference:**

Please see the Amazon QuickSight user guide titled **Working with Data Sources in Amazon QuickSight** (<https://docs.aws.amazon.com/quicksight/latest/user/working-with-data-sources.html>), and the Amazon QuickSight user guide titled **Supported Data Sources** (<https://docs.aws.amazon.com/quicksight/latest/user/supported-data-sources.html>)

Question: 36

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Select the appropriate data visualization solution for a given scenario**

**Domain:** Analysis and Visualization

**Question text**:

You work as a data scientist for a car manufacturer that has started to venture into the fully electric vehicle market. Your company currently has two models of electric cars in the US market. These cars have many sensors on them that emit data back to your EMR cluster where you store the data in your S3 data lake. You need to be able to be able to generate interactive visualizations in real-time of the sensor data for your management team so that they can get insight into the use of their new electric models.

What is the most efficient way to build your real-time visualization for your management team?

1. Create Presto tables in your Hive metastore and import your data into SPICE so you can create QuickSight visualizations using Presto queries.
2. Create Presto tables in your Hive metastore then create QuickSight visualizations using Presto queries.
3. Create Presto tables in your Hive metastore and specify the Glue data catalog as the default Hive metastore for Presto. Create QuickSight visualizations using Presto queries.
4. Create Presto tables in your Hive metastore and specify the Glue data catalog as the default Hive metastore for Presto. Use a Glue ETL job to load the data from Presto onto your S3 bucket in the CSV format. Create QuickSight visualizations using the CSV files.

**Answer:** B

**Explanation:**

Option A is incorrect. Loading your data from Presto into SPICE makes your data static. You need real-time visualizations.

Option B is correct. You can directly query your data using Presto queries from your QuickSight visualizations. This gives you real-time visualizations as the data arrives in your EMR cluster.

Option C is incorrect. There is no need to use the Glue data catalog as your Hive metastore since you can just use the Hive metastore. Using the Glue data catalog over complicates the solution making this option less efficient.

Option D is incorrect. There is no need to use the Glue data catalog as your Hive metastore since you can just use the Hive metastore. Also, using a Glue ETL job to convert your data to CSV before using the data in your visualizations is unnecessary, since you can query your data directly from QuickSight using Presto queries. This over complicates the solution making this option less efficient.

**Reference:**

Please see the Amazon QuickSight user guide titled **Creating a Data Source Using Presto** (<https://docs.aws.amazon.com/quicksight/latest/user/create-a-data-source-presto.html>), the Amazon EMR release guide titled **Using Presto with the AWS Glue Data Catalog** (<https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-presto-glue.html>), the Amazon QuickSight user guide titled **Supported Data Sources** (<https://docs.aws.amazon.com/quicksight/latest/user/supported-data-sources.html>), and the AWS Big Data blog titled **Visualize Big Data with Amazon QuickSight, Presto, and Apache Spark on Amazon EMR** (<https://aws.amazon.com/blogs/big-data/visualize-big-data-with-amazon-quicksight-presto-and-apache-spark-on-amazon-emr/>)

Question: 37

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Select appropriate authentication and authorization mechanisms**

**Domain:** Security

**Question text**:

You work as a data analytics specialist for a banking company where you are building a data warehouse to store data about customers and their accounts. This data warehouse will be used by analysts building customer insight analytics applications.

Since the data warehouse you are creating houses Personally Identifiable Information (PII), you need to restrict access to your Redshift data warehouse tables. In the simplest way possible, how can you restrict access so that users only can retrieve the information they need at the most granular level?

1. Create IAM roles for each user group/type that have granular permissions defined.
2. Build stored procedures with the DEFINER security attribute in Redshift that control the access to the data needed for each function.
3. Create IAM policies that define granular access permissions and attach these policies to each user group/type.
4. Build stored procedures with the INVOKER security attribute in Redshift that control the access to the data needed for each function.

**Answer:** D

**Explanation:**

Option A is incorrect. While you could define IAM roles and permissions to control access to Redshift, you could not control down to the level of which data elements in each column your users can access.

Option B is incorrect. Using a stored procedure in Redshift you can give your users access to the data they need, and only the data they need, without giving the users access to the underlying tables. However, the DEFINER security attribute for a stored procedure in Redshift is for use by the stored procedure’s owner. The INVOKER security attribute for Redshift stored procedures is the default security attribute, where the procedure runs under the permissions of the user who calls it, thereby restricting access to only retrieving the results of the stored procedure.

Option C is incorrect. While you could define IAM policies defining permissions to control access to Redshift, you could not control down to the level of which data elements in each column your users can access.

Option D is correct. Using a stored procedure in Redshift you can give your users access to the data they need, and only the data they need, without giving the users access to the underlying tables. The INVOKER security attribute for Redshift stored procedures is the default security attribute, where the procedure runs under the permissions of the user who calls it, thereby restricting access to only retrieving the results of the stored procedure.

**Reference:**

Please see the Amazon Redshift Database developer guide titled **Creating stored procedures in Amazon Redshift** (<https://docs.aws.amazon.com/redshift/latest/dg/stored-procedure-overview.html>), the Amazon Redshift Database developer guide titled **Overview of stored procedures in Amazon Redshift** (<https://docs.aws.amazon.com/redshift/latest/dg/stored-procedure-create.html>), and the Amazon Redshift Database developer guide titled **Security and privileges for stored procedures** (<https://docs.aws.amazon.com/redshift/latest/dg/stored-procedure-security-and-privileges.html>)

Question: 38

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Apply data protection and encryption techniques**

**Domain:** Security

**Question text**:

You work as a data analytics specialist for a healthcare company that has a data warehouse that was built to hold data about their healthcare provider network. This data was encrypted using AWS Key Management Service (KMS) encryption. You have now been tasked with storing healthcare customer medical history in the same data warehouse. Since the customer medical history data is highly confidential, what can you do to further increase the security of your Redshift cluster?

1. Modify your Redshift cluster to use an AWS Key Management Service (KMS) customer managed key (CMK) for encryption.
2. Modify your Redshift cluster to enable hardware security module (HSM) encryption.
3. Create a new Redshift cluster that is HSM enabled and migrate your data to the new cluster.
4. Modify your Redshift cluster to use an AWS Key Management Service (KMS) AWS managed key for encryption.

**Answer:** C

**Explanation:**

Option A is incorrect. Changing from an AWS managed key to a customer managed key for encryption will not increase the security of your Redshift cluster.

Option B is incorrect. You cannot directly modify a Redshift cluster to use HSM. You must first create a new Redshift cluster that is HSM enabled and migrate your data to the new cluster.

Option C is correct. Hardware security module (HSM) encryption is the most secure encryption you can use on your Redshift cluster. You cannot directly modify a Redshift cluster to use HSM. You must first create a new Redshift cluster that is HSM enabled and migrate your data to the new cluster.

Option D is incorrect. Changing from a customer managed key to an AWS managed key for encryption will not increase the security of your Redshift cluster.

**Reference:**

Please see the Amazon Redshift Cluster Management guide titled **Amazon Redshift database encryption** (<https://docs.amazonaws.cn/en_us/redshift/latest/mgmt/working-with-db-encryption.html>), and the Amazon Redshift Cluster Management guide titled **Changing cluster encryption** (<https://docs.amazonaws.cn/en_us/redshift/latest/mgmt/changing-cluster-encryption.html>)

Question: 39

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Apply data protection and encryption techniques**

**Domain:** Security

**Question text**:

You work as a data analytics specialist for a financial services firm where you are building their EMR clusters for use in your data lake. You plan to extract and store data from your firm’s accounting, customer, securities, and banking systems for use in your EMR ingestion clusters. Your EMR query clusters will be used as the source for business intelligence (BI) analytics applications.

You need to make sure your EMR cluster data is encrypted at rest. Which of the following options is NOT a valid encryption approach for EMR encryption at rest?

1. SSE-S3
2. SSE-KMS
3. CSE-KMS
4. AES-256

**Answer:** D

**Explanation:**

Option A is incorrect. SSE-S3 is a valid choice for encryption at rest for EMR data on S3.

Option B is incorrect. SSE-KMS is a valid choice for encryption at rest for EMR data on S3.

Option C is incorrect. CSE-KMS is a valid choice for encryption at rest for EMR data on S3.

Option D is correct. AES-256 is not a valid choice for encryption at rest for EMR data on S3.  
  
Here is a screen shot from the AWS console showing the valid choices:



**Reference:**

Please see the AWS Big Data blog titled **Best Practices for Securing Amazon EMR** (<https://aws.amazon.com/blogs/big-data/best-practices-for-securing-amazon-emr/>), the Amazon EMR management guide titled **Encryption Options** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-data-encryption-options.html>), and the AWS Cryptography services and tools guide titled **Cryptographic algorithms** (<https://docs.aws.amazon.com/crypto/latest/userguide/concepts-algorithms.html>)

Question: 40

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Apply data governance and compliance controls**

**Domain:** Security

**Question text**:

You work as a data analytics specialist for a software company that develops Software as a Service options for the healthcare industry. In order to provide very fast access to your systems data using the simplest caching mechanism, you have chosen to implement the Elasticache Memcached cache engine.

Because your company develops software solutions for the healthcare industry, you are required by Health Insurance Portability and Accountability Act (HIPAA) regulations to record all changes to your data stores for auditing purposes. To this end, you need to record each time your Elasticache Memcached cache engine scales out or in, adding and removing nodes based on changes to the demand on your system. How would you most efficiently implement this monitoring in your infrastructure?

1. Use the AWS Config elasticache-memcached-scale-check managed rule.
2. Write a Lambda function that is triggered and records each time your Memcached nodes scale in or out.
3. Create a custom AWS Config rule to monitor your Memcached nodes and record each time they scale in or out.
4. Use the AWS Config elasticache-memcached-fanout-check managed rule.

**Answer:** C

**Explanation:**

Option A is incorrect. There is no elasticache-memcached-scale-check AWS Config managed rule.

Option B is incorrect. Creating a lambda function outside the AWS Config rules will put much of the burden of managing the configuration check on you. The AWS Config custom rule option takes much of this work off your hands.

Option C is correct. The most efficient way to monitor your resource use, such as monitoring your Elasticache Memcached node autoscaling, is to leverage the AWS Config custom rule capability.

Option D is incorrect. There is no elasticache-memcached-fanout-check AWS Config managed rule.  
  
**Reference:**

Please see the AWS Config developer guide titled **Evaluating Resources with Rules** (<https://docs.aws.amazon.com/config/latest/developerguide/evaluate-config.html>), the AWS Config API reference titled **ConfigRule** (<https://docs.aws.amazon.com/config/latest/APIReference/API_ConfigRule.html>), the AWS Config developer guide titled **AWS Config Managed Rules** (<https://docs.aws.amazon.com/config/latest/developerguide/evaluate-config_use-managed-rules.html>), the AWS Config developer guide titled **List of AWS Config Managed Rules** (<https://docs.aws.amazon.com/config/latest/developerguide/managed-rules-by-aws-config.html>), and the Amazon Elasticache for Redis user guide titled **Comparing Memcached and Redis** (<https://docs.aws.amazon.com/AmazonElastiCache/latest/red-ug/SelectEngine.html>)

Question: 41

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Select a collection system that handles the frequency, volume, and source of data**

**Domain:** Collection

**Question text**:

You work as a data analytics specialist for a mobile software company that is moving its customer database from their on-premises database instances to DMS in their AWS account. You are on the team creating the Database Migration Service (DMS) tasks to move your data tables and views to your RDS instance.

Which option is the most efficient and performant method of migrating your database resources?

1. Create a DMS task that captures ongoing changes to your database tables and views after you complete your initial (full-load) migration to your RDS instance.
2. Create a DMS task that first does a full-load of your database and then starts a CDC after the full-load completes.
3. Create a DMS task that does a full-load only task.
4. Create a DMS task that performs an ongoing replication.

**Answer:** C

**Explanation:**

Option A is incorrect. This option describes the DMS Change Data Capture (CDC) task. The CDC process only migrates tables, it will not migrate views.

Option B is incorrect. This option describes first performing a full load and then starting a CDC. With this approach only tables will be migrated, not views.

Option C is correct. The only way to migrate your tables and views using DMS is to use the full-load only task.

Option D is incorrect. An ongoing replication is another name for a DMS Change Data Capture CDC task.  
  
**Reference:**

Please see the AWS Database Migration Service user guide titled **Creating Tasks for Ongoing Replication Using AWS DMS** (<https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Task.CDC.html>), and the AWS Database Migration Service user guide titled **Using Table Mapping to Specify Task Settings** (<https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Tasks.CustomizingTasks.TableMapping.html#CHAP_Tasks.CustomizingTasks.TableMapping.SelectionTransformation>)

Question: 42

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Select a collection system that handles the frequency, volume, and source of data**

**Domain:** Collection

**Question text**:

You work as a data analytics specialist for a multinational publishing conglomerate that has just purchased another publishing firm in mid-April of 2020. You and your team are now in the process of integrating the newly acquired publisher’s PostgreSQL datastore source endpoint into your RDS PostgreSQL target endpoint using DMS. The acquisition migration will take several months to complete so you’ll need to perform ongoing replication from the source endpoint to the target endpoint for several months until the acquired firm's systems are decommissioned.

You have been instructed to start your DMS Change Data Capture (CDC) task from the starting point of 2 weeks after the acquisition migration began, which is May 15, 2020. What CDC task should you use to accomplish your ongoing replication activity?

1. Perform your replication of the source PostgreSQL endpoint from a CDC recovery checkpoint generated on 5/15/2020.
2. Perform your replication of the source PostgreSQL endpoint from a custom start time of the timestamp 5/15/2020.
3. Perform your replication of the source PostgreSQL endpoint from the native start point in the purchased firm’s PostgreSQL transaction log at the timestamp of 5/15/2020.
4. Create a DMS task that does a full-load only task starting on 5/15/2020.

**Answer:** A

**Explanation:**

Option A is correct. When using DMS to migrate a PostgreSQL source endpoint you need to use a CDC native start point. To do this you can use a CDC recovery checkpoint in your source endpoint.

Option B is incorrect. When using DMS to migrate a PostgreSQL source endpoint you need to use a CDC native start point. When using DMS with PostgreSQL as a source endpoint custom start times are not supported.

Option C is incorrect. When using DMS to migrate a PostgreSQL source endpoint you need to use a CDC native start point. To do this you can use a CDC recovery checkpoint in your source endpoint, but not a timestamp.

Option D is incorrect. A full-load only CDC task on the start day will not record all subsequent days activity while your firm works over the several months to decommission the acquired firm’s systems.  
  
**Reference:**

Please see the AWS Database Migration Service user guide titled **Creating Tasks for Ongoing Replication Using AWS DMS** (<https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Task.CDC.html>), and the AWS Database Migration Service user guide titled **Creating a Task** (<https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Tasks.Creating.html>)

Question: 43

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine an appropriate system for cataloging data and managing metadata**

**Domain:** Storage and Data Management

**Question text**:

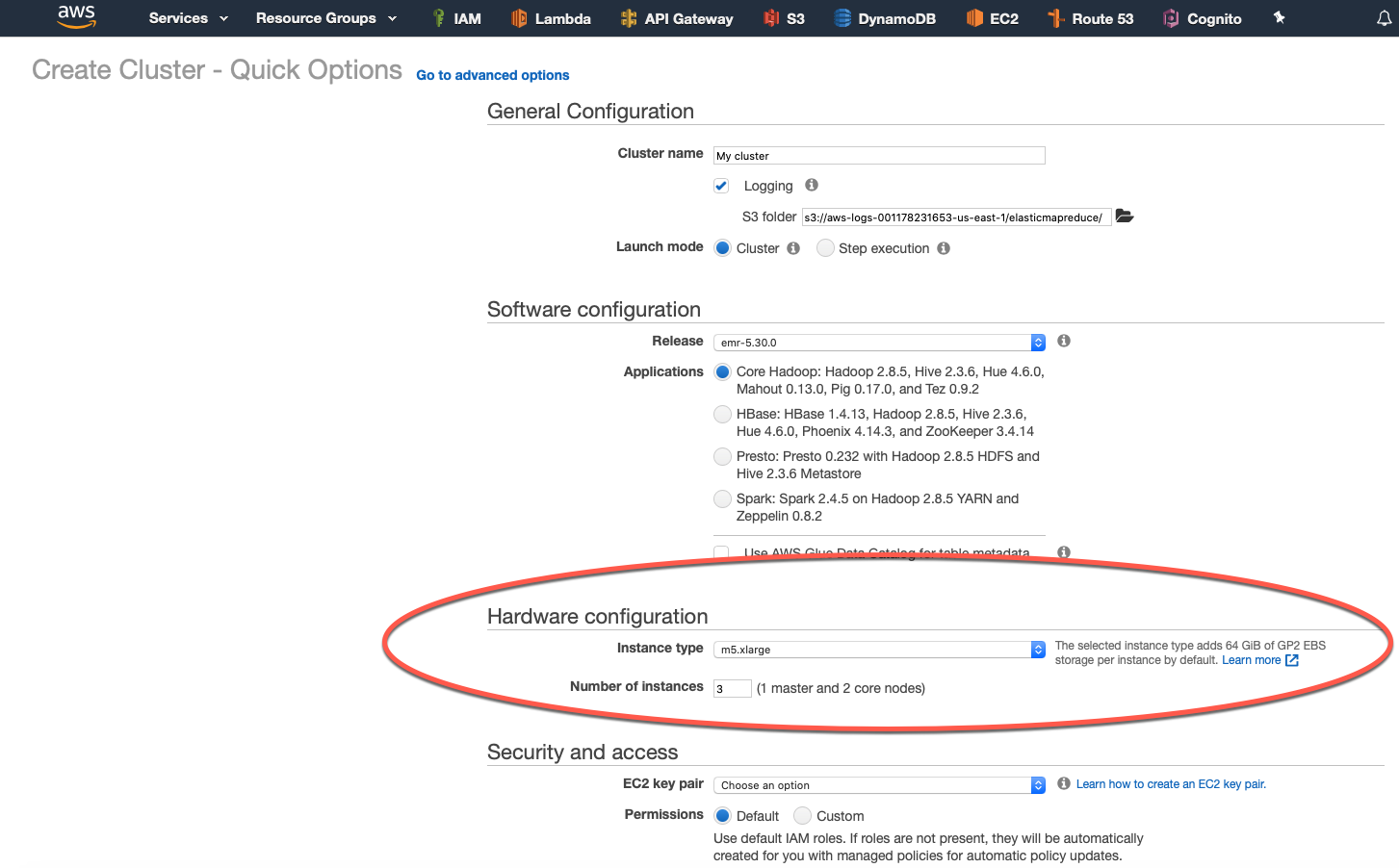
You work as a data analytics specialist for a regional bank that is building a data lake to hold information about their customers, their customers purchasing habits, and their customers accounts. They wish to use this data lake as the source for their real-time analytics that identifies potential fraudulent customer purchases as close to immediately as possible.

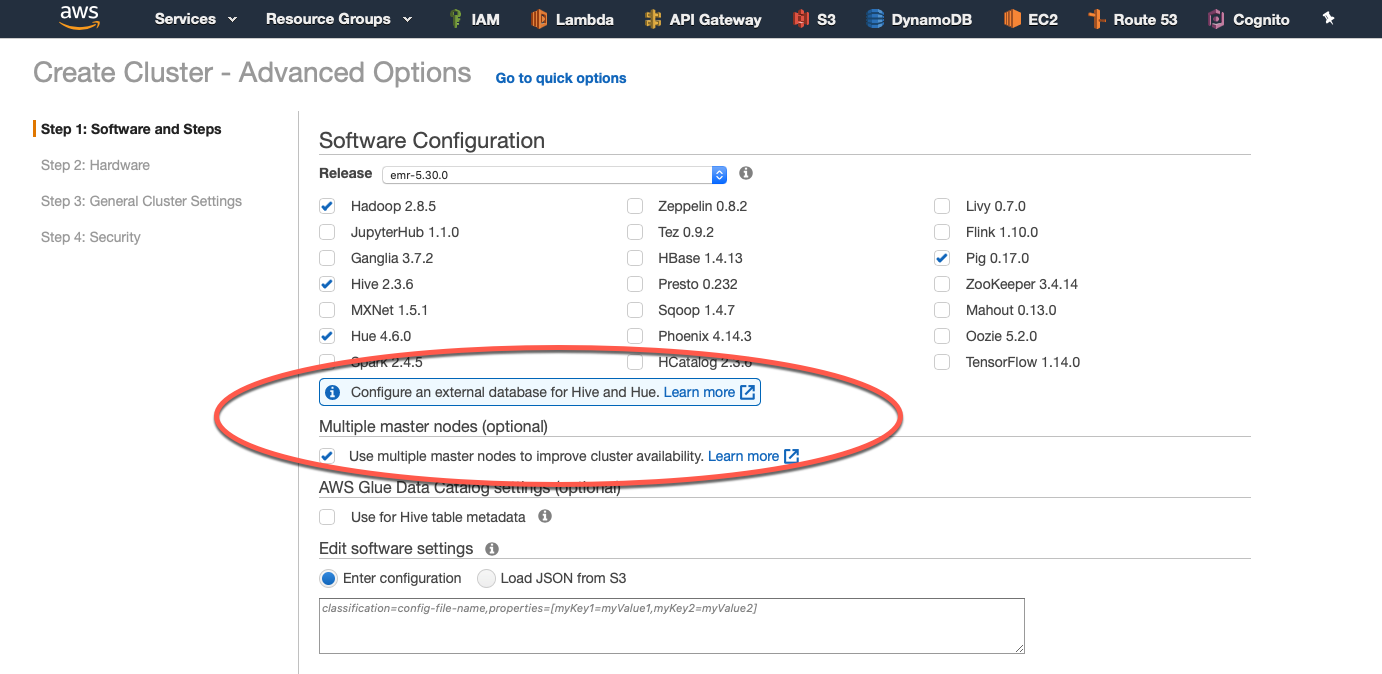
Due to the high availability and response time requirements of your data lake you have decided to use EMR as your big data platform. What features of EMR can you use to make your EMR cluster highly available in the most cost effective manner?

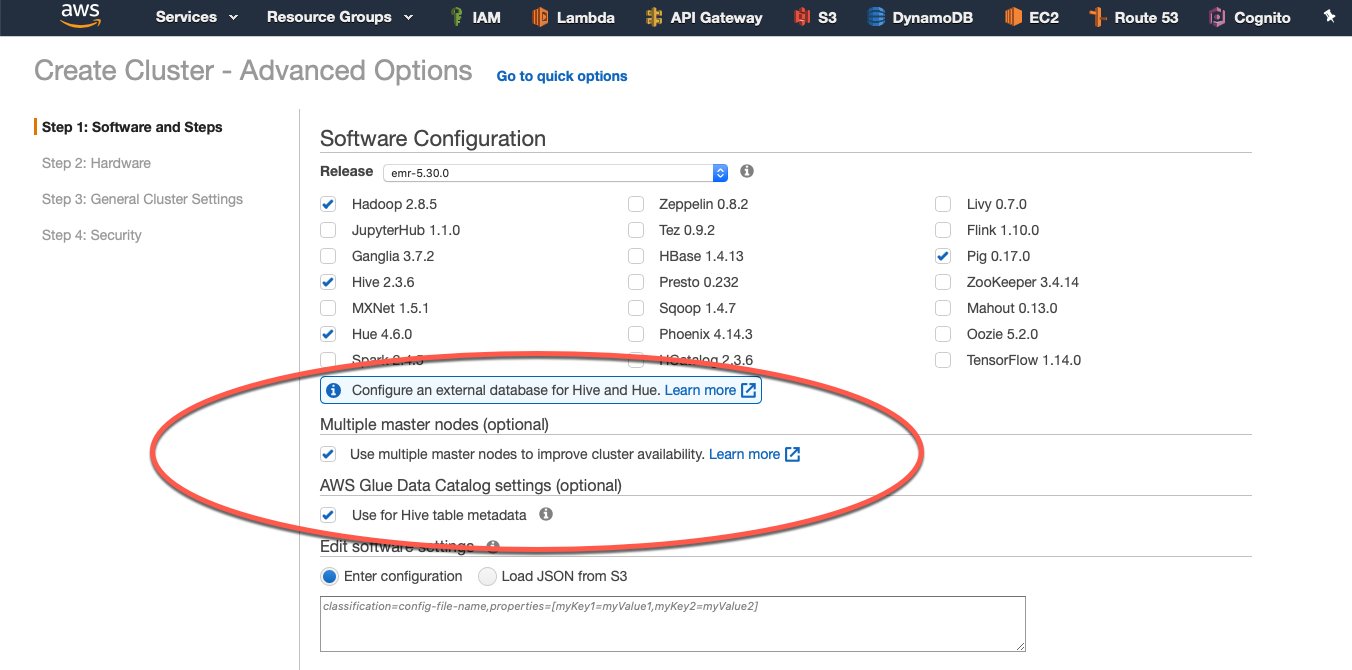
1. Create your EMR cluster with the default master and core node configuration. Use the Hadoop, Hive, Hue, and Pig software on your EMR cluster, taking advantage of Hadoop’s high availability feature.
2. Create your EMR cluster with multiple master nodes using the default Hive metastore configuration. Use the Hadoop, Hive, Hue, and Pig software on your EMR cluster, taking advantage of Hadoop’s high availability feature.
3. Create your EMR cluster with multiple master nodes using Glue as your external Hive metastore configuration. Use the Hadoop, Hive, Hue, and Pig software on your EMR cluster, taking advantage of Hadoop’s high availability feature.
4. Create your EMR cluster with multiple master nodes using Aurora as your external Hive metastore configuration. Use the Hadoop, Hive, Hue, and Pig software on your EMR cluster, taking advantage of Hadoop’s high availability feature.

**Answer:** C

**Explanation:**

Option A is incorrect. The default EMR cluster configuration uses one master node and two core core nodes. With only one master node your EMR cluster is not configured to be highly available. If the master node fails, your cluster will be inoperable.   


Option B is incorrect. The multiple master option will give you 3 master nodes, one active and two standby master nodes. If the active node fails, EMR automatically fails over to one of the standby nodes, making it the new master. However, if you use the default configuration for your Hive metastore, specifically storing the metadata in a MySQL instance on the active master node’s file system, if the active master fails, you will lose your Hive metadata.  


Option C is correct. The multiple master option will give you 3 master nodes, one active and two standby master nodes. If the active node fails, EMR automatically fails over to one of the standby nodes, making it the new master. Selecting Glue as your external Hive metastore prevents the active master node from becoming a single point of failure.  


Option D is incorrect. While using Aurora to hold your external Hive metastore is a viable option, it is not as cost effective as using Glue as your external Hive metastore.  
  
**Reference:**

Please see the Amazon EMR management guide titled **Plan and Configure Master Nodes** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-ha.html>), the Amazon EMR management guide titled **Supported Applications and Features** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-ha-applications.html>), the Amazon EMR release guide titled **Configuring an External Metastore for Hive** (<https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-metastore-external-hive.html>), the **Apache Hadoop HDFS High Availability page** (<https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HDFSHighAvailabilityWithNFS.html>), the **Apache Hadoop ResourceManager High Availability page** (<https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/ResourceManagerHA.html>), and the **Amazon EMR overview page** (<https://aws.amazon.com/emr/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc>)

Question: 44

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Automate and operationalize a data processing solution**

**Domain:** Processing

**Question text**:

You work as a data scientist for an automobile manufacturer that has an EMR cluster that is used to populate your S3 data lake with automobile information such as sales across regions and countries, parts source partner performance, etc.

You need to create Apache Spark applications and run interactive queries on your EMR cluster to gain insight into your company’s sales and partner performance with minimal effort. You also need to give your fellow data scientists access to create Apache Spark applications.

Which option gives you the simplest, most cost effective operation for your data processing solution?

1. Create a SageMaker Jupyter notebook instance and attach it to your Spark cluster running in your EMR configuration.
2. Use Athena to directly query your data using Spark applications.
3. Create a stand-alone EMR notebook and use Jupyter notebooks to run your Spark applications.
4. Create EMR notebooks within your EMR cluster and use Jupyter notebooks to run your Spark applications.

**Answer:** D

**Explanation:**

Option A is incorrect. You could run your Spark applications from within your SageMaker Jupyter notebooks, but this option is not as cost effective as using an EMR notebook since you already have an EMR cluster.

Option B is incorrect. Athena cannot run Spark applications. Athena runs SQL queries, but cannot run Spark applications.

Option C is incorrect. An EMR notebook must execute its code from within an EMR cluster. You cannot run an EMR notebook as “stand-alone.”

Option D is correct. Running your Spark applications in EMR notebooks running on your EMR cluster’s master node(s) is the most cost effective option listed.  
  
**Reference:**

Please see the Amazon EMR management guide titled **Using Amazon EMR Notebooks** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-managed-notebooks.html>), the **Amazon EMR FAQs page** (<https://aws.amazon.com/emr/faqs/>), the Amazon Machine Learning blog titled **Build Amazon SageMaker notebooks backed by Spark in Amazon EMR** (<https://aws.amazon.com/blogs/machine-learning/build-amazon-sagemaker-notebooks-backed-by-spark-in-amazon-emr/>), and the **Amazon Athena FAQs page** (<https://aws.amazon.com/athena/faqs/>)

Question: 45

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Determine appropriate data processing solution requirements**

**Domain:** Processing

**Question text**:

You work as a data analytics specialist for a large logistics company that ships containers all over the world via ships, planes, and trucks. These shipments need to be tracked and optimal routes need to be determined by solving the “traveling salesman problem” operations research problem. The traveling salesman problem is defined as "Given a list of locations and the distances between each pair of locations, what is the shortest possible route that visits each location and returns to the origin location?"

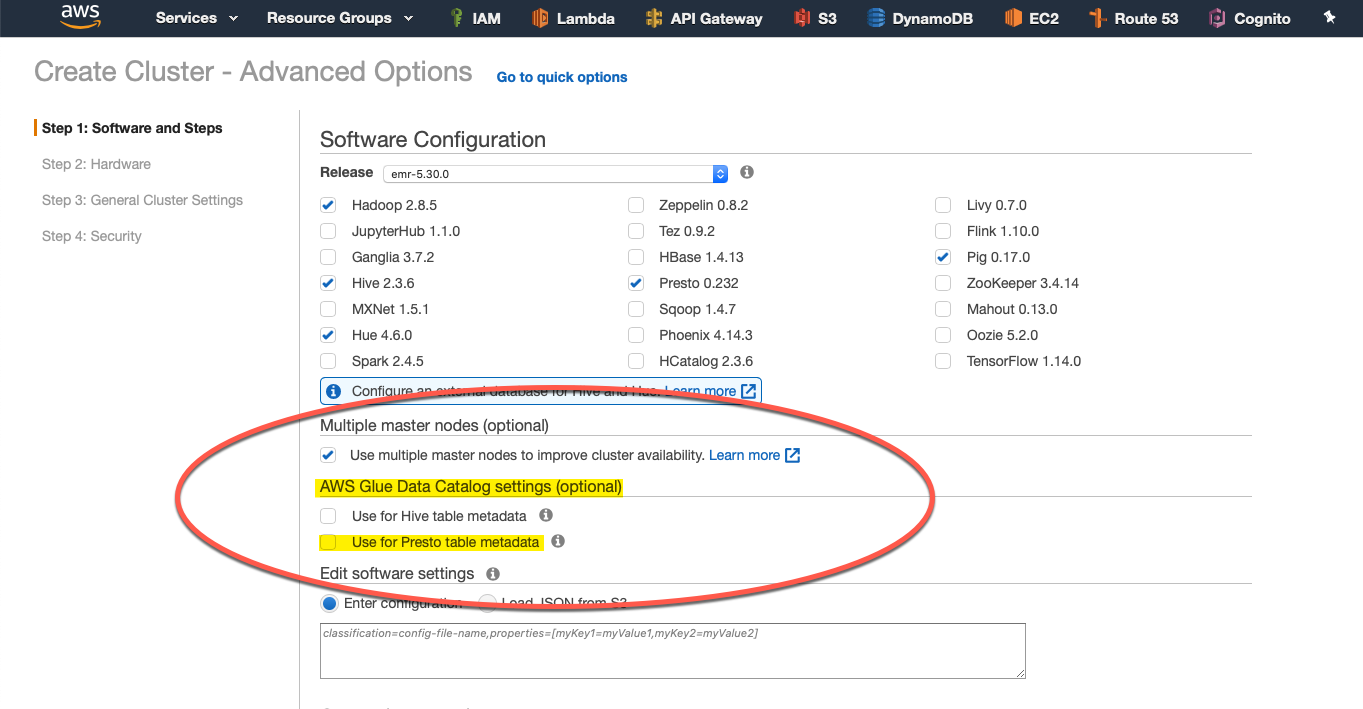
You need data from your containers, their transport facility (plane, boat, truck) and other key data for your real-time analytics applications that attempt to solve the logistics optimization problem for each container. These applications also need to update the optimized routes based on real-time changes in route attributes, such as obstructions, etc.

How should you define your data processing solution so that it is highly available and provides real-time query capabilities across your very large data lake?

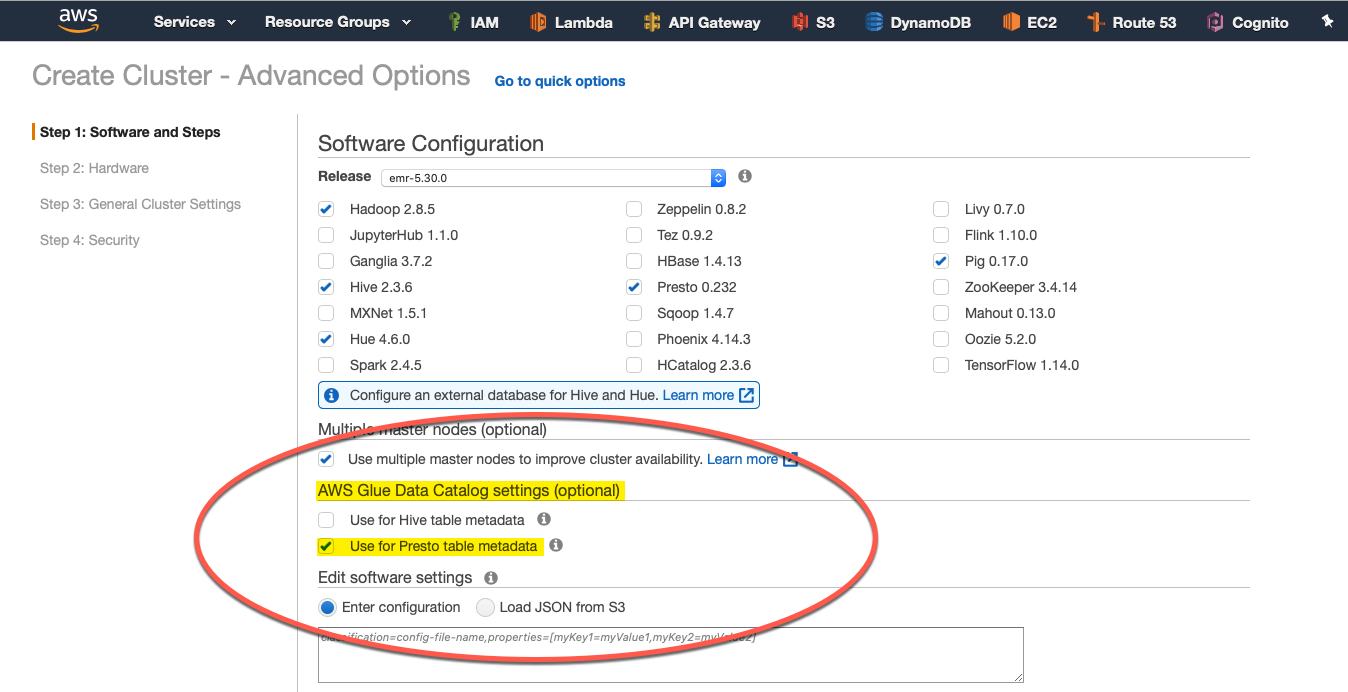
1. Use EMR in multi-master mode with at least 2 master nodes for high availability. Use Presto for real-time query capability with the default metastore configuration for persistent metastore.
2. Use EMR in multi-master mode with 1 master node for high availability. Use Presto for real-time query capability with the default metastore configuration for persistent metastore.
3. Use EMR in multi-master mode with 5 master nodes for high availability. Use Presto for real-time query capability with Hive for the Presto external table metastore for persistent metastore.
4. Use EMR in multi-master mode with 3 master nodes for high availability. Use Presto for real-time query capability with Glue for the Presto external table metastore for persistent metastore.

**Answer:** D

**Explanation:**

Option A is incorrect. When using the multi-master node feature of EMR you are given 3 master nodes. You cannot choose another number of master nodes as this option suggests “at least 2 master nodes.” Also, if you use the default metastore configuration for Presto you will not have a persistent Presto metastore, you need to select to use Glue for the Presto external table metastore.  


Option B is incorrect. When using the multi-master node feature of EMR you cannot select 1 master node. Additionally, having only 1 master node doesn’t give you high availability for your cluster since if your only master node fails your cluster becomes unavailable. Also, if you use the default metastore configuration for Presto you will not have a persistent Presto metastore, you need to select to use Glue for the Presto external table metastore.

Option C is incorrect. Three master nodes will give you high availability for your EMR cluster, there is no option for more than 3 master nodes. Also, Hive is not a valid option for an external Presto metastore. The correct option for Presto external metastore is Glue.  


Option D is correct. Three master nodes will give you high availability for your EMR cluster. The correct option for Presto external metastore is Glue.  
  
**Reference:**

Please see the Amazon EMR management guide titled **Plan and Configure Master Nodes** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-ha.html>), the Amazon EMR management guide titled **Supported Applications and Features** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-ha-applications.html>), the Amazon EMR release guide titled **Using Presto with the AWS Glue Data Catalog** (<https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-presto-glue.html>), the Amazon EMR release guide titled **Considerations with Presto on Amazon EMR** (<https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-presto-considerations.html>), the Wikipedia page titled **Travelling salesman problem** (<https://en.wikipedia.org/wiki/Travelling_salesman_problem#:~:text=The%20travelling%20salesman%20problem%20(also,NP%2Dhard%20problem%20in%20combinatorial>)

Question: 46

**Main​ ​Topic​ ​:​** Data Analytics

**Sub​ ​Topic​ ​:​ Design a solution for transforming and preparing data for analysis**

**Domain:** Processing

**Question text**:

You work as a data analytics specialist for a company that processes massive amounts of data through your EMR cluster. You need to transform this data with as little latency as possible. The data arrives in Avro files in one of your S3 buckets at a rate of hundreds of files (from approximately 100 to up to 500) per second. You need to process these files as close to real-time as you can.

What option is the most efficient and most performant that gives you the transformation of your data for analysis at the processing rate you need?

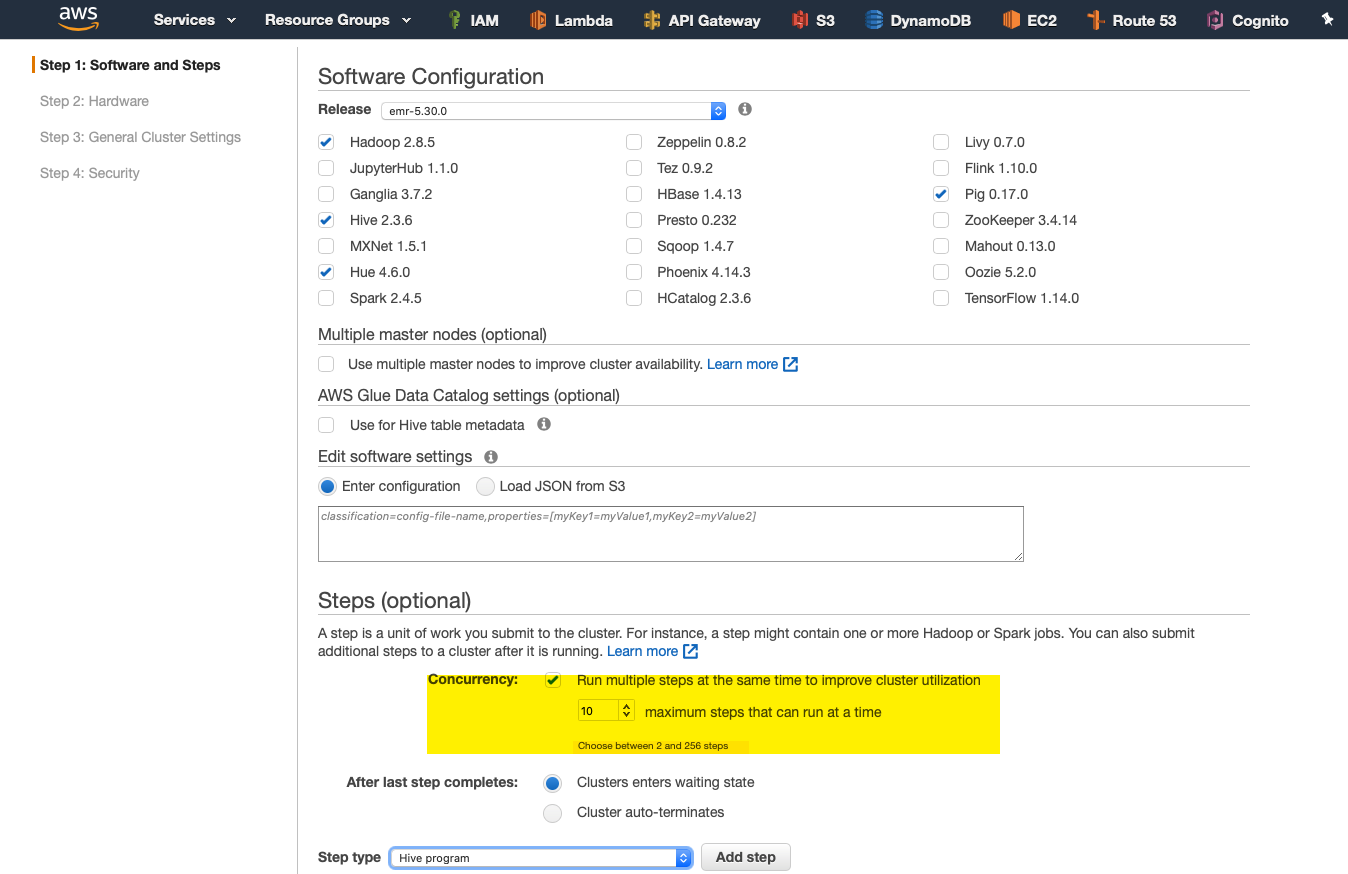
1. Create a step in your EMR cluster to execute a Hue script to transform the file data as each Avro file arrives.
2. Create a step in your EMR cluster to execute a Mahout script to transform the file data as each Avro file arrives.
3. Create a step in your EMR cluster to execute a Hive script with concurrency set to 500 to transform the file data as the Avro files arrive.
4. Create a step in your EMR cluster to execute a Hive script with concurrency set to 250 to transform the file data as the Avro files arrive.

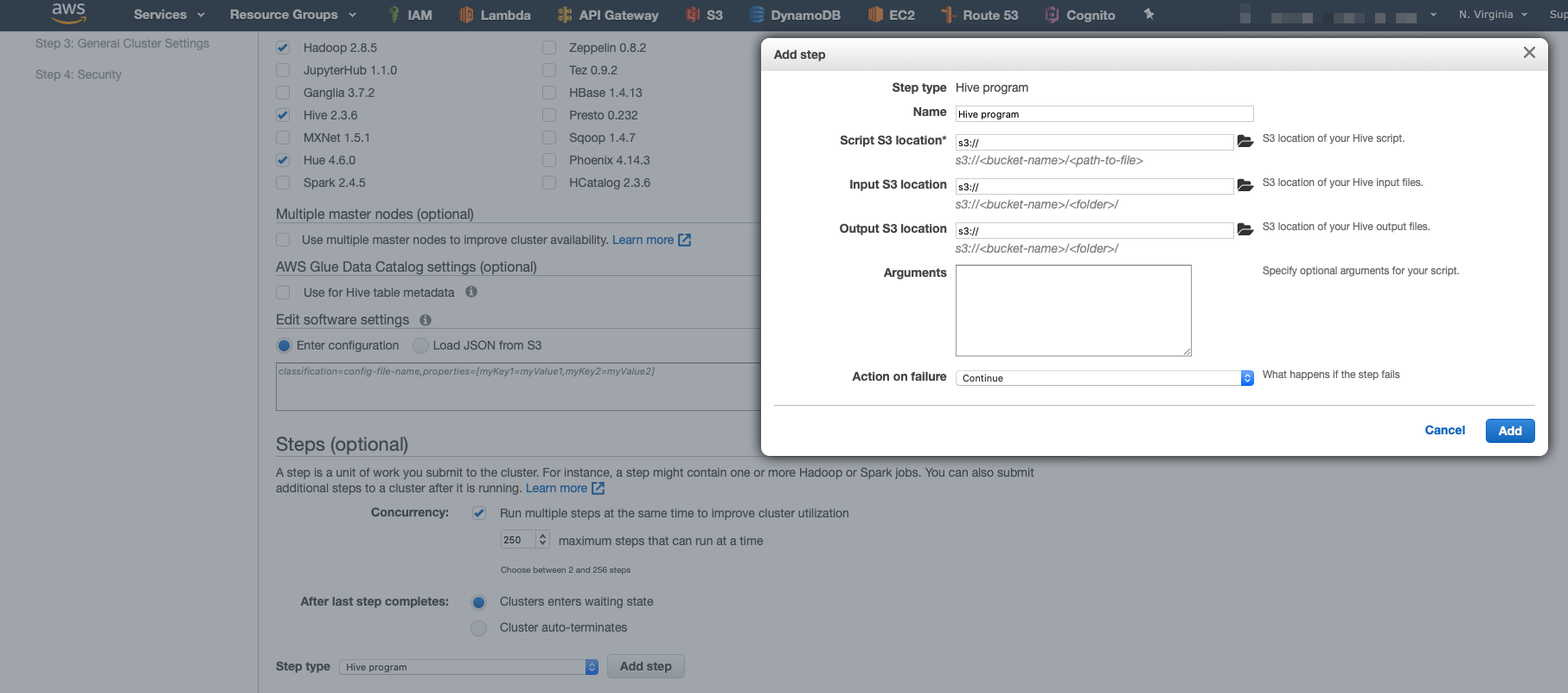
**Answer:** D

**Explanation:**

Option A is incorrect. Hue is a web based user interface for Hadoop. It is not a scripting language that you can use for transformation of data.

Option B is incorrect. Mahout is a machine learning framework for Hadoop. It is not a scripting language that you would use for transformation of data.

Option C is incorrect. When using concurrency with Hive scripts you can choose from 2 to 256 steps to run at the same time. So you can’t choose 500 as your concurrency value.  


Option D is correct. Concurrency, set to 250, with a Hive script will allow you to process your Avro files as they arrive using your Hive script running many instances in parallel.  
  
  
**Reference:**

Please see the Amazon EMR management guide titled **Step 4: Process Data By Running The Hive Script as a Step** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs-process-sample-data.html>), the AWS What’s New page titled **Amazon EMR now allows you to run multiple steps in parallel, cancel running steps, and integrate with AWS Step Functions** (<https://aws.amazon.com/about-aws/whats-new/2019/11/amazon-emr-now-allows-you-to-run-multiple-steps-in-parallel-cancel-running-steps-and-integrate-with-aws-step-functions/>), the Amazon EMR release guide titled **Apache Mahout** (<https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-mahout.html>), and the Amazon EMR release guide titled **Hue** (<https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hue.html>)