**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>)