**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:** Storage and Data Management

**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:** Storage and Data Management

**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>)