**AWS Certified Data Analytics** - **Specialty Practice Questions**

**Requirement**: Share 10 DA Specialty practice questions.

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

**Delivery Timeline**: April-4

**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 DA Specialty 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 the number of expected answers. e.g. (Select TWO) or (Select THREE) etc. For single answers this is NOT required>

1. Option A...
2. Option B...
3. Option C...
4. 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].

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

**DA 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 financial services company that has several relational databases, data warehouses, and NoSQL databases that hold transactional information about their financial trades and operational activities. The company wants to manage their financial counterparty risk through using their real-time trading/operational data to perform risk analysis and build risk management dashboards.

You need to build a data repository that combines all of these disparate data sources so that your company can perform their Business Intelligence (BI) analysis work on the complete picture of their risk exposure.

What collection system best fits this use case?

1. Financial data sources data -> Kinesis Data Firehose -> S3 -> Glue -> S3 Data Lake -> Athena
2. Financial data sources data -> Kinesis Data Firehose -> Kinesis Data Analytics -> Kinesis Data Firehose -> Redshift -> QuickSight
3. Financial data sources data -> Database Migration Service -> S3 -> Glue -> S3 Data Lake -> Athena
4. Financial data sources data -> Kinesis Data Streams -> Kinesis Data Analytics -> S3 Data Lake -> QuickSight

**Answer:** C

**Explanation:**

Option A is incorrect. This data collection system architecture is best suited to batch consumption of stream data. You are trying to build a real-time financial risk management analytics collection architecture. You have several databases and data warehouses generating your data stream from their changed data. This approach is called ongoing replication or change data capture (CDC) within the Database Migration Service. A collection architecture using the Database Migration Service will be the most optimal for this use case.

Option B is incorrect. This data collection system architecture is suited to real-time consumption of data, but a collection architecture using the Database Migration Service would better fit this use case. You have several databases and data warehouses generating your data stream from their changed data. This approach is called ongoing replication or change data capture (CDC) within the Database Migration Service. A collection architecture using the Database Migration Service will be the most optimal for this use case.

Option C is correct. This type of data collection infrastructure is best used for streaming transactional data from existing relational data stores. You create a task within the Database Migration Service that collects ongoing changes within your various operational data stores, an approach called ongoing replication or change data capture (CDC). These changes are streamed to an S3 bucket where a Glue job is used to transform the data and move it to your S3 data lake.

Option D is incorrect. Kinesis Data Analytics cannot write directly to S3; it only writes to a Kinesis data stream, a Kinesis Data Firehose delivery stream, or a Lambda function. Also, this collection architecture does not take advantage of the Database Migration Service ongoing replication or change data capture (CDC) technique.

**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>), the AWS Schema Conversion Tool user guide titled **What Is the AWS Schema Conversion Tool?** (<https://docs.aws.amazon.com/SchemaConversionTool/latest/userguide/CHAP_Welcome.html>),

the Amazon Kinesis Data Analytics for SQL Applications developer guide titled **Configuring Application Output**

(<https://docs.aws.amazon.com/kinesisanalytics/latest/dev/how-it-works-output.html>), the AWS Streaming Data page titled **What is Streaming Data?** (<https://aws.amazon.com/streaming-data/>), the **AWS Database Migration Service FAQs** (<https://aws.amazon.com/dms/faqs/>), the **Amazon Kinesis Data Analytics FAQs (**[**https://aws.amazon.com/kinesis/data-analytics/faqs/**](https://aws.amazon.com/kinesis/data-analytics/faqs/)), the **Amazon Kinesis Data Streams FAQs (**[**https://aws.amazon.com/kinesis/data-streams/faqs/**](https://aws.amazon.com/kinesis/data-streams/faqs/)), , the Amazon Kinesis Data Firehose developer guide titles **What is Amazon Kinesis Data Firehose? (**[**https://docs.aws.amazon.com/firehose/latest/dev/what-is-this-service.html#data-flow-diagrams**](https://docs.aws.amazon.com/firehose/latest/dev/what-is-this-service.html#data-flow-diagrams)), the AWS Glue developer guide titled **AWS Glue Concepts** (<https://docs.aws.amazon.com/glue/latest/dg/components-key-concepts.html>), and the **Amazon Kinesis Data Firehose FAQs (**[**https://aws.amazon.com/kinesis/data-firehose/faqs/**](https://aws.amazon.com/kinesis/data-firehose/faqs/))

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 on a project where you have two large tables (orders and products) that you need to load into Redshift from one of your S3 buckets. Your table files, which are both several million rows large, are currently on an EBS volume of one of your EC2 instances in a directory titled $HOME/myredshiftdata.

Since your table files are so large, what is the most efficient approach to first copy them to S3 from your EC2 instance?

1. Load your orders.tbl and products.tbl using the command: ‘aws s3 cp $HOME/myredshiftdata s3://dataanalytics/myredshiftdata --recursive’
2. Load your orders.tbl and products.tbl by first splitting each tbl file into smaller parts using the command: ‘split -d -l 5000000 -a 4 orders.tbl orders.tbl’ and ‘split -d -l 10000000 -a 4 products.tbl products.tbl’
3. Load your orders.tbl and products.tbl by first getting a count of the number of rows in each using the commands: ‘wc -l orders.tbl’ and ‘wc -l products.tbl’. Then splitting each tbl file into smaller parts using the command: ‘split -d -l # -a 4 orders.tbl orders.tbl’ and ‘split -d -l # -a 4 products.tbl products.tbl’ where # is replaced by the result of your wc command divided by 4.
4. Load your orders.tbl and products.tbl by first getting a count of the number of rows in each using the commands: ‘wc -l orders.tbl’ and ‘wc -l products.tbl’. Then splitting each tbl file into smaller parts using the command: ‘split -d -l # -a 4 orders.tbl orders.tbl-’ and ‘split -d -l # -a 4 products.tbl products.tbl-’ where # is replaced by the result of your wc command divided by 4.

**Answer:** D

**Explanation:**

Option A is incorrect because using the commands in this answer you don’t reduce the size of your tbl files before attempting to move them to S3. Also, when you attempt to move these files into Redshift from your S3 bucket the process will be less efficient because you haven’t split your files into more manageable sizes.

Option B is incorrect because when you attempt to split your files you haven’t determined the actual number of rows of each file. Therefore, your random selection of a split size will more than likely not be an efficient selection.

Option C is incorrect because your split command does not have a trailing ‘-’ at the end of the command. Therefore your smaller files on your S3 bucket will have names like ‘orders.tbl0001’ versus the more readable and manageable ‘orders.tbl-0001’ if you use a trailing ‘-’ in the split command.

Option D is correct because you have used the wc command to find the number of rows in each tbl file, and you have used the split command with the trailing ‘-’ to get the proper file name format on your S3 bucket in preparation for loading into Redshift.

**Reference:**

Please see the AWS Redshift Developer Guide titled **Tutorial: Loading Data from Amazon S3** (<https://docs.aws.amazon.com/redshift/latest/dg/tutorial-loading-data.html>), specifically step 2: Download the Data Files and Step 5: Run the Copy Commands where you’ll see that having the ‘-’ at the end of your split command will allow you to make your Redshift copy command more efficient.

Question: 3

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Processing

**Question text**:

You are working on a project where you need to perform real-time analytics on your application server logs. Your application is split across several EC2 servers in an auto-scaling group and is behind an application load balancer as depicted in this diagram:

[](https://www.draw.io/?page-id=6_lzF5P4isDOIafqsJO-&scale=auto#G1GNODpLc0DVlcFq7mT79w1ym0bAFvR-pw)

You need to perform some transformation on the log data, such as joining rows of data, before you stream the data to your real-time dashboard.

What is the most efficient and performant solution to aggregate your application logs?

1. Install the Kinesis Agent on your application servers to watch your logs and use Kinesis Data Firehose to stream the logs directly to S3. Use Kinesis Data Analytics queries to build your real-time analytics dashboard.
2. Write a Kinesis Data Streams producer application that reads the application logs and pushes the data directly into your Kinesis data stream. Use Kinesis Data Analytics queries to build your real-time analytics dashboard.
3. Install the Kinesis Agent on your application servers to watch your logs and ingest the log data. Write a Kinesis Data Analytics application that reads the application log data from the agent, performs the required transformations, and pushes the data into your Kinesis data output stream. Use Kinesis Data Analytics queries to build your real-time analytics dashboard.
4. Use a CloudWatch dashboard that uses your application’s CloudWatch logs as the data source.

**Answer:** C

**Explanation:**

Option A is incorrect because with this approach you don’t have a capability to perform the required transformations. You could write a lambda function to perform the transformations but the answer doesn’t specify these details.

Option B is incorrect because the answer is missing the Kinesis Agent part of the solution. You could write your Kinesis producer application to read the application log files, but using the Kinesis Agent is much more efficient.

Option C is correct. The Kinesis Agent ingests the application log data, the Kinesis Analytics application transforms the data, and Kinesis Analytics queries are used to build your dashboard.

Option D is incorrect since while a CloudWatch dashboard could be used to build this solution simply, it lacks the real-time capability. CloudWatch high-resolution metrics log in intervals of 1 second, 5 seconds, 10 seconds, 30 seconds, or multiples of 60 seconds. Also, this solution lacks the ability to perform the required transformations of the log data.

**Reference:**

Please see the **Amazon CloudWatch FAQs** (<https://aws.amazon.com/cloudwatch/faqs/>), 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 blog titled **Implement Serverless Log Analytics Using Kinesis Analytics** (<https://aws.amazon.com/blogs/big-data/implement-serverless-log-analytics-using-amazon-kinesis-analytics/>), and the **Amazon Kinesis Data Streams overview page** (<https://aws.amazon.com/kinesis/data-streams/>)

Question: 4

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Processing

**Question text**:

You are a data scientist on a team where you are responsible for ingesting IoT streamed data into a data lake for use in an EMR cluster. The data in the data lake will be used to allow your company to do business intelligence analytics on the IoT data. Due to the large amount of data being streamed to your application you will need to compress the data on the fly as you process it into your EMR cluster.

How would you most efficiently collect the data from your IoT devices?

1. Use the Kinesis REST API to get the IoT device records from the IoT devices and stream the data to your data lake through Kinesis Data Streams, then use Apache DistCp to move the data from S3 to your EMR cluster.
2. Use the AWS IoT service to get the device data from the IoT devices, use Kinesis Data Firehose to stream the data to your data lake, then use S3DistCp to move the data from S3 to your EMR cluster.
3. Use the Kinesis Producer Library to create a Kinesis producer application that reads the data from the IoT devices and stream the data to your data lake through Kinesis Data Streams, then use S3DistCp to move the data from S3 to your EMR cluster.
4. Use the Kinesis Client Library to get the device data from the IoT devices, use Kinesis Data Firehose to stream the data to your data lake, then use Apache DistCp to move the data from S3 to your EMR cluster.

**Answer:** B

**Explanation:**

Option A is incorrect because the Kinesis REST API is not the most efficient way to gather the IoT device data from your set of devices. Also, Apache DistCp does not offer the compression option that S3DistCp offers.

Option B is correct. The AWS IoT service ingests the device data, Kinesis Data Firehose streams the data to your S3 data lake, then the S3DistCp command is used to compress and move the data inno your EMR cluster

Option C is incorrect. The Kinesis Producer Library is not the most efficient way to gather the IoT device data from your set of devices.

Option D is incorrect. The Kinesis Client Library is used to consume Kinesis Stream data, not to produce data for consumption into the data stream. Also, Apache DistCp does not offer the compression option that S3DistCp offers.

**Reference:**

Please see the **AWS IoT overview page** (<https://aws.amazon.com/iot/>), the Amazon Premium Support Knowledge Center article titled **How can I copy large amounts of data from Amazon S3 into HDFS on my Amazon EMR cluster?**

(<https://aws.amazon.com/premiumsupport/knowledge-center/copy-s3-hdfs-emr/>), the Amazon EMR Release Guide titled **S3DistCp (s3-dist-cp)**

(<https://docs.aws.amazon.com/emr/latest/ReleaseGuide/UsingEMR_s3distcp.html>), the AWS Big Data blog titled **Seven Tips for Using S3DistCp on Amazon EMR to Move Data Efficiently Between HDFS and Amazon S3** (<https://aws.amazon.com/blogs/big-data/seven-tips-for-using-s3distcp-on-amazon-emr-to-move-data-efficiently-between-hdfs-and-amazon-s3/>), and the AWS Solutions page titled **Real-Time IoT Device Monitoring with Kinesis Data Analytics** (<https://aws.amazon.com/solutions/real-time-iot-device-monitoring-with-kinesis/>)

Question: 5

**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 rental car company that has fleets of rental cars across the globe. Each car is equipped with IoT sensors that report important information about the car’s functioning, location, service levels, mileage, etc.

You have been tasked with determining how rental efficiency has changed over time for fleets in certain cities across the US. This solution requires you to look back at several years of historical data collected by your IoT sensors.

Your management team wishes to perform Key Performance Indicator (KPI) analysis on the rental car data through visualization using business intelligence (BI) tools. They will use this analysis and visualization to make management decisions on how to keep their fleet of rental cars at optimum levels of service and use. They will also use the KPI analysis to assess the performance of their regional management teams for each city for which you collect data.

What collection system best fits this use case?

1. IoT device sensor data -> Kinesis Data Firehose -> S3 -> Glue -> S3 Data Lake -> Athena
2. IoT device sensor data -> Kinesis Data Firehose -> Kinesis Data Analytics -> Kinesis Data Firehose -> Redshift -> QuickSight
3. IoT device sensor data -> RDS -> Database Migration Service -> S3 -> Glue -> S3 Data Lake -> Athena
4. IoT device sensor data -> Kinesis Data Streams -> Kinesis Data Analytics -> S3 Data Lake -> QuickSight

**Answer:** A

**Explanation:**

Option A is correct. This data collection system architecture is best suited to batch consumption of stream data. Crawling the S3 data using Glue and then using a Glue job to write the data to an S3 data lake to then be queried by Athena allows you to produce aggregate data analytics. These data can help you build your KPI dashboard.

Option B is incorrect. This data collection system architecture is best suited to real-time consumption of data. Batch sensor data is better processed with a Glue ETL job versus a Kinesis Data Analytics application.

Option C is incorrect. This type of data collection infrastructure is best used for streaming transactional data from existing relational data stores. There is no need for an RDS instance in this data collection system since we can use a data lake to house the historical data and use Amazon Athena to query the data lake.

Option D is incorrect. Kinesis Data Analytics cannot write directly to S3; it only writes to a Kinesis data stream, a Kinesis Data Firehose delivery stream, or a Lambda function.

**Reference:**

Please see the Amazon Kinesis Data Analytics for SQL Applications developer guide titled **Configuring Application Output**

(<https://docs.aws.amazon.com/kinesisanalytics/latest/dev/how-it-works-output.html>), the AWS Streaming Data page titled **What is Streaming Data?** (<https://aws.amazon.com/streaming-data/>), the **AWS Database Migration Service FAQs** (<https://aws.amazon.com/dms/faqs/>), the **Amazon Kinesis Data Analytics FAQs (**[**https://aws.amazon.com/kinesis/data-analytics/faqs/**](https://aws.amazon.com/kinesis/data-analytics/faqs/)), the **Amazon Kinesis Data Streams FAQs (**[**https://aws.amazon.com/kinesis/data-streams/faqs/**](https://aws.amazon.com/kinesis/data-streams/faqs/)), , the Amazon Kinesis Data Firehose developer guide titles **What is Amazon Kinesis Data Firehose? (**[**https://docs.aws.amazon.com/firehose/latest/dev/what-is-this-service.html#data-flow-diagrams**](https://docs.aws.amazon.com/firehose/latest/dev/what-is-this-service.html#data-flow-diagrams)), the AWS Glue developer guide titled **AWS Glue Concepts** (<https://docs.aws.amazon.com/glue/latest/dg/components-key-concepts.html>), and the **Amazon Kinesis Data Firehose FAQs (**[**https://aws.amazon.com/kinesis/data-firehose/faqs/**](https://aws.amazon.com/kinesis/data-firehose/faqs/))

Question: 6

**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 mobile gaming company that is developing a new mobile gaming app that will need to handle thousands of messages per second arriving in your application data store. Due to the user interactivity of your game, all changes to the game datastore must be recorded with a before-change and after-change view of the data record. These data store changes will be used to deliver a near-real-time usage dashboard of the app for your management team.

What application collection system infrastructure best delivers these capabilities in the most performant and cost effective way?

1. Kinesis Firehose -> S3 -> EMR with Spark -> S3 -> Redshift -> QuickSight
2. DynamoDB -> DynamoDB Streams -> Lambda -> Kinesis Firehose -> Redshift -> QuickSight
3. Kinesis Firehose -> Aurora MySQL -> Lambda -> Kinesis Firehose -> Redshift -> QuickSight
4. Kinesis Data Streams -> Aurora MySQL -> Lambda->Kinesis Firehose -> Redshift -> QuickSight

**Answer:** B

**Explanation:**

Option A is incorrect because none of the collection systems listed easily allow for the before-change and after-change views of your applications data store changes. Also, there is no data store other than S3 in the listed collection system components. S3 is not the most cost effective data store for this type of application.

Option B is correct. Your application will write its game activity data to your DynamoDB table which will have DynamoDB streams enabled. DynamoDB Streams will record both the new and old (or before and after) images of any item in the DynamoDB table that is changed. Your Lambda function will be triggered by DynamoDB Streams. Your Lambda function will use the Firehose client to write to your Firehose stream. Firehose will stream your data to Redshift. Quicksite will visualize your data in near-real-time.

Option C is incorrect. Kinesis Firehose does not have the capability to write directly to Aurora. You would have to write your stream data to S3 then write a Lambda function, triggered on each write, to consume the data stream and then write the stream data to your Aurora data store. You could also use the Amazon Database Migration service to move your data from S3 to Aurora. Also, you would have to write custom code to record the before-change information.

Option D is incorrect. Kinesis Data Streams does not have the capability to write directly to Aurora. You would have to write a Kinesis consumer client using the Kinesis Consumer Library (KCL) to consume the data stream and then write the stream data to your Aurora data store. Also, you would have to write custom code to record the before-change information.

**Reference:**

Please see the Amazon DynamoDB developer guide titled **Capturing Table Activity with DynamoDB Streams**

(<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Streams.html#Streams.Processing>), the Medium.com article titled **Data Transfer Dynamodb to Redshift**

(<https://medium.com/@ananthsrinivas/data-transfer-dynamodb-to-redshift-5424d7fdf673>), the **Amazon Redshift overview page** (<https://aws.amazon.com/redshift/>), the AWS Database blog titled **Stream data into an Aurora PostgreSQL Database using AWS DMS and Amazon Kinesis Data Firehose** (<https://aws.amazon.com/blogs/database/stream-data-into-an-aurora-postgresql-database-using-aws-dms-and-amazon-kinesis-data-firehose/>), the AWS Database blog titled **Capturing Data Changes in Amazon Aurora Using AWS Lambda**

(<https://aws.amazon.com/blogs/database/capturing-data-changes-in-amazon-aurora-using-aws-lambda/>), the **Kinesis Data Firehose overview page** (<https://aws.amazon.com/kinesis/data-firehose/>), and the **Kinesis Data Streams overview page** (<https://aws.amazon.com/kinesis/data-streams/>)

Question: 7

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Storage and Data Management

**Question text**:

You are a data scientist working for an online retail electronics chain. Their website receives very heavy traffic during certain months of the year, but these heavy traffic periods fluctuate over time. Your firm wants to get a better understanding of these patterns. Therefore, they have decided to build a traffic prediction machine learning model based on click-stream data.

Your task is to capture the click-stream data and store it in S3 for use as training and inference data in the machine learning model. You have built a streaming data capture system using Kinesis Data Streams and its Kinesis Producer Library (KPL) for your click-stream data capture component. You are using collection batching in your KPL code to improve performance of your collection system. Exception and failure handling is very important to your collection process, since losing click-stream data will compromise the integrity of your machine learning model data.

How can you best handle failures in your KPL component?

1. For each record processed by your KPL component trigger a Lambda function that ensures proper sequencing of the records processed
2. Kinesis Data Streams synchronously replicates your data across three availability zones. Take advantage of this to recover from failed record processing with retry logic.
3. With the KPL PutRecords operation, if a put fails, the record is automatically put back into the KPL buffer and retried.
4. With the KPL PutRecords operation, if a put fails, the record is automatically rolled back, giving you the option to use retry logic in your KPL code.

**Answer:** C

**Explanation:**

Option A is incorrect because this implementation would be very inefficient. Also, you would be writing logic that the KPL gives you in its PutRecords operation.

Option B is incorrect. While Kinesis Data Streams does synchronously replicate your data across three availability zones, this capability would not give you the opportunity to recover from failed record puts into the stream since the failed records would not be replicated across the three availability zones.

Option C is correct. You would use the Kinesis Producer Library (KPL) PutRecords method in your KPL code to send click-stream records into your Kinesis Data Streams stream. The KPL PutRecords automatically adds any failed records back into the KPL buffer so it can be retried.

Option D is incorrect. The KPL PutRecords automatically adds any failed records back into the KPL buffer so it can be retried. You don’t need to implement retry logic in your code since the failed record is placed back into the KPL buffer. Your normal buffer processing logic will process the KPL buffer data without changes needed for retry.

**Reference:**

Please see the Amazon Kinesis Streams developer guide titled **KPL Key Concepts**

(<https://docs.aws.amazon.com/streams/latest/dev/kinesis-kpl-concepts.html>), the Amazon Kinesis 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 Streams developer guide titled **KPL Retries and Rate Limiting** (<https://docs.aws.amazon.com/streams/latest/dev/kinesis-producer-adv-retries-rate-limiting.html>), the AWS Big Data blog titled **Implementing Efficient and Reliable Producers with the Amazon Kinesis Producer Library**

(<https://aws.amazon.com/blogs/big-data/implementing-efficient-and-reliable-producers-with-the-amazon-kinesis-producer-library/>), the **AWS Real-time Analytics on AWS overview page** (<https://aws.amazon.com/big-data/real-time-analytics-featured-partners/>), and 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/**](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/))

Question: 8

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Analysis and Visualization

**Question text**:

You are a data scientist working for a large city that has implemented an electric scooter ride sharing system. Each electric scooter is equipped with IoT sensors that report the scooter’s location, whether it is currently rented out, current renter, battery level, speed of travel, etc.

You have been tasked with determining scooter density of location throughout the city and redistributing scooters if some areas of the city are overpopulated with scooters while other areas are underpopulated. This solution requires real-time IoT data to be ingested into your data collection system.

Your management team wishes to perform real-time analysis on the scooter data through visualization using business intelligence (BI) tools. They will use this analysis and visualization to make management decisions on how to keep their fleet of scooters at optimum levels of service and use.

What collection system best fits this use case?

1. IoT device sensor data -> Kinesis Data Firehose -> S3 -> Glue -> S3 Data Lake -> Athena
2. IoT device sensor data -> Kinesis Data Firehose -> Kinesis Data Analytics -> Kinesis Data Firehose -> Redshift -> QuickSight
3. IoT device sensor data -> RDS -> Database Migration Service -> S3 -> Glue -> S3 Data Lake -> Athena
4. IoT device sensor data -> Kinesis Data Streams -> Kinesis Data Analytics -> S3 Data Lake -> QuickSight

**Answer:** B

**Explanation:**

Option A is incorrect. This data collection system architecture is better suited to batch consumption of stream data. Crawling the S3 data using Glue and then using a Glue job to write the data to an S3 data lake to then be queried by Athena would not allow you to produce real-time analytics. While Glue can process micro-batches, it does not handle streaming data.

Option B is correct. You can use a Kinesis Data Firehose stream to ingest the IoT data, then analyze and filter your data with Kinesis Data Analytics, then direct the analyzed data to another Kinesis Data Firehose stream to load the data into your data warehouse in RedShift. Finally, use QuickSight to produce your visualization and dashboard for your management team.

Option C is incorrect. This type of data collection infrastructure is best used for streaming transactional data from existing relational data stores. There is no need for an RDS instance in this data collection system since the data is transitory in nature.

Option D is incorrect. Kinesis Data Analytics cannot write directly to S3; it only writes to a Kinesis data stream, a Kinesis Data Firehose delivery stream, or a Lambda function.

**Reference:**

Please see the Amazon Kinesis Data Analytics for SQL Applications developer guide titled **Configuring Application Output**

(<https://docs.aws.amazon.com/kinesisanalytics/latest/dev/how-it-works-output.html>), the AWS Streaming Data page titled **What is Streaming Data?** (<https://aws.amazon.com/streaming-data/>), the **AWS Database Migration Service FAQs** (<https://aws.amazon.com/dms/faqs/>), the **Amazon Kinesis Data Analytics FAQs (**[**https://aws.amazon.com/kinesis/data-analytics/faqs/**](https://aws.amazon.com/kinesis/data-analytics/faqs/)), the **Amazon Kinesis Data Streams FAQs (**[**https://aws.amazon.com/kinesis/data-streams/faqs/**](https://aws.amazon.com/kinesis/data-streams/faqs/)), the Amazon Kinesis Data Firehose developer guide titles **What is Amazon Kinesis Data Firehose? (**[**https://docs.aws.amazon.com/firehose/latest/dev/what-is-this-service.html#data-flow-diagrams**](https://docs.aws.amazon.com/firehose/latest/dev/what-is-this-service.html#data-flow-diagrams)), the AWS Glue developer guide titled **AWS Glue Concepts** (<https://docs.aws.amazon.com/glue/latest/dg/components-key-concepts.html>), and the **Amazon Kinesis Data Firehose FAQs (**[**https://aws.amazon.com/kinesis/data-firehose/faqs/**](https://aws.amazon.com/kinesis/data-firehose/faqs/))

Question: 9

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Security

**Question text**:

You are a data scientist working for a medical services company that has a suite of apps available for patients and their doctors to share their medical data. These apps are used to share patient details, MRI and XRAY images, appointment schedules, etc. Because of the importance of this data and its inherent Personally Identifiable Information (PII), your data collection system needs to be secure and the system cannot suffer lost data, process data out of order, or duplicate data.

Which data collection system(s) gives you the security and data integrity your requirements demand? (SELECT 2)

1. Apache Kafka/Amazon MSK
2. SQS (FIFO)
3. SQS (Standard)
4. Kinesis Data Firehose
5. Kinesis Data Streams
6. DynamoDB Streams

**Answers:** B, F

**Explanation:**

Option A is incorrect. Apache Kafka/Amazon MSK allows you to process streaming data. It guarantees the correct order of delivery of your data messages, but it uses the “at-least-once” delivery method. At-least-once delivery means that the message will not be lost, but the message may be delivered to a consumer more than once.

Option B is correct. SQS in the FIFO mode guarantees the correct order of delivery of your data messages and it uses the “exactly-once” delivery method. Exactly-once means that all messages will be delivered exactly one time. No message losses, no duplicate data.

Option C is incorrect. SQS in the Standard mode does not guarantee the correct order of delivery of your data messages and it uses the “at-least-once” delivery method. At-least-once delivery means that the message will not be lost, but the message may be delivered to a consumer more than once.

Option D is incorrect. Kinesis Data Firehose does not guarantee the correct order of delivery of your data messages and it uses the “at-least-once” delivery method. At-least-once delivery means that the message will not be lost, but the message may be delivered to a consumer more than once.

Option E is incorrect. Kinesis Data Streams guarantees the correct order of delivery of your data messages, but it uses the “at-least-once” delivery method. At-least-once delivery means that the message will not be lost, but the message may be delivered to a consumer more than once.

Option F is correct. DynamoDB Streams guarantees the correct order of delivery of your data messages and it uses the “exactly-once” delivery method. Exactly-once means that all messages will be delivered exactly one time. No message losses, no duplicate data.

**Reference:**

Please see the **Amazon Managed Streaming for Apache Kafka (Amazon MSK)** overview page (<https://aws.amazon.com/msk/>), the Amazon Simple Queue Service developer guide titled **Amazon SQS Standard Queues** (<https://docs.aws.amazon.com/AWSSimpleQueueService/latest/SQSDeveloperGuide/standard-queues.html>), the Amazon Simple Queue Service developer guide titled **Amazon SQS FIFO (First-In-First-Out) Queues** (<https://docs.aws.amazon.com/AWSSimpleQueueService/latest/SQSDeveloperGuide/FIFO-queues.html>), the Amazon DynamoDB developer guide titled **Capturing Table Activity with DynamoDB Streams** (<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Streams.html>), the Amazon Kinesis Data Streams developer guide titled **Handling Duplicate Records** (<https://docs.aws.amazon.com/streams/latest/dev/kinesis-record-processor-duplicates.html>), the **Amazon Kinesis Data Firehose FAQs** (<https://aws.amazon.com/kinesis/data-firehose/faqs/>), and the **Amazon Kinesis Data Streams FAQs** (<https://aws.amazon.com/kinesis/data-streams/faqs/>)

Question: 10

**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 for a ski resort corporation. Your company is developing a lift ticket system for mobile devices that allows skiers and snowboarders to use their phone as their lift ticket. The ski resort corporation owns many resorts around the world. The lift ticketing system needs to handle users who move from resort to resort throughout any given time period. Resort customers can also purchase packages where they can ski or snowboard at a defined list (a subset of the total) of several different resorts across the globe as part of their package.

The storage system for the lift ticket mobile application has to handle large fluctuations in volume. The data collected from the devices and stored in the data store is small in size, but the system must provide the data at low latency and high throughput. It also has to authenticate users through their mobile device registered facial recognition service, so that users can’t share a lift ticket by sharing their mobile devices.

What storage system is the best fit for this system?

1. Neptune
2. RDS
3. DynamoDB
4. ElastiCache
5. Redshift
6. S3

**Answer:** C

**Explanation:**

Option A is incorrect. Neptune is a graph database engine optimized for storing billions of relationships and querying the graph data. Graph databases like Neptune are best leveraged for use cases like social networking, recommendation engines, and fraud detection, where you need to create relationships between data and quickly query these relationships. Your application is more operational in nature and therefore requires a database that fits that profile.

Option B is incorrect. While RDS is operational in nature, it is bounded by instance and storage size limits. Also, while offering a multi-availability zone (multi-AZ) capability, RDS does not scale globally as easily as DynamoDB. Therefore, DynamoDB is a better choice for your global availability requirements.

Option C is correct. DynamoDB offers single-digit millisecond latency at scale. It also scales horizontally for high performance at any size data store. Finally, DynamoDB offers global tables for multi-region replication of your data, which you’ll need for your globally dispersed user base and ski resort locations.

Option D is incorrect. ElastiCache is an in-memory caching system that, alone, would not have the persistence needed for your system.

Option E is incorrect. Redshift is a columnar storage database best used for data warehouse use cases. Since your application requires an operational data store, Redshift would not be the correct choice.

Option F is incorrect. S3 is used for structured and unstructured data. Querying S3 using Athena or Redshift Spectrum allow for relatively quick queries, but not fast enough for an operational application like your ski resort mobile application requirements.

**Reference:**

Please see the **Amazon DynamoDB FAQs** (<https://aws.amazon.com/dynamodb/faqs/>), the **Amazon Neptune overview** page (<https://aws.amazon.com/neptune/>), the Amazon DynamoDB developer guide titled **Global Tables: Multi-Region Replication with DynamoDB** (<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/GlobalTables.html>), the **Amazon RDS FAQs** (<https://aws.amazon.com/rds/faqs/>), the **Amazon S3 FAQs** (<https://aws.amazon.com/s3/faqs/>), the **Amazon Redshift FAQs** (<https://aws.amazon.com/redshift/faqs/>), and the **Amazon ElastiCache FAQs** (<https://aws.amazon.com/elasticache/faqs/>)

Question: 11

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Storage and Data Management

**Question text**:

You work for a mobile gaming company that has developed a word puzzle game that allows multiple users to challenge each other to complete a crossword puzzle type of game board. This interactive game works on mobile devices and web browsers. You have a world-wide user base that can play against each other no matter where each player is located.

You now need to create a leaderboard component of the game architecture where players can look at the daily point leaders for the day, week, or other timeframes. Each time a player accumulates points, the points counter for that player needs to be updated in real-time. This leaderboard data is transient in that it only needs to be stored for a limited duration.

Which of the following architectures best suits your data access and retrieval patterns using the simplest, most efficient approach?

1. Data sources -> Kinesis Data Streams -> Spark Streaming on EMR -> ElastiCache Redis -> DynamoDB
2. Data Sources -> Kinesis Data Firehose -> S3 -> Athena
3. Data sources -> Kinesis Data Streams -> Spark Streaming on EMR -> ElastiCache Memcached -> DynamoDB
4. Data sources -> Kinesis Data Streams -> Spark Streaming on EMR -> ElastiCache Redis
5. Data sources -> Kinesis Data Firehose -> Spark Streaming on EMR -> ElastiCache Redis -> S3

**Answer:** D

**Explanation:**

Option A is incorrect. While Kinesis Data Streams is the appropriate streaming solution for gathering the streaming player data and loading it onto your EMR cluster, then using Spark Streaming to transform the data into a format that is efficiently stored in ElastiCache Redis. There is no need for DynamoDB based on your data access and retrieval patterns for your application since your leaderboard application data is transient.

Option B is incorrect. Streaming your player data from Kinesis Data Firehose straight to S3 without any caching or transformation won’t give you your leaderboard functionality.

Option C is incorrect. While Kinesis Data Streams is the appropriate streaming solution for gathering the streaming player data and loading it onto your EMR cluster, then using Spark Streaming to transform the data into a format that is efficiently stored in ElastiCache. The Memcached version of ElastiCache does not allow you to easily implement the leaderboard functionality that ElastiCache Redis gives you. So this option is much less efficient.

Option D is correct. Kinesis Data Streams is the appropriate streaming solution for gathering the streaming player data and loading it onto your EMR cluster, then using Spark Streaming to transform the data into a format that is efficiently stored in ElastiCache Redis. You can use the Redis INCR and DECR functions to keep track of user points and the Redis Sorted Set data structure to maintain the leader list sorted by player. You can maintain your real-time ranked leader list by updating each user's score each time it changes.

Option E is incorrect. Based on your data access and retrieval patterns, there is no need for an S3 storage layer in this architecture.

**Reference:**

Please see the **Amazon ElastiCache for Redis overview page** (<https://aws.amazon.com/elasticache/redis/>), the **Amazon ElastiCache for Redis User Guide** (<https://docs.aws.amazon.com/AmazonElastiCache/latest/red-ug/redis-ug.pdf>), the **RedisLabs Leaderboards page** (<https://redislabs.com/redis-enterprise/use-cases/leaderboards/>), the AWS Database Blog page titled **Build a real-time gaming leaderboard with Amazon ElastiCache for Redis** (<https://aws.amazon.com/blogs/database/building-a-real-time-gaming-leaderboard-with-amazon-elasticache-for-redis/>), and the Amazon ElastiCache for Redis user guide titled **Common ElastiCache Use Cases and How ElastiCache Can Help** (<https://docs.aws.amazon.com/AmazonElastiCache/latest/red-ug/elasticache-use-cases.html>)

Question: 12

**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 for a car manufacturer who has implemented many sensors into their vehicles such as GPS, lane-assist, braking-assist, temperature/humidity, etc. These cars continuously transmit their structured and unstructured sensor data. You need to build a data collection system to capture their data for use in ad-hoc analytics applications to understand the performance of the cars, the locations traveled to and from, the effectiveness of the lane and brake assist features, etc. You also need to filter and transform the sensor data depending on rules based on parameters such as temperature readings. The sensor data needs to be stored indefinitely, however you only wish to pay for the analytics processing when you use it.

Which of the following architectures best suits your data lifecycle and usage patterns using the simplest, most efficient approach?

1. Sensor data -> Kinesis Data Streams -> IoT Core -> S3 -> Athena
2. Sensor data -> Kinesis Data Firehose -> IoT Core -> S3 -> Athena
3. Sensor data -> Kinesis Data Streams -> IoT Core -> Kinesis Data Firehose -> RedShift -> QuickSight
4. Sensor data -> IoT Core -> S3 -> Athena
5. Sensor data -> Kinesis Data Firehose -> S3 -> Athena

**Answer:** D

**Explanation:**

Option A is incorrect. While Kinesis Data Streams can be used to ingest IoT sensor data, it is an unnecessary component in your data collection architecture since IoT Core can do the sensor data ingestion task.

Option B is incorrect. While Kinesis Data Firehose can be used to ingest IoT sensor data, it is an unnecessary component in your data collection architecture since IoT Core can do the sensor data ingestion task.

Option C is incorrect. This data collection architecture has unnecessary components. While Kinesis Data Streams can be used to ingest IoT sensor data, it is an unnecessary component in your data collection architecture since IoT Core can do the sensor data ingestion task. RedShift is not the optimal data store for your IoT sensor data in this scenario. RedShift is better suited for storing structured data, but you have both structured and unstructured data.

Option D is correct. The simplest data collection architecture that meets your data lifecycle and usage patterns uses IoT Core to ingest the sensor data. Also, IoT Core is used to run a rules-based filtering and transformation set of functions. IoT Core then streams the sensor data to S3 where you house your data lake. You then use Athena to run your ad-hoc queries on your sensor data, taking advantage of Athena’s serverless query service so that you only pay for the service when you use it.

Option E is incorrect. This data collection architecture gives you a simple process flow to get your sensor data into your S3 data lake. However, it lacks the rules-based filtering and transformation set of functions. You would have to implement these functions in a Lambda function, which would make this data collection architecture less efficient than using the IoT Core service to address this requirement.

**Reference:**

Please see the **AWS IoT Core overview page** (<https://aws.amazon.com/iot-core/>), the AWS Big Data blog titled **Integrating IoT Events into Your Analytic Platform**

(<https://aws.amazon.com/blogs/big-data/integrating-iot-events-into-your-analytic-platform/>), the blog titled **Athena Vs Redshift: An Amazonian Battle Or Performance And Scale** (<https://blog.panoply.io/an-amazonian-battle-comparing-athena-and-redshift>), and the **Amazon Athena overview page** (<https://aws.amazon.com/athena/>)

Question: 13

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Analysis and Visualization

**Question text**:

You work for a public health governmental organization where you are responsible for building out a data warehouse to hold infectious disease information based on the data found at the World Health Organization’s Global Health Observatory data repository. You expect your initial data warehouse to hold less than TBs of data. However, you expect that the data stored in your warehouse will grow rapidly based on the state of world-wide infectious disease progression in the near future.

Your organization plans to use the data stored in your data warehouse to visualize disease progression across the various states in your country as infectious diseases progress through their lifecycle. These analyses will be used to make important decisions about citizen interaction and mobility.

Which of the following data warehouse configurations best suits your data analysis scenario using the simplest, most cost effective approach?

1. Redshift with RA3 nodes
2. Redshift with DC2 nodes
3. S3 with SSD volumes
4. S3 with HDD volumes
5. Redshift with DS2 nodes

**Answer:** A

**Explanation:**

Option A is correct. Redshift is the best choice for your data warehouse. Also, when configuring your Redshift warehouse, if you have less than 10 TBs of data DC2 nodes are the best price performer. However, if you expect your data to rapidly grow, as in this scenario, then RA3 nodes are the most cost effective choice.

Option B is incorrect. Redshift is the best choice for your data warehouse. Also, when configuring your Redshift warehouse, if you have less than 10 TBs of data DC2 nodes are the best price performer. However, if you expect your data to rapidly grow, as in this scenario, then RA3 nodes are the most cost effective choice.

Option C is incorrect. S3 is not a good choice for a data warehouse. Also, you do not choose the volume type when you create your S3 buckets.

Option D is incorrect. S3 is not a good choice for a data warehouse. Also, you do not choose the volume type when you create your S3 buckets.

Option E is incorrect. Redshift is the best choice for your data warehouse. Also, when configuring your Redshift warehouse, if you have less than 10 TBs of data DC2 nodes are the best price performer. However, if you expect your data to rapidly grow, as in this scenario, then RA3 nodes are the most cost effective choice. The DS2 node type is now classified as a legacy node choice by Amazon. Amazon no longer recommends that you build new Redshift data warehouses using the DS2 node type.

**Reference:**

Please see the Data Lakes and Analytics on AWS page titled **What is a Data Lake?** (<https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/>), the **Amazon Redshift Pricing page** (<https://aws.amazon.com/redshift/pricing/>) and the World Health Organization **Global Health Observatory data repository page** (<https://apps.who.int/gho/data/node.home>)

Question: 14

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Storage and Data Management

**Question text**:

You work for a large city police department as a data scientist. You have been given the task of tracking crime by city district for each criminal committing the given crime. You have created a DynamoDB table to track the crimes across your city’s districts. The table has this configuration: for each crime the table contains a CriminalId (the partition key), CityDistrict, and CrimeDate the crime was reported. Your police department wants to create a dashboard of the crimes reported by district and date.

What is the most cost effective way to retrieve the crime data from your DynamoDB table to build your crimes reported by district and date?

1. Create a local secondary index with CriminalId as the partition key and CrimeDate as the sort key
2. Create a global secondary index with CityDistrict as the partition key and CrimeDate as the sort key
3. Scan the table and use the ProjectionExpression parameter to return the crimes reported by district and date
4. Scan the secondary index and use the ProjectionExpression parameter to return the crimes reported by district and date

**Answer:** B

**Explanation:**

Option A is incorrect. Since you are looking to use the CityDistrict and CrimeDate to retrieve your dashboard data, the combination of CityDistrict and CrimeDate won’t always be unique. A global secondary index is the best choice for this use case since the combination of primary key attributes does not require unique values.

Option B is correct. Since you are looking to use the CityDistrict and CrimeDate to retrieve your dashboard data, the combination of CityDistrict and CrimeDate won’t always be unique. A global secondary index is the best choice for this use case since the combination of primary key attributes does not require unique values.

Option C is incorrect. Scanning the entire table and then using the ProjectionExpression parameter to filter the returned data will be a much more expensive operation than using a secondary index.

Option D is incorrect. Scanning a secondary index and then using the ProjectionExpression parameter to filter the returned data will be a much more expensive operation than just using a secondary index. Also, the scenario doesn’t state that you have created a secondary index, so how could you scan it if you haven’t yet created it?

**Reference:**

Please see the Amazon DynamoDB developer guide titled **Using Global Secondary Indexes in DynamoDB** (<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/GSI.html>), the Amazon DynamoDB developer guide titled **Working with Scans in DynamoDB** (<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Scan.html>), and the Amazon DynamoDB developer guide titled **Local Secondary Indexes** (<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/LSI.html>)

Question: 15

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Storage and Data Management

**Question text**:

You work for a large retail and wholesale business with a significant ecommerce web presence. Your company has just acquired a new ecommerce clothing line and needs to build a data warehouse for this new line of business. The acquired ecommerce business sells clothing to a niche market of men’s casual and business attire. You have chosen to use Amazon Redshift for your data warehouse. The data that you’ll initially load into the warehouse will be relatively small. However, you expect the warehouse data to grow as the niche customer base expands once the parent company makes a significant investment in advertising.

What is the most cost effective and best performing Redshift strategy that you should use when you create your initial tables in Redshift?

1. Use the KEY distribution strategy
2. Use the EVEN distribution strategy
3. Use the ALL distribution strategy
4. Use the AUTO distribution strategy

**Answer:** D

**Explanation:**

Option A is incorrect. With the KEY distribution strategy the Redshift leader node distributes the rows relative to the values in one column. This strategy is good for situations where you need to do joins across tables, but since your initial table sizes are small and will grow over time, there are better performing and more cost effective strategies you can use.

Option B is incorrect. With the EVEN distribution strategy, the Redshift leader node distributes the rows of your tables across the compute node slices using a round robin approach. This is not the best strategy if your tables need to participate in joins. This may be a good strategy for your tables once your tables increase in size as your new business grows, but since your initial table sizes are small, there are better performing and more cost effective strategies you can use.

Option C is incorrect. With the ALL distribution strategy, the Redshift leader node distributes the entire table to every compute node. Thus multiplying the storage required by the number of compute nodes you have configured in your Redshift cluster. This strategy is a good choice for tables that are not updated often and that are not updated with large change sets. This may be a good choice when you first create your tables, but since you expect rapid growth in your tables, this choice would not give you the optimum performance and cost over the life of your Redshift cluster.

Option D is correct. The AUTO distribution strategy Redshift assigns the best distribution strategy based on the table size. It then changes the distribution strategy as the changing table activity and size demands. So Redshift may initially assign an ALL distribution strategy to your table since it is small, then change the distribution strategy to EVEN as your table grows in size. When Redshift changes the distribution strategy the change happens very quickly (a few seconds) in the background.

**Reference:**

Please see the Amazon Redshift Database developer guide titled **Choosing a Distribution Style** (<https://docs.aws.amazon.com/redshift/latest/dg/t_Distributing_data.html>), 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 Management Overview** (<https://docs.aws.amazon.com/redshift/latest/mgmt/overview.html>)

Question: 16

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Storage and Data Management

**Question text**:

You are a data scientist working for a multinational conglomerate corporation that has many data stores for which you need to provide a common repository. All of your company’s systems need to use this common repository to store and retrieve metadata to work with the data stored in all of the data siolos throughout the organization. You also need to provide the ability to query and transform the data in the organization’s data silos. This common repository will be used for data analytics by your data scientist team to produce dashboards and KPIs for your management team.

You are using AWS Glue to build your common repository as depicted in this diagram:

[](https://www.draw.io/?page-id=64pGz7Uf3Qyy9XCqssY2&scale=auto#G159eW6hO6VsoGQwnyvseA_rsiYjI0UnKJ)

As you begin to create this common repository you notice that you aren’t getting the inferred schema for some of your data stores. You have run your crawler against your data stores using your custom classifiers. What might be the problem with your process?

1. The username you provided to your JDBC connection to your S3 buckets does not have SELECT permission to retrieve metadata from the S3 bucket data store
2. The username you provided to your JDBC connection to your Redshift clusters does not have SELECT permission to retrieve metadata from the Redshift data store
3. You did not use the Glue built-in classifiers in your crawler job
4. The username you provided to your JDBC connection to your DynamoDB tables does not have SELECT permission to retrieve metadata from the DynamoDB data store

**Answer:** B

**Explanation:**

Option A is incorrect. You do not need to use a JDBC connector to crawl S3 data stores. Your crawler can crawl S3 data stores through the native S3 interface.

Option B is correct. For data stores such as Redshift and RDS, you need to use a JDBC connector to crawl these types of data stores. If the username you provide to your JDBC connection does not have the appropriate permissions to access the data store, the connection will fail and Glue will not produce the inferred schema for that data store.

Option C is incorrect. Glue automatically runs its built-in classifiers if none of your custom classifiers return a certainty number equal to 1.

Option D is incorrect. You do not need to use a JDBC connector to crawl DynamoDB data stores. Your crawler can crawl DynamoDB data stores through the native DynamoDB interface.

**Reference:**

Please see the AWS Glue developer guide titled **Populating the AWS Glue Data Catalog** (<https://docs.aws.amazon.com/glue/latest/dg/populate-data-catalog.html>), the AWS Glue developer guide titled **Adding Classifiers to a Crawler** (<https://docs.aws.amazon.com/glue/latest/dg/add-classifier.html>)

Question: 17

**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 are a data scientist working for a retail chain that stores information about their supply chain partners (partner metadata) and their interaction with these partners (products produced, payments processed, competing partners, etc.). You are tasked with building a data store and associated data lifecycle management system for this partner data. The data will be used for analytics in managing these partners to maximize profitability for your supply chain.

You need to manage the data lifecycle according to the various access patterns defined for each type while maintaining storage cost efficiency. The partner metadata is less frequently accessed than the partner interaction data. You need to manage your storage costs so that high frequency accessed data (such as your partner interaction data) is available at very fast response times (sub-second), less frequently accessed data (such as your partner metadata) is available in minutes, and your rarely accessed data (such as historical data on former partners) is available within hours.

Which storage lifecycle best fits your usage patterns and business requirements?

1. Partner interaction data (sub-second response) stored in Redshift, partner metadata (minutes response) stored in S3 Standard, and former partner data (hours response) in S3 Intelligent-Tiering.
2. Use a Redshift cluster for all of your data. Create RA3 nodes in your cluster for your partner interaction data (sub-second response), create DC2 nodes for your partner metadata (minutes response), and DS2 nodes for your former partner data (hours response).
3. Partner interaction data (sub-second response) stored in Redshift, partner metadata (minutes response) stored in S3 Standard, and former partner data (hours response) in S3 Glacier.
4. Partner interaction data (sub-second response) stored in RDS Aurora, partner metadata (minutes response) stored in S3 Standard, and former partner data (hours response) in S3 Glacier.

**Answer:** C

**Explanation:**

Option A is incorrect. Redshift is a good choice for your partner interaction data because it requires sub-second response times. S3 Standard is a good choice for your partner metadata because it offers good response times (in minutes) at a much lower cost than Redshift. S3 Intelligent-Tiering is not the best choice for your former partner data because it is less cost optimized than the S3 Glacier tier for this type of infrequently accessed data. For example, when a data object is retrieved from the S3 Intelligent-Tier infrequently accessed tier, that object is moved to the frequently accessed tire. It then stays in the frequently accessed tier for 30 days.

Option B is incorrect. Using Redshift for all of your data storage and relying on cluster node types to optimize storage costs based on frequency is not a best practice use case for Redshift. This option will cost much more to maintain than the option with Redshift, S3 Standard, and S3 Glacier.

Option C is correct. Redshift is a good choice for your partner interaction data because it requires sub-second response times. S3 Standard is a good choice for your partner metadata because it offers good response times (minutes) at a much lower cost than Redshift. S3 Glacier is a good choice for your former partner data (hours) because the Glacier tier of S3 is the most inexpensive option for storing data like this that has very infrequent access and response times of an hour can be tolerated.

Option D is incorrect. Using RDS Aurora for your partner interaction data for this inherently data analytics warehouse type of use case is highly inefficient. Also, Redshift’s compressed, partitioned columnar storage format of your database tables optimizes your solution (and response times) for analytic query performance. This (analytics access) is listed as a requirement in the scenario.

**Reference:**

Please see the **Amazon Redshift features page** (<https://aws.amazon.com/redshift/features/>), the **Amazon Redshift FAQs page** (<https://aws.amazon.com/redshift/faqs/>), the Amazon Simple Storage Service developer guide titled **Amazon S3 Storage Classes** (<https://docs.aws.amazon.com/AmazonS3/latest/dev/storage-class-intro.html>), the **Amazon Redshift Pricing page** (<https://aws.amazon.com/redshift/pricing/>), the Amazon Redshift Cluster Management Guide titled **Amazon Redshift Clusters** (<https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html>), the **Amazon Aurora overview page** (<https://aws.amazon.com/rds/aurora/>), and the Amazon Redshift Database developer guide titled **Columnar Storage** (<https://docs.aws.amazon.com/redshift/latest/dg/c_columnar_storage_disk_mem_mgmnt.html>)

Question: 18

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Processing

**Question text**:

You are a data analyst working for a scientific research and data science company that is building a large scale data lake on EMR to house research data for ongoing research projects. Some of the projects have data processing requirements that need hot data set access, while others require less-hot data set access. For example, analysis for political polling related projects requires hot data set access due to the pressing nature of understanding political analytics and trends in real-time. Infrastructure and materials projects have less-hot data set access requirements since these projects have the option of producing their analysis on a daily basis versus a real-time basis.

Additionally, the real-time analytics projects require fast performance, their data is considered timely but temporary. However, the less-hot data projects don’t require real-time analytics, they require persistent data storage.

Which data processing solution best fits your usage patterns and business requirements?

1. S3 BFS for the hot data sets, S3 Glacier for the less-hot data sets
2. S3 EMRFS for the hot data sets, HDFS for the less-hot data sets
3. HDFS for the hot data sets, S3 EMRFS for the less-hot data sets
4. S3 BFS for the hot data sets, HDFS for the less-hot data sets

**Answer:** C

**Explanation:**

Option A is incorrect. S3 BFS (Block File System) is a legacy storage system and is no longer recommended by AWS. One reason: it can cause race conditions within your EMR cluster.

Option B is incorrect. S3 EMRFS is good for Hadoop file systems that need fast access for analytics, however the HDFS Hadoop file system is faster. Also, choosing HDFS for your data sets that require persistence is not a good option since HDFS is ephemeral, its storage is reclaimed when your EMR cluster is terminated.

Option C is correct. Use the HDFS Hadoop file system for your hot data sets that are temporary in nature, use the S3 EMRFS Hadoop file system for less-hot data sets that require persistence.

Option D is incorrect. S3 BFS (Block File System) is a legacy storage system and is no longer recommended by AWS. Choosing HDFS for your data sets that require persistence is not a good option since HDFS is ephemeral, its storage is reclaimed when your EMR cluster is terminated.

**Reference:**

Please see the **Amazon EMR FAQs page** (<https://aws.amazon.com/emr/faqs/>), the Amazon EMR Management guide titled **Working with Storage and File Systems** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-file-systems.html>), the **Amazon EMR Features page** (<https://aws.amazon.com/emr/features/>), and the Amazon EMR Management guide titled **Supported Applications and Features** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-ha-applications.html>)

Question: 19

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Processing

**Question text**:

You are a data scientist working for a large transportation company that manages its distribution data across all of its distribution lines: trucking, shipping, airfreight, etc. This data is stored in a data warehouse in Redshift. The company ingests all of the distribution data into an EMR cluster before loading the data into their data warehouse in Redshift. The data is loaded from EMR to Redshift on a schedule, once per day.

How might you lower the operational costs of running your EMR cluster? (CHOOSE 2)

1. EMR Transient Cluster
2. EMR Long-running Cluster
3. EMR Core Nodes as spot instances
4. EMR Task Nodes as spot instances
5. EMR cluster launched via AWS CLI using defaults

**Answers:** A, D

**Explanation:**

Option A is correct. EMR Transient Clusters automatically terminate after all steps are complete. This will lower your operational costs by not leaving the EMR nodes running when they are not in use.

Option B is incorrect. EMR Long-running clusters must be manually terminated when they are no longer needed, therefore this option will not give you the same cost effectiveness as a Transient Cluster.

Option C is incorrect. EMR Core Nodes run HDFS and therefore if a Code Node is terminated through the spot instance process, you will lose your data stored in HDFS.

Option D is correct. EMR Task Nodes do not store data in HDFS. If you lose your Task Node through the spot instance process you will not lose data stored on HDFS.

Option E is incorrect. When you launch an EMR cluster via the AWS CLI, the default is to have auto-terminate disabled. This will in effect create a long running cluster.

**Reference:**

Please see the Amazon Redshift Database developer guide titled **Loading Data from Amazon EMR** (<https://docs.aws.amazon.com/redshift/latest/dg/loading-data-from-emr.html>), the Amazon EMR Management Guide titled **Benefits of Using Amazon EMR** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-overview-benefits.html>), the Amazon EMR Management Guide titled **Configuring a Cluster to Auto-Terminate or Continue** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-longrunning-transient.html>), and 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>)

Question: 20

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Processing

**Question text**:

You are a data scientist working for an online retail company that wishes to catalog all of their products in a data lake. They also want to load their product data from their data lake into a data warehouse that they can use for business intelligence (BI) dashboards and analytics with QuickSight.

How would you automate and operationalize the data processing to get the company’s product data from their data lake to their data warehouse in the most efficient, cost effective manner?

1. Product data put to S3 data lake -triggers-> Lambda -runs-> Glue Crawler -> on completion CloudWatch event rule -triggers-> Lambda which runs Glue ETL job that transforms data to JSON -> S3 -triggers-> Lambda which runs COPY command to move data to Redshift
2. Product data put to S3 data lake -triggers-> Lambda -runs-> Glue Crawler -> on completion CloudWatch event rule -triggers-> Lambda which runs Glue ETL job that transforms data to Parquet -> S3 -triggers-> Lambda which runs COPY command to move data to Redshift
3. Product data put to S3 data lake -triggers-> Lambda -runs-> Glue Crawler -> on completion CloudWatch event rule -triggers-> Lambda which runs Glue ETL job that transforms data to JSON -> S3 -triggers-> Lambda which runs COPY command to move data to RDS Aurora
4. Product data put to S3 data lake -triggers-> Lambda -runs-> Glue Crawler -> on completion CloudWatch event rule -triggers-> Lambda which runs Glue ETL job that transforms data to CSV -> S3 -triggers-> Lambda which runs COPY command to move data to Redshift

**Answer:** B

**Explanation:**

Option A is incorrect. JSON is not the most efficient format to use when using the COPY command to load data files into Redshift. Apache Parquet and ORC are better choices for loading data files into Redshift. Parquet and ORC are columnar data formats that allow you to copy your data more efficiently and cost-effectively into Redshift.

Option B is correct. Apache Parquet and ORC are better choices for loading data files into Redshift. Parquet and ORC are columnar data formats that allow you to copy your data more efficiently and cost-effectively into Redshift.

Option C is incorrect. RDS Aurora is not a good choice for housing your data warehouse. Redshift is better suited for data warehouse analytic applications.

Option D is incorrect. CSV is not the most efficient format to use when using the COPY command to load data files into Redshift. Apache Parquet and ORC are better choices for loading data files into Redshift. Parquet and ORC are columnar data formats that allow you to copy your data more efficiently and cost-effectively into Redshift.

**Reference:**

Please see the AWS What’s New article titled **Amazon Redshift Can Now COPY from Parquet and ORC File Formats** (<https://aws.amazon.com/about-aws/whats-new/2018/06/amazon-redshift-can-now-copy-from-parquet-and-orc-file-formats/>), the Amazon QuickSight user guide titled **Creating a Dataset from a Database** (<https://docs.aws.amazon.com/quicksight/latest/user/create-a-database-data-set.html>), and the Amazon Redshift Database developer guide titled **COPY from Columnar Data Formats** (<https://docs.aws.amazon.com/redshift/latest/dg/copy-usage_notes-copy-from-columnar.html>)

Question: 21

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Processing

**Question text**:

You work as a data scientist at a large hedge fund. Your firm produces analytics dashboard data for all of its traders. The data that you use is extracted from several trading systems, then transformed by removing canceled trades and classifying trades that remain open as pending. Quite often there are exotic trade types that your analytics application has not processed in past runs. When this happens your data processing solution needs to handle these new types of trades without having to modify the transformation code or the downstream data store.

This process is run at the end of each trading day for each trader in the firm. How would you automate and operationalize this data processing flow in the most efficient, cost effective manner?

1. A Glue schedule trigger runs at the end of the day which starts two Glue transformation jobs: remove\_canceled\_trades and classify\_open\_trades\_as\_pending. When both of these jobs have completed an event trigger starts a Glue crawler that crawls the transformed data and updates the schema. Upon completion of the crawler schema update, a Glue ETL job runs and uses the COPY command to move the data to Redshift. Analytics dashboards are built using Redshift data.
2. A Glue schedule trigger runs at the end of the day which starts two Glue transformation jobs: remove\_canceled\_trades and classify\_open\_trades\_as\_pending. When both of these jobs have completed an event trigger starts a Glue crawler that crawls the transformed data and updates the schema. Upon completion of the crawler schema update, a Glue ETL job runs and uses the UNLOAD command to move the data to Redshift. Analytics dashboards are built using Redshift data.
3. A cron job schedule trigger runs at the end of the day which starts two Glue transformation jobs: remove\_canceled\_trades and classify\_open\_trades\_as\_pending. When both of these jobs have completed a cron job schedule trigger starts a Glue crawler that crawls the transformed data and updates the schema. Upon completion of the crawler schema update, a Glue ETL job runs and uses the COPY command to move the data to Redshift. Analytics dashboards are built using Redshift data.
4. A Glue schedule trigger runs at the end of the day which starts two Glue transformation jobs: remove\_canceled\_trades and classify\_open\_trades\_as\_pending. When both of these jobs have completed an event trigger starts a Glue crawler that crawls the transformed data and updates the schema. Upon completion of the crawler schema update, a Glue ETL job runs and uses the PUT command to move the data to Redshift. Analytics dashboards are built using Redshift data.

**Answer:** A

**Explanation:**

Option A is correct. AWS Glue allows you to create workflows using extract, transform, and load (ETL) activities using as many crawlers, jobs, and triggers nas you need. The Glue job that runs at the completion of the schema update uses the Redshift COPY command to load the trade data into Redshift.

Option B is incorrect. AWS Glue allows you to create workflows using extract, transform, and load (ETL) activities using as many crawlers, jobs, and triggers nas you need. The Glue job that runs at the completion of the schema update should use the Redshift COPY command to load the trade data into Redshift. The UNLOAD command is used to retrieve data from Redshift, not to move data into Redshift.

Option C is incorrect. Adding cron jobs to the workflow over complicates the data processing solution. The use of cron jobs is unnecessary since Glue workflows can orchestrate your entire workflow.

Option D is incorrect. AWS Glue allows you to create workflows using extract, transform, and load (ETL) activities using as many crawlers, jobs, and triggers nas you need. The Glue job that runs at the completion of the schema update should use the Redshift COPY command to load the trade data into Redshift. There is no PUT command to move data to or from Redshift. The commands used to move data to and from Redshift are COPY and UNLOAD.

**Reference:**

Please see the AWS Glue developer guide titled **Overview of Workflows in AWS Glue** (<https://docs.aws.amazon.com/glue/latest/dg/workflows_overview.html>), the AWS Glue developer guide titled **Performing Complex ETL Activities Using Workflows in AWS Glue** (<https://docs.aws.amazon.com/glue/latest/dg/orchestrate-using-workflows.html>), and the AWS Glue developer guide titled **Moving Data to and from Amazon Redshift** (<https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming-etl-redshift.html>)

Question: 22

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Processing

**Question text**:

You work as a data scientist at a large global bank. Your bank receives loan information in the form of weekly files from several different loan processing and credit verification agencies. You need to automate and operationalize a data processing solution to take these weekly files, transform them and then finish up by combining them into one file to be ingested into your Redshift data warehouse. The files arrive at different times every week, but the delivering agencies attempt to meet their service level agreement (SLA) of 1:00 AM to 4:00 AM. Unfortunately, the agencies frequently miss their SLAs. You have a tight batch time frame into which you have to squeeze all of this processing.

How would you build a data processing system that allows you to gather the agency files and process them for your data warehouse in the most efficient manner and in the shortest time frame?

1. Agency files arrive on an S3 bucket. An ETL Lambda function is triggered as each file arrives. The ETL Lambda function transforms the data and writes the transformed file to another S3 bucket. After all of the agency files have been processed by the ETL Lambda function, another Lambda function is triggered to combine the agency file data into one parquet file and write it to another S3 bucket. Then a last Lambda function is triggered to run the COPY command to load the parquet file data into Redshift.
2. Agency files arrive on an S3 bucket. Use CloudWatch events to schedule a weekly Step Functions state machine. The Step Functions state machine calls a Lambda function to verify that the agency files have arrived. The state machine then starts several Glue ETL jobs in parallel to transform the agency data. Once the agency file transformation jobs have completed the state machine starts another Glue ETL job to combine the transformed agency files and convert the data to a parquet file. The parquet file is written to an S3 bucket. Then the state machine finally runs a last Glue ETL job to run the COPY command to load the parquet file data into Redshift.
3. Agency files arrive on an S3 bucket. An ETL Lambda function is triggered as each file arrives. The ETL Lambda function transforms the data and writes the transformed file to another S3 bucket. After all of the agency files have been processed by the ETL Lambda function, another Lambda function is triggered to combine the agency file data into one CSV file and write it to another S3 bucket. Then a last Lambda function is triggered to run the UNLOAD command to load the CSV file data into Redshift.
4. Agency files arrive on an S3 bucket. Use CloudWatch events to schedule a weekly Step Functions state machine. The Step Functions state machine calls a Lambda function to verify that the agency files have arrived. The state machine then starts several Glue ETL jobs in parallel to transform the agency data. Once the agency file transformation jobs have completed the state machine starts another Glue ETL job to combine the transformed agency files and convert the data to an ORC file. The ORC file is written to an S3 bucket. Then the state machine finally runs a last Glue ETL job to run the UNLOAD command to load the ORC file data into Redshift.

**Answer:** B

**Explanation:**

Option A is incorrect. This Lambda based data processing solution would work but it is less efficient and will take longer to run than using Step Functions state machines to run the several ETL transformation jobs in parallel.

Option B is correct. Using Step Functions state machines to orchestrate this data processing workflow allows you to take advantage of processing all of your transformation ETL jobs in parallel. This makes your data processing workflow efficient and allows it to fit within your tight batch window.

Option C is incorrect. It is less efficient and will take longer to run than using Step Functions state machines to run the several ETL transformation jobs in parallel. Also, using a CSV file to load data into your Redshift cluster is slower and less efficient than using either the ORC or parquet formats. Finally, you use the COPY command to load data into your Redshift cluster, not the UNLOAD command.

Option D is incorrect. You use the COPY command to load data into your Redshift cluster, not the UNLOAD command.

**Reference:**

Please see the AWS Glue developer guide titled **Performing Complex ETL Activities Using Workflows in AWS Glue** (<https://docs.aws.amazon.com/glue/latest/dg/orchestrate-using-workflows.html>), the AWS Big Data blog titled **Orchestrate multiple ETL jobs using AWS Step Functions and AWS Lambda** (<https://aws.amazon.com/blogs/big-data/orchestrate-multiple-etl-jobs-using-aws-step-functions-and-aws-lambda/>), the AWS Glue developer guide titled **Moving Data to and from Amazon Redshift** (<https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming-etl-redshift.html>), the AWS announcement titled **Amazon Redshift Can Now COPY from Parquet and ORC File Formats** (<https://aws.amazon.com/about-aws/whats-new/2018/06/amazon-redshift-can-now-copy-from-parquet-and-orc-file-formats/>), and the AWS Big Data blog titled **Orchestrate Amazon Redshift-Based ETL workflows with AWS Step Functions and AWS Glue** (<https://aws.amazon.com/blogs/big-data/orchestrate-amazon-redshift-based-etl-workflows-with-aws-step-functions-and-aws-glue/>)

Question: 23

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Processing

**Question text**:

You work as a cloud architect for a cloud consultancy practice at a major IT consulting firm. Your latest client has a series of data processing Apache Spark ELT jobs that they want to run in a pipeline on EMR. Thay have asked you which set of data processing tools and techniques will best suit their pipeline needs. The jobs have a specified sequence. Your client wants to manage their costs. Therefore, they want to keep the solution simple, they don’t want to build an application to run these jobs, and they don’t want to incur any additional costs on virtual servers to run their pipeline. Also, they plan on integrating their Apache Spark pipeline with other AWS services in the future.

Which orchestration tool set best suits your client’s pipeline requirements?

1. Apache Oozie to schedule and run the Spark jobs
2. Apache Airflow to schedule and run the Spark jobs
3. AWS Step Functions to schedule and run the Spark jobs
4. AWS Lambda to schedule and run the Spark jobs
5. AWS DMS to schedule and run the Spark jobs

**Answer:** C

**Explanation:**

Option A is incorrect. Apache Oozie is a popular workflow scheduler for Hadoop jobs, but it has limited integration with AWS services and requires XML configuration which makes using it more complex than using Step Functions, thereby increasing the cost of the solution.

Option B is incorrect. Apache Airflow integrates with several AWS services, but it requires your client to run it on a server that they’ll also have to maintain. This will increase the cost compared to using Step Functions.

Option C is correct. Using Step Functions will allow your client to run their workflow as a serverless pipeline that runs their Spark ETL jobs using the Apache Livy REST service. This will allow for very quick development time and pay-as-you-use costs, which will be far less expensive than the other options.

Option D is incorrect. You could use Lambda to string together a pipeline. While this approach gives you a serverless pipeline, it lacks the job flow coordination features that Step Functions has. Your client would have to write these capabilities themselves, increasing the cost of their solution.

Option E is incorrect. AWS Database Migration Service (DMS) is primarily used to migrate databases to AWS. Your client could use DMS to load data from existing databases into S3 and then use Glue to run Spark ETL jobs, but this is not what the scenario describes.

**Reference:**

Please see the AWS Big Data blog titled **Orchestrate Apache Spark applications using AWS Step Functions and Apache Livy** (<https://aws.amazon.com/blogs/big-data/orchestrate-apache-spark-applications-using-aws-step-functions-and-apache-livy/>), the AWS News blog titled **New – Using Step Functions to Orchestrate Amazon EMR Workloads** (<https://aws.amazon.com/blogs/aws/new-using-step-functions-to-orchestrate-amazon-emr-workloads/>), the **Apache Livy overview page** (<https://livy.apache.org/>), and the AWS Big Data blog titled **Load ongoing data lake changes with AWS DMS and AWS Glue** (<https://aws.amazon.com/blogs/big-data/loading-ongoing-data-lake-changes-with-aws-dms-and-aws-glue/>)

Question: 24

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Processing

**Question text**:

You work as a cloud architect for a gaming company that is building an analytics platform for their gaming data. This analytics platform will ingest game data from current games being played by users of their mobile game platform. The game data needs to be loaded into a data lake where business intelligence (BI) tools will be used to build analytics views of key performance indicators (KPIs). You load your data lake from an EMR cluster where you run Glue ETL jobs to perform the transformation of the incoming game data to the parquet file format. Once transformed, the parquet files are stored in your S3 data lake. From there you can run BI tools, such as Athena, to build your KPIs.

You want to handle EMR step through recovery logic. What is the simplest way to build retry logic into your data processing solution?

1. CloudTrail event rule sends a text message via a Simple Notification Service (SNS) topic, a support engineer reruns the failed EMR step.
2. CloudWatch event rule sends a text message via a Simple Notification Service (SNS) topic, a support engineer reruns the failed EMR step.
3. CloudTrail event rule triggers a Lambda function via a Simple Notification Service (SNS) topic which retries the EMR step.
4. CloudWatch event rule triggers a Lambda function via a Simple Notification Service (SNS) topic which retries the EMR step.
5. CloudWatch event rule triggers a retry of the Spark step via a Simple Notification Service (SNS) topic.

**Answer:** D

**Explanation:**

Option A is incorrect. CloudTrail does not have event rules.

Option B is incorrect. While this would work, it is not as efficient as having automated retry logic via a Lambda function.

Option C is incorrect. CloudTrail does not have event rules.

Option D is correct. Using SNS to trigger a Lambda function on failure allows you to use automated retry logic in your data processing solution.

Option E is incorrect. This option would require you to build some mechanism to allow Spark jobs to be initiated via an SNS topic. This would not be as simple as writing a Lambda function and having it triggered by the SNS topic.

**Reference:**

Please see the Amazon Simple Notification Service developer guide titled **Using Amazon SNS for system-to-system messaging with an AWS Lambda function as a subscriber** (<https://docs.aws.amazon.com/sns/latest/dg/sns-lambda-as-subscriber.html>), 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/>), and the AWS Lambda developer guide titled **Using AWS Lambda with Amazon SNS** (<https://docs.aws.amazon.com/lambda/latest/dg/with-sns.html>)

Question: 25

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Security

**Question text**:

You work as a cloud security architect for a financial services company. Your company has an EMR cluster that is integrated with their AWS Lake Formation managed data lake. You use the Lake Formation service to enforce column-level access control driven by policies you have defined. You need to implement a real-time alert and notification system if authenticated users run the TerminateJobFlows, DeleteSecurityConfiguration, or CancelSteps actions within EMR.

How would you implement this real-time alert mechanism in the simplest way possible?

1. Create a CloudTrail trail and enable continuous delivery of events to an S3 bucket. Use the **aws cloudtrail create-trail** CLI command to create an SNS topic. When an event occurs a Simple Queue Service (SQS) queue that subscribes to the SNS topic will receive the message. Use a Lambda function triggered by SQS to filter the messages for the TerminateJobFlows, DeleteSecurityConfiguration, or CancelSteps actions. The Lambda function will notify security alert subscribers via another SNS topic.
2. Create a CloudWatch event and enable continuous delivery of events to an S3 bucket. Use the **aws cloudwatch create-event** CLI command to create an SNS topic. When an event occurs for the TerminateJobFlows, DeleteSecurityConfiguration, or CancelSteps actions subscribers to the SNS topic will be notified.
3. Create a Lambda function that subscribes to an SNS topic that you define. The Lambda function will be triggered every time a TerminateJobFlows, DeleteSecurityConfiguration, or CancelSteps action is written to the EMR logs.
4. Create a CloudTrail trail and enable continuous delivery of events to an S3 bucket. Use the **aws cloudtrail create-trail** CLI command to create an SNS topic. When an event occurs for the TerminateJobFlows, DeleteSecurityConfiguration, or CancelSteps actions SNS will notify security alert subscribers.

**Answer:** A

**Explanation:**

Option A is correct. With CloudTrail you can configure your trail to use SNS topics. You use the **aws cloudtrail create-trail** CLI command to create the SNS topic. When events occur you use a Lambda function triggered by an SQS queue which receives the alert. The Lambda function filters for the events for which you are concerned. If you don’t filter the events you’ll receive alerts for every event generated by CloudTrail.

Option B is incorrect. CloudWatch is not the service to use when you are monitoring API calls. CloudTrail is the service to use for this purpose.

Option C is incorrect. This answer lacks the linkage of the SNS topic with the logging of events in the EMR logs.

Option D is incorrect. This answer lacks the filtering of CloudTrail messages. If you don’t filter the events you’ll receive alerts for every event generated by CloudTrail. This would make it hard for you to act on the events for which you are concerned.

**Reference:**

Please see the Amazon EMR Management guide titled **Logging Amazon EMR API Calls in AWS CloudTrail** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/logging_emr_api_calls.html>), the AWS CloudTrail user guide titled **Configuring Amazon SNS Notifications for CloudTrail** (<https://docs.aws.amazon.com/awscloudtrail/latest/userguide/configure-sns-notifications-for-cloudtrail.html>), the AWS CLI Command Reference titled **create-trail** (<https://docs.aws.amazon.com/cli/latest/reference/cloudtrail/create-trail.html>), the AWS CloudTrail user guide titled **Configuring CloudTrail to Send Notifications** (<https://docs.aws.amazon.com/awscloudtrail/latest/userguide/configure-sns-notifications-for-cloudtrail.html#configure-cloudtrail-to-send-notifications>), and the Amazon EMR Management guide titled **Conceptual Overview of Amazon EMR Integration with Lake Formation** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-lf-conceptual.html>)

Question: 26

**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 scientist for a medical data processing company. Your company receives patient data via file feeds into one of your S3 buckets. The data is formatted as a nested JSON document similar to this:

[

{

**"id"**: **"796"**,

**"category"**: **"Epidemiology"**,

**"info"**: {

**"subcategory"**: **"Neuroepidemiology"**,

**"questionType"**: **"multiple choice 1"**,

**"question"**: **"What is the reference to pi?"**,

**"answers"**: [

**"First three digits"**,

**"Infinite number of digits"**,

**"Digits after the decimal point"**,

**"Digits before the decimal point"**

],

**"correctAnswer"**: [**"Digits after the decimal point"**]

}

}

]

After performing data engineering on some sample files you have noticed occasional inconsistencies in the data types in the JSON.

What is the most performant and cost effective way to clean your semi-structured JSON data?

1. Run an AWS Batch job that uses the dirtyjson library
2. Use an EMR job that uses the Spark native DataFrame API
3. Trigger a Lambda function that uses the json.load and json.loads libraries
4. Run a Glue job that uses the DynamicFrame extension

**Answer:** D

**Explanation:**

Option A is incorrect. You could write an AWS Batch job that uses the dirtyjson library to clean your JSON, but you would have to spend development time building the code that leverages the dirtyjson library, costing you development time.

Option B is incorrect. The Spark DataFrame API requires a schema to know before you load your data. It also doesn’t handle cleaning up data as well as the Glue DynamicFrame class. The Spark DataaFrame makes two passes over the JSON dataset, costing you operational time and performance costs. Also, setting up an EMR cluster to run your job will cost you development time and you’ll have to pay for EC2 instances to run your EMR cluster, costing you infrastructure expenses.

Option C is incorrect. The json.load and json.loads libraries would not give you the capability to clean your semi-structured data. These libraries give you the capability to convert your JSON data into python objects. You would then have to write custom code to actually clean up the inconsistencies.

Option D is correct. The Glue DynamicFrame extension requires no schema; Glue determines the schema in real-time while handling schema inconsistencies using the resolveChoice, unnest, split\_rows, relationalize, and other transforms.

**Reference:**

Please see the RealPython article titled **Working with JSON Data in Python** (<https://realpython.com/python-json/>), the **pip project description of dirtyjson** (<https://pypi.org/project/dirtyjson/>), the **AWS Batch API Reference** (<https://docs.aws.amazon.com/batch/latest/APIReference/batch-api.pdf>), the AWS Glue developer guide titled **DynamicFrame Class** (<https://docs.aws.amazon.com/glue/latest/dg/aws-glue-api-crawler-pyspark-extensions-dynamic-frame.html>), and the Spark SQL guide titled **Spark SQL, DataFrames and Datasets Guide** (<https://spark.apache.org/docs/latest/sql-programming-guide.html>)

Question: 27

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Collection

**Question text**:

You work as a data scientist for a rideshare company. Rideshare request data is collected in one of the company’s S3 buckets (inbound bucket). This data needs to be processed (transformed) very quickly, within seconds of being put onto the S3 bucket. Once transformed, the rideshare request data must be put into another S3 bucket (transformed bucket) where it will be processed to link rideshare drivers with rideshare requesters.

You have already written Spark jobs to do the transformation. You need to control costs and minimize data latency for the rideshare request transformation operationalization of your data collection system. Which option best meets your requirements?

1. Lambda function triggered when the rideshare data request is put onto the inbound S3 bucket. Lambda sends an SNS topic to an SQS queue. Another Lambda function polls the queue every minute and when it finds a message it launches an EMR cluster and submits a Spark job to process the request.
2. Lambda function triggered when the rideshare data request is put onto the inbound S3 bucket. The Lambda function passes the request data to a Spark job in Glue.
3. Lambda function triggered when the rideshare data request is put onto the inbound S3 bucket. The Lambda function launches an EMR cluster and submits the job using the EMR Steps API to process the request.
4. Build an EMR cluster that runs Apache Livy. Lambda function triggered when the rideshare data request is put onto the inbound S3 bucket. The Lambda function passes the request data to a Spark job on the EMR cluster.

**Answer:** D

**Explanation:**

Option A is incorrect. This approach will be too slow in transforming and then moving the request data to the transformed bucket. Starting up an EMR cluster and then submitting the Spark job will take far longer than using a long running EMR cluster.

Option B is incorrect. A Spark job running in Glue is batch oriented. You can only schedule ETL jobs at 5 minute intervals or greater. This option will be far slower than using a long running EMR cluster.

Option C is incorrect. This approach will be too slow in transforming and then moving the request data to the transformed bucket. Starting up an EMR cluster and then submitting the EMR Steps API job will take far longer than using a long running EMR cluster.

Option D is correct. A Livy server on a long running EMR cluster will handle requests much faster than starting an EMR cluster with each request or using an SQS polling structure.

**Reference:**

Please see the **AWS Glue FAQs** (<https://aws.amazon.com/glue/faqs/>), the Amazon EMR Release Guide titled **Apache Livy** (<https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-livy.html>), the AWS Big Data blog titled **Build a Concurrent Data Orchestration Pipeline Using Amazon EMR and Apache Livy** (<https://aws.amazon.com/blogs/big-data/build-a-concurrent-data-orchestration-pipeline-using-amazon-emr-and-apache-livy/>), and the Apache Livy Getting Started guide (<https://livy.incubator.apache.org/get-started/>)

Question: 28

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Analysis and Visualization

**Question text**:

You are a data scientist working on a preventative health research project using the Global Health Observatory data repository. This repository contains the Body Mass Index (BMI) dataset which is based on several thousand observations from around the globe from 1975 to 2016. You need to analyze this dataset using QuickSite. One of the visuals you’ve been asked to create is to show the prevalence of thinness by country across the globe from 1975 to 2016 at 5-year increments. What is the best visual type to use to display this data?

1. Geospatial chart
2. Bubble chart
3. Heat map
4. Tree map

**Answer:** A

**Explanation:**

Option A is correct. You are looking for the historical prevalence of thinness data for each country around the world. This is a perfect use of geospatial charts, where you want to show differences in data values across a geographical map.

Option B is incorrect. Use bubble charts to compare values for items in a dimension. A bubble is displayed on the chart at the point where the measures for an item intersect within a dimension.

Option C is incorrect. Heat maps are used to show how two dimensions intersect. You use colors to show the range of distribution. You are looking to show the distribution of data values across a geographic map.

Option D is incorrect. Use tree maps to show how one or two data points in a dimension using rectangles. Every rectangle displayed on the tree shows one data point in the dimension. You are looking to show the distribution of data values across a geographic map.

**Reference:**

Please see the Amazon QuickSight user guide titled **Using Tree Maps** (<https://docs.aws.amazon.com/quicksight/latest/user/tree-map.html>), the Amazon QuickSight user guide titled **Using Heat Maps** (<https://docs.aws.amazon.com/quicksight/latest/user/heat-map.html>), the Amazon QuickSight user guide titled **Using Geospatial Charts (Maps)** (<https://docs.aws.amazon.com/quicksight/latest/user/geospatial-charts.html>), the Amazon QuickSight user guide titled **Using Scatter Plots** (<https://docs.aws.amazon.com/quicksight/latest/user/scatter-plot.html>), the **Amazon QuickSight overview page** (<https://aws.amazon.com/quicksight/?c=a&sec=srv>), and the World Health Organization’s **Global Health Observatory data repository** (<https://apps.who.int/gho/data/view.main.NCDBMIMINUS210-19Cv?lang=en>)

Question: 29

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Analysis and Visualization

**Question text**:

You are a data scientist working for the Fédération Internationale de Football Association (FIFA). Your management team has asked you to select the appropriate data analysis solution to analyze streaming football data in near real-time. You need to use this data to build interactive results through graphics and interactive charts for the FIFA management team. The football streaming events are based on time series that are unordered and may frequently be duplicated. You also need to transform the football data before you store it. You’ve been instructed to focus on providing high quality functionality based on fast data access.

Which solution best fits your needs?

1. Kinesis data firehose -> ORC files -> S3 -> Athena
2. Kinesis data firehose -> Lambda -> Elasticsearch Cluster -> Kibana
3. Kinesis data firehose -> RDS -> QuickSight
4. Kinesis Data Streams -> parquet files -> S3 -> Redshift Spectrum

**Answer:** B

**Explanation:**

Option A is incorrect. This option does not give you the ability to analyze your football data in near real-time because you are transforming the data into ORC format, storing it on S3, and then attempting to query it across multiple ORC files.

Option B is correct. You can leverage a Lambda function together with Kinesis Data Firehose to transform your streaming football data prior to storage on the Elasticsearch cluster storage volumes. You can then use Elasticsearch together with Kibana to perform near real-time analytics on your streaming football data.

Option C is incorrect. Kinesis Data Firehose doesn’t have the capability to write its streaming data to RDS. It can write streaming data to S3, Redshift, Elasticsearch, and Splunk

Option D is incorrect. This option does not give you the ability to analyze your football data in near real-time because you are transforming the data into parquet format, storing it on S3, and then attempting to query it across multiple parquet files.

**Reference:**

Please see the **Amazon Elasticsearch Service FAQs** (<https://aws.amazon.com/elasticsearch-service/faqs/>), the **Amazon Athena FAQs** (<https://aws.amazon.com/athena/faqs/>), the **Kinesis Data Firehose overview page** (<https://aws.amazon.com/kinesis/data-firehose/>), the **Kinesis Data Streams overview page** (<https://aws.amazon.com/kinesis/data-streams/>), the AWS Big Data blog titled **Perform Near Real-time Analytics on Streaming Data with Amazon Kinesis and Amazon Elasticsearch Service** (<https://aws.amazon.com/blogs/big-data/perform-near-real-time-analytics-on-streaming-data-with-amazon-kinesis-and-amazon-elasticsearch-service/>**)**, the **Kibana overview page** (<https://www.elastic.co/kibana>)**,** and the **Wikipedia page on FIFA** (<https://en.wikipedia.org/wiki/FIFA>)

Question: 30

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Security

**Question text**:

You are a data scientist working for a large bank where you are building out an EMR cluster for their customer information data lake. Due to the Personally Identifiable Information (PII) stored in the data lake, you need to lock down all environments (dev, engineering, test, perf, prod) to make sure only the appropriate users and user groups have access to the data lake.

To accomplish this goal you have created this IAM policy and attached it to your users and user groups who will be working with your EMR cluster:

{

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

"Statement": [

{

"Sid": "Stmt7645587658758",

"Effect": "Allow",

"Action": [

"elasticmapreduce:DescribeCluster",

"elasticmapreduce:ListSecurityConfigurations",

"elasticmapreduce:ListSteps",

"elasticmapreduce:TerminateJobFlows",

"elasticmapreduce:ModifyCluster",

"elasticmapreduce:PutAutoScalingPolicy",

"elasticmapreduce:ListInstances",

"elasticmapreduce:SetTerminationProtection",

"elasticmapreduce:DescribeStep"

],

"Resource": [

"\*"

],

"Condition": {

"StringEquals": {

"elasticmapreduce:ResourceTag/department": ["dev", “eng”]

}

}

}

]

}

How does this policy protect your EMR cluster that contains the company’s customer PII data?

1. It prevents users from performing the actions listed in the Action part of the Statement
2. It allows users to perform only the actions listed in the Action part of the Statement
3. It only allows users to perform the actions listed in the Action part of the Statement if the EMR cluster they are attempting to access is tagged as department of dev or eng
4. It only allows users to perform the actions listed in the Action part of the Statement if the EMR cluster they are attempting to access is NOT tagged as department of dev or eng

**Answer:** C

**Explanation:**

Option A is incorrect. The actions listed in the Action part of the Statement are associated with the Effect of Allow, so these actions aren’t prevented unilaterally by the policy. The StringEquals condition controls the access.

Option B is incorrect. The actions listed in the Action part of the Statement are associated with the Effect of Allow, but those actions aren’t allowed unilaterally by the policy. The StringEquals condition controls the access.

Option C is correct. The StringEquals condition attempts to match dev or eng to the value of the department tag. If the department tag was not added to the EMR cluster or the department tag does not have the value of either dev or eng, then the policy does not apply and the user/group can’t perform the actions on the EMR cluster.

Option D is incorrect. The StringEquals condition attempts to match dev or eng to the value of the department tag. If the department tag has the value of either dev or eng, then the policy applies and the user/group can perform the actions on the EMR cluster.

**Reference:**

Please see the Amazon EMR management guide titled **IAM Policies for Tag-Based Access to Clusters and EMR Notebooks** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-fine-grained-cluster-access.html>), and the Amazon Elastic Map Reduce API reference titled **Actions** (<https://docs.aws.amazon.com/emr/latest/APIReference/API_Operations.html>)

Question: 31

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Security

**Question text**:

You are a data scientist working for a financial services firm where you are building out an EMR cluster used to house the data lake used for your company’s proprietary machine learning models that predict market movement in global markets. The data in this data lake is considered to be a fundamental part of the company’s knowledge capital so it can only be accessed by users in the Quantitative Equity Group defined within IAM. You need to lock down all environments (dev, engineering, test, perf, prod) to make sure only the users in the Quantitative Equity Group have access to the data lake.

To accomplish this goal you have created this IAM policy and attached it to your users in the Quantitative Equity Group IAM group who will be working with your confidential EMR cluster:

{

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

"Statement": [

{

"Sid": "Stmt7645587658758",

"Effect": "Allow",

"Action": [

"elasticmapreduce:DescribeCluster",

"elasticmapreduce:ListSecurityConfigurations",

"elasticmapreduce:ListSteps",

"elasticmapreduce:TerminateJobFlows",

"elasticmapreduce:ModifyCluster",

"elasticmapreduce:PutAutoScalingPolicy",

"elasticmapreduce:ListInstances",

"elasticmapreduce:SetTerminationProtection",

"elasticmapreduce:DescribeStep"

],

"Resource": [

"\*"

],

"Condition": {

"StringEquals": {

"elasticmapreduce:ResourceTag/department":

["dev", “eng”, “test”, “perf”, “prod”]

}

}

}

]

}

You then created this policy and attached it to all users to further lockdown the EMR cluster environments:

{

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

"Statement": [

{

"Effect": "Deny",

"Action": [

"elasticmapreduce:AddTags",

"elasticmapreduce:RemoveTags"

],

"Condition": {

"StringNotEquals": {

"elasticmapreduce:ResourceTag/department":

["dev", “eng”, “test”, “perf”, “prod”]

}

},

"Resource": [

"\*"

]

}

]

}

What further protection does this policy give you (SELECT TWO)?

1. It prevents users from adding or removing tags on an EMR cluster that is tagged with the department of dev, eng, test, perf, or prod. This prevents users from giving themselves access to the EMR cluster by adding a tag to the cluster that their user profile allows
2. It prevents users from adding or removing tags on an EMR cluster that is tagged with the department of dev, eng, test, perf, or prod. This prevents users from giving anyone access to the EMR cluster by adding tags to the cluster, opening up access to any user
3. It prevents users from adding or removing tags on an EMR cluster that is tagged with the department of dev, eng, test, perf, or prod. This prevents users from giving themselves access to the EMR cluster by removing the tags on the cluster, opening up access to any user even if they aren’t in the Quantitative Equity Group IAM group
4. It prevents users from adding or removing tags on an EMR cluster that is tagged with the department of dev, eng, test, perf, or prod. This prevents users from giving themselves access to the EMR cluster by removing a tag on the cluster, opening up access to any user who is allowed to act on resources with that tag
5. It prevents users from adding or removing tags on an EMR cluster that is tagged with the department of dev, eng, test, perf, or prod. This prevents users from creating a new EMR cluster by cloning the Quantitative Equity Group owned cluster into the new cluster

**Answers:** A, C

**Explanation:**

Option A is correct. If a user adds a tag to the cluster to which their user policy allows access, they can circumvent your security policies. This policy prevents adding new tags to the EMR cluster that would open up access to users with access to resources with that tag.

Option B is incorrect. Adding any tag to your confidential EMR cluster won’t allow access to any user. It will only allow access to users who have a policy associated with their IAM user account that allows access to resources that allow access via that tag

Option C is correct. Removing the tags from your confidential EMR cluster will allow access to users even if they are not in the Quantitative Equity Group IAM group because the policy will no longer have any tags.

Option D is incorrect. The logic of this option is flawed. By removing a tag from your confidential EMR cluster, any user who is allowed access to resources that have that tag will be denied access to your confidential EMR cluster.

Option E is incorrect. The Deny AddTags/RemoveTags statement does not prevent users from creating new EMR clusters, nor does it prevent users from attempting to clone a new cluster from another.

**Reference:**

Please see the Amazon EMR management guide titled **IAM Policies for Tag-Based Access to Clusters and EMR Notebooks** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-fine-grained-cluster-access.html>), the Amazon Elastic Map Reduce API reference titled **Actions** (<https://docs.aws.amazon.com/emr/latest/APIReference/API_Operations.html>), the Amazon EMR management guide titled **Tag Clusters** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-tags.html>), and the Amazon EMR management guide titled **Adding Tags to an Existing Cluster** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-tags-add.html>)

Question: 32

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Security

**Question text**:

You are a data scientist working for a medical services firm where you are building out an EMR cluster used to house the data lake used for your company’s client healthcare protected health information (PHI) data. The storage of this type of data is highly regulated through the Health Insurance Portability and Accountability Act (HIPAA). Specifically, HIPAA requires that healthcare companies, like your company, encrypt their client’s PHI data using encryption technology.

You have set up your EMR cluster to use the default of using the EMRFS to read and write your client’s PHI data to and from S3. You need to encrypt your client’s PHI data before you send it to S3.

Which option is the best encryption technique to use for your EMR cluster configuration?

1. SSE-S3
2. SSE-KMS
3. CSE-KMS
4. SSE-C

**Answer:** C

**Explanation:**

Option A is incorrect. When you use SSE-S3 to encrypt your data, EMR first sends your data to S3, then S3 encrypts the data with a CMK. Your requirement is to encrypt the data before you send it to S3.

Option B is incorrect. When you use SSE-KMS to encrypt your data, EMR first sends your data to S3, then S3 encrypts the data with a CMK. Your requirement is to encrypt the data before you send it to S3.

Option C is correct. When you use CSE-KMS to encrypt your data, EMR first encrypts the data with a CMK, then sends it to Amazon S3 for storage. This meets your requirement of encrypting your data before you send it to S3.

Option D is incorrect. EMR does not have an encryption mode that uses SSE-C.

**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 AWS Key Management Service developer guide titled **Encrypting data on the EMR file system (EMRFS)** (<https://docs.aws.amazon.com/kms/latest/developerguide/services-emr.html>), the Amazon EMR Management guide titled **Encryption Options** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-data-encryption-options.html>), and the blog titled **HIPAA Data at Rest Encryption Requirements** (<https://www.zettaset.com/blog/hipaa-data-at-rest-encryption-requirements/>)

Question: 33

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Security

**Question text**:

You are a data scientist working for a retail clothing manufacturer that has a large online presence through their retail website. The website gathers Personally Identifiable Information (PII), such as credit card numbers, when customers complete their purchases on the website. Therefore, your company must adhere to the Payment Card Industry Data Security Standard (PCI DSS). Your company wishes to store the client data and purchase information data gathered through these transactions in their data warehouse, running on Redshift, where they intend to build Key Performance Indicator (KPI) dashboards using QuickSight.

You and your security department know that your data collection system needs to obfuscate the PII (credit card) data, gathered through your data collection system. How should you protect the highly sensitive credit card data in order to meet the PCI DSS requirements while keeping your data collection system as efficient and cost effective as possible?

1. AWS Shield Advanced for website traffic
2. AWS WAF for website traffic
3. Tokenization PII data
4. Use GuardDuty for website traffic
5. KMS encryption in transit and at rest

**Answer:** C

**Explanation:**

Option A is incorrect. You can use AWS Shield to protect your retail website from distributed denial of service (DDoS) attacks, and Shield Advanced gives you a DDoS response team. However, Shield and Shield advanced won’t protect your PII credit card data from being exposed. You need to either encrypt your data or tokenize your customer’s PII data.

Option B is incorrect. You can use WAF to control the way traffic reaches your retail website by creating security rules that block common website attacks. However, WAF won’t protect your PII credit card data from being exposed. You need to either encrypt your data or tokenize your customer’s PII data.

Option C is correct. You can use tokenization instead of encryption when you only need to protect specific highly sensitive data for regulatory compliance requirements, such as PCI DSS.

Option D is incorrect. You can use GuardDuty to systematically monitor network traffic to detect anomalies in the behavior of your website users by using machine learning. However, GuardDuty won’t protect your PII credit card data from being exposed. You need to either encrypt your data or tokenize your customer’s PII data.

Option E is incorrect. Using KMS and encrypting your data in transit and at rest is more complex and costly than using tokenization on the specific PII data, the credit card data.

**Reference:**

Please see the AWS Big Data blog titled **Best practices for securing sensitive data in AWS data stores** (<https://aws.amazon.com/blogs/database/best-practices-for-securing-sensitive-data-in-aws-data-stores/>), the **AWS WAF overview page** (<https://aws.amazon.com/waf/?nc=bc&pg=pr>), the Wikipedia page titled **Payment Card Industry Data Security Standard** (<https://en.wikipedia.org/wiki/Payment_Card_Industry_Data_Security_Standard>), the **AWS GuardDuty** overview page (<https://aws.amazon.com/guardduty/>), and the **AWS Shield overview page** (<https://aws.amazon.com/shield/>)

Question: 34

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Security

**Question text**:

You are a data scientist working for a sports gambling company that produces sports betting data for inclusion in mainstream sports websites. Your company’s data is proprietary and needs to be protected for copyright purposes. You have been tasked with creating a data lake on S3 and also loading a relational database that stores your sports data. Any parameters (such as database connection information) used when building analytics applications used to access the data lake and/or database need to be stored in a secure service that encrypts the parameters. Your management team also has the requirement that parameters like database connection information be rotated automatically.

What AWS service should you use to protect the media content and metadata?

1. AWS IAM
2. AWS Systems Manager Parameter Store
3. KMS encryption
4. AWS Secrets Manager

**Answer:** D

**Explanation:**

Option A is incorrect. IAM can be used to manage passwords for AWS user accounts, but it is not a good choice for managing parameters like database connection information, and you can’t take advantage of encryption of your parameters and secrets with IAM without additional work on your part. A parameter or secrets management service such as Secrets Manager is a better choice.

Option B is incorrect. Systems Manager Parameter Store is great for storing parameters and even passwords. It can encrypt all parameters it stores, but the Systems Manager Parameter Store does not have the capability to automatically rotate your database connection information.

Option C is incorrect. KMS encryption is an obvious choice for encrypting your data, but it does not have the parameter and secret management capabilities that Secrets Manager gives you.

Option D is correct. Secrets Manager gives you the capability to encrypt your parameters, randomly generate passwords, and automatically rotate your database connection information.

**Reference:**

Please see the **AWS Secrets Manager FAQs** (<https://aws.amazon.com/secrets-manager/faqs/>), the **AWS Systems Manager FAQs** (<https://aws.amazon.com/systems-manager/faq/#Parameter_Store>), the AWS Identity and Access Management user guide titled **Managing Passwords** (<https://docs.aws.amazon.com/IAM/latest/UserGuide/id_credentials_passwords.html>), the Linux Academy article titled **An Inside Look at AWS Secrets Manager vs Parameter Store** (<https://linuxacademy.com/blog/amazon-web-services-2/an-inside-look-at-aws-secrets-manager-vs-parameter-store/>), and the AWS Security Blog titled **Rotate Amazon RDS database credentials automatically with AWS Secrets Manager** (<https://aws.amazon.com/blogs/security/rotate-amazon-rds-database-credentials-automatically-with-aws-secrets-manager/>)

Question: 35

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Security

**Question text**:

You are a data scientist working for a healthcare company that needs to comply with Health Insurance Portability and Accountability Act (HIPAA) regulations. Your company needs to take all of their patient’s data, including test diagnostic data, wearable sensor data, diagnostic data from all doctor visits, etc. and store it in a data lake. They then want to use Athena and other Business Intelligence (BI) tools to query the patient data to enable their healthcare providers to give optimal service to their patients.

In order to apply the appropriate data governance and compliance controls, what AWS service(s) will allow you to provide the appropriate (HIPAA) reports? Also, what AWS service(s) will allow you to monitor changes to your data lake S3 bucket ACLs and bucket policies to scan for public read/write access violations?

1. CloudTrail to gather the data for the Business Associate Addendum (BAA) HIPAA compliance report. Use custom rules in AWS Config to track and report on S3 ACL and/or bucket policy changes that violate your security policies.
2. CloudWatch to gather the data for the Business Associate Addendum (BAA) HIPAA compliance report. Use custom rules in AWS Resource Access Manager to track and report on S3 ACL and/or bucket policy changes that violate your security policies.
3. AWS Artifact to retrieve the Business Associate Addendum (BAA) HIPAA compliance report. Use custom rules in AWS Config to track and report on S3 ACL and/or bucket policy changes that violate your security policies.
4. AWS Artifact to retrieve the Business Associate Addendum (BAA) HIPAA compliance report. Use custom rules in AWS Resource Access Manager to track and report on S3 ACL and/or bucket policy changes that violate your security policies.

**Answer:** C

**Explanation:**

Option A is incorrect. CloudTrail logs all API access to your cloud resources, but it does not give you the information you need to produce the Business Associate Addendum (BAA) HIPAA compliance report.

Option B is incorrect. CloudWatch logs many important metrics and alerts regarding your AWS resources and services, but it does not give you the information you need to produce the Business Associate Addendum (BAA) HIPAA compliance report. Also, AWS Resource Access Manager gives you the ability to securely share AWS resources with another AWS account within your company, but it doesn’t allow you to monitor changes to your data lake S3 bucket ACLs and bucket policies for public read/write access violations.

Option C is correct. AWS Artifact gives you the capability to retrieve the Business Associate Addendum (BAA) HIPAA compliance report directly from AWS. Also, AWS Config monitors your AWS resource configuration changes. It allows you to take action or alert, using custom rules, on configuration changes that violate your policies.

Option D is incorrect. AWS Artifact gives you the capability to retrieve the Business Associate Addendum (BAA) HIPAA compliance report directly from AWS. However, AWS Resource Access Manager gives you the ability to securely share AWS resources with another AWS account or within your company, but it doesn’t allow you to monitor changes to your data lake S3 bucket ACLs and bucket policies for public read/write access violations.

**Reference:**

Please see the **AWS Artifact FAQs** (<https://aws.amazon.com/artifact/faq/>), the **AWS Resource Access Manager overview page** (<https://aws.amazon.com/ram/?c=sc&sec=srv>), the **AWS Config overview page** (<https://aws.amazon.com/config/>), and the AWS Security Blog titled **How to Use AWS Config to Monitor for and Respond to Amazon S3 Buckets Allowing Public Access** (<https://aws.amazon.com/blogs/security/how-to-use-aws-config-to-monitor-for-and-respond-to-amazon-s3-buckets-allowing-public-access/>)

Question: 36

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Security

**Question text**:

You are a data scientist working for a company that provides credit card verification services to banks and insurance companies. Your client credit card data is streamed into your S3 data lake on a daily basis in the form of large sets of JSON files. Due to the Personally Identifiable Information (PII) data contained in these JSON files, your company must adhere to the regulations defined in the Payment Card Industry Data Security Standard (PCI DSS). This means you must encrypt the data at rest in your S3 buckets. You also need to recognize and take action on any abnormal data access activity.

Which option best satisfies your data governance and compliance controls in the most cost effective manner?

1. Store the credit card JSON data in DynamoDB with encryption enabled on your tables. Use a Lambda function to determine if any of your compliance rules are violated by scanning the DynamoDB tables. When compliance rule violations are found, send alerts using Simple Notification Service (SNS).
2. Store the credit card JSON data in buckets in S3 with encryption enabled. Use a Lambda function to determine if any of your compliance rules are violated by scanning the S3 buckets. When compliance rule violations are found, send alerts using Simple Notification Service (SNS).
3. Store the credit card JSON data in buckets in S3 with encryption enabled. Use the AWS Macie service to determine if any of your compliance rules are violated by scanning the S3 buckets. When compliance rule violations are found, use CloudWatch events to trigger alerts sent via Simple Notification Service (SNS).
4. Store the credit card JSON data in DynamoDB with encryption enabled on your tables. Use the AWS Macie service to determine if any of your compliance rules are violated by scanning the DynamoDB tables. When compliance rule violations are found, use CloudWatch events to trigger alerts sent via Simple Notification Service (SNS).

**Answer:** C

**Explanation:**

Option A is incorrect. This option is not the most cost effective; using DynamoDB instead of using your data lake S3 buckets to store the data adds another layer of complexity and data storage cost. Also, writing your compliance rules into a Lambda function is not as cost effective or scalable as using the AWS Macie service.

Option B is incorrect. This option is not cost effective because you would have to write your compliance rules into a Lambda function, which is not as cost effective or scalable as using the AWS Macie service.

Option C is correct. Use the AWS Macie service to guard against security violations by continuously scanning your S3 bucket data and your account settings. Macie uses machine learning to properly classify your PII data. Macie also monitors access activity for your data, looking for access abnormalities and data leaks.

Option D is incorrect. The AWS Macie service works with data stored in S3, not DynamoDB. Also this option is not the most cost effective; using DynamoDB instead of using your data lake S3 buckets to store the data adds another layer of complexity and data storage cost.

**Reference:**

Please see the AWS News blog titled **New Amazon S3 Encryption and Security Features** (<https://aws.amazon.com/blogs/aws/new-amazon-s3-encryption-security-features/>), the **Amazon Macie overview page** (<https://aws.amazon.com/macie/>), and the **Amazon Macie FAQs page** (<https://aws.amazon.com/macie/faq/>)

Question: 37

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Security

**Question text**:

You are a data scientist working for a company that processes industrial machine operational data for various industrial manufacturers around the globe. You receive streaming data via Kinesis Data Firehose from the various manufacturers. You want to ingest the data into your Splunk cluster to deliver operational intelligence analysis, security analytics, and business performance KPIs for your manufacturing clients.

You have installed your Splunk cluster within your VPC. However, you have noticed that the ingestion process of moving your data from Kinesis Data Firehose to your Splunk cluster is failing. Which configuration option will allow your Kinesis Data Firehose stream to move your data into your Splunk cluster?

1. Change the Splunk cluster security group to allow access from Kinesis service IPs
2. Change your S3 bucket policy to allow Kinesis Data Firehose to write to your S3 bucket to allow access from service IPs
3. Verify the IAM role assigned to your Kinesis Data Firehose stream allows access from service IPs
4. Change your Splunk cluster ACL to allow access from Kinesis service IPs

**Answer:** A

**Explanation:**

Option A is correct. Since your Splunk cluster is in a VPC, you need to make your Splunk cluster publically accessible with a public IP address. Additionally, you need to unblock the Kinesis Data Firehose address. Kinesis Data Firehose has a set group of IP addresses depending on which region in which you have configured your VPC. For example, if your VPC is in US East Virginia, then the IP address is one of these CIDR blocks: 34.238.188.128/26, 34.238.188.192/26, or 34.238.195.0/26

Option B is incorrect. Kinesis Data Firehose does send your streaming data to S3, the bucket policy will not allow you to open port access on the Splunk cluster.

Option C is incorrect. The Kinesis Data Firehose IAM role will not allow you to open port access on the Splunk cluster.

Option D is incorrect. The Splunk cluster ACL is used to control the IP addresses that can access your Splunk cluster. You need to open the security group housing your Splunk cluster to the Kinesis Data Firehose service address.

**Reference:**

Please see the Kinesis Data Firehose developer guide titled **Controlling Access with Amazon Kinesis Data Firehose** (<https://docs.aws.amazon.com/firehose/latest/dev/controlling-access.html#using-iam-splunk-vpc>), the AWS Big Data blog titled **Power data ingestion into Splunk using Amazon Kinesis Data Firehose** (<https://aws.amazon.com/blogs/big-data/power-data-ingestion-into-splunk-using-amazon-kinesis-data-firehose/>), and the Splunk docs page titled **Securing Splunk Enterprise** (<https://docs.splunk.com/Documentation/Splunk/8.0.1/Security/Useaccesscontrollists>)

Question: 38

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Security

**Question text**:

You are a data scientist working for a large hedge fund. Your hedge fund managers rely on analytics data produced from the S3 data lake you have built that houses trade data produced by the firm’s various traders. You are configuring a public Elasticsearch domain that will allow your hedge fund managers to gain access to your trade data stored in your data lake. You have given your hedge fund managers Kibana to allow them to use visualizations you’ve produced to manage their traders activity.

When your hedge fund managers first test out your Kibana analytics visualizations, you find that Kibana cannot connect to your Elasticsearch cluster. Which options are ways to securely give your hedge fund managers access to your Elasticsearch cluster via their Kibana running on their local desktop? (SELECT TWO)

1. Configure a proxy server that acts as an intermediary between your Kibana users and your Elasticsearch cluster. Add an IP-based access IAM policy which allows requests from your user’s IP address to gain access to your Elasticsearch cluster through the proxy server’s IP address.
2. Configure a proxy server that acts as an intermediary between your Kibana users and your Elasticsearch cluster. Configure an open access IAM policy which allows requests from your user’s IP address to gain access to your Elasticsearch cluster through the proxy server’s IP address.
3. Configure a proxy server that acts as an intermediary between your Kibana users and your Elasticsearch cluster. Configure a security group which allows requests from your user’s IP address to gain access to your Elasticsearch cluster through the proxy server’s IP address.
4. Setup Amazon Cognito by creating a user pool and an identity pool to authenticate your Kibana users
5. Add an IP-based access IAM policy and use a security group which allows requests from your user’s IP address to gain access to your Elasticsearch cluster through the proxy server’s IP address.

**Answers:** A, D

**Explanation:**

Option A is correct. You can use a proxy server to avoid having to include all of your hedge fund manager’s IP addresses in your access policy. You only include the proxy server’s IP address in your IAM access policy with a policy statement segment like this:

{

...

"Effect": "Allow",

"Principal": {

"AWS": "\*"

},

"Action": "es:\*",

"Condition": {

"IpAddress": {

"aws:SourceIp": [

"57.201.547.32"

]

}

}

...

}

Where 57.201.547.32 is the IP address of your proxy server.

Option B is incorrect. An open access IAM policy will allow any user on the internet to make requests to put, get, post, and delete data from your Elasticsearch domain. This option is not secure.

Option C is incorrect. You can only use security groups to control access to Elasticsearch domains that are configured in a VPC. Your Elasticsearch domain is a public domain Elasticsearch cluster.

Option D is correct. You can use Cognito and its user pools and identity pools to provide username and password access for Kibana users.

Option E is incorrect. You can only use security groups to control access to Elasticsearch domains that are configured in a VPC. Your Elasticsearch domain is a public domain Elasticsearch cluster.

**Reference:**

Please see the Amazon Elasticsearch Service developer guide titled **Using a Proxy to Access Amazon ES from Kibana** (<https://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/es-kibana.html#es-kibana-proxy>), the Amazon Elasticsearch Service developer guide titled **Amazon Cognito Authentication for Kibana** (<https://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/es-cognito-auth.html>), the **Kibana overview page** (<https://aws.amazon.com/elasticsearch-service/the-elk-stack/kibana/>), the Kibana Docs guide titled **Explore Kibana using sample data** (<https://www.elastic.co/guide/en/kibana/current/tutorial-sample-data.html>), and the AWS Database blog titled **Set Access Control for Amazon Elasticsearch Service** (<https://aws.amazon.com/blogs/database/set-access-control-for-amazon-elasticsearch-service/>)

Question: 39

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Security

**Question text**:

You have just landed a new job as a data scientist for a worldwide retail and wholesale business with distribution centers located all around the globe.Your first assignment is to build a data collection system that stores all of the company’s product distribution performance data from all of their distribution centers into S3. You have been given the requirement that the data collected from the distribution centers must be encrypted at rest. You also have to load your distribution center data into your company’s analytics EMR cluster on a daily basis so that your management team can produce daily Key Performance Indicators (KPIs) for the various regional distribution centers.

Which option best meets your encryption at rest requirement?

1. Use an AWS KMS customer master key (CMK) for server side encryption when writing your distribution center performance data to S3. Create an IAM role that allows your analytics EMR cluster to have permission to access your S3 buckets and to use the AWS KMS CMK.
2. Use a customer provided key (SSE-C) for server side encryption when writing your distribution center performance data to S3. Create an IAM role that allows your analytics EMR cluster to have permission to access your S3 buckets and to use the SSE-C key.
3. Use client side encryption before you write your distribution center performance data to your S3 buckets. Save the client side encryption key in your analytics EMR cluster. Create an IAM role that allows your analytics EMR cluster to have permission to access your S3 buckets and to use the client side key.
4. Write your distribution center performance data to S3 and encrypt the data using a Kinesis Analytics job as it’s being written to the S3 buckets using server side encryption. Save the server side encryption key in one of your S3 buckets. Create an IAM role that allows your analytics EMR cluster to have permission to access your S3 buckets and to use the server side key.

**Answer:** A

**Explanation:**

Option A is correct. Using server side encryption through an AWS KMS CMK gives you the most secure encryption of the options provided.

Option B is incorrect. EMR does not support SSE-C

Option C is incorrect. You can not save encryption keys in your EMR cluster.

Option D is incorrect. Attempting to use Kinesis Analytics to encrypt your data is not a valid use of Kinesis Analytics. Also, you can not save your encryption keys in an S3 bucket.

**Reference:**

Please see the Amazon EMR management guide titled **Create Keys and Certificates for Data Encryption** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-encryption-enable.html#emr-awskms-keys>), the AWS Key Management Service developer guide titled **AWS Key Management Service concepts** (<https://docs.aws.amazon.com/kms/latest/developerguide/concepts.html#master_keys>), the AWS Key Management Service developer guide titled **How Amazon EMR uses AWS KMS** (<https://docs.aws.amazon.com/kms/latest/developerguide/services-emr.html>), and the AWS Big Data blog titled **Process Encrypted Data in Amazon EMR with Amazon S3 and AWS KMS** (<https://aws.amazon.com/blogs/big-data/process-encrypted-data-in-amazon-emr-with-amazon-s3-and-aws-kms/>)

Question: 40

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Collection

**Question text**:

You work as a data architect for a sports media data provider. Your company supplies sports data to sports gambling and sports gaming companies. These partner companies use the data your company provides to give their applications the detailed sports information needed to create reliable betting and realistic game simulation. These partners distribute their product as web and mobile applications. Your company currently gathers the data needed to create your sports data media content through a set of EC2 instances running in an auto-scaling group in your AWS account. All of the real-time ingestion, transformation, processing, and visualization of the data for your internal analysts is completed on these EC2 instances.

You need to improve this architecture by decoupling the real-time data collection system components because your company frequently experiences failures where important data is lost.

Which is the most cost effective and performant way to improve your architecture while decoupling your data collection components?

1. Load your sports data into S3 using Storage Gateway. Use EMR running Spark to ingest the data into your EMR cluster. Use the Redshift INSERT command to load the data from EMR into your Redshift cluster.
2. Load your sports data into S3 using Snowball@Edge. Use EMR running Spark to ingest the data into your EMR cluster. Use the Redshift UPDATE command to load the data from EMR into your Redshift cluster.
3. Load your sports data into S3 using Kinesis Data Firehose. Use EMR running Spark to ingest the data into your EMR cluster. Use the Redshift INSERT command to load the data from EMR into your Redshift cluster.
4. Load your sports data into S3 using Kinesis Data Firehose. Use EMR running Spark to ingest the data into your EMR cluster. Use the Redshift COPY command to load the data from EMR into your Redshift cluster.

**Answer:** D

**Explanation:**

Option A is incorrect. Storage Gateway is used to move data from your data center to S3. You would not use Storage Gateway to ingest real-time streaming data. Also, using the INSERT Redshift command will be much slower than using the Redshift COPY command.

Option B is incorrect. Snowball@Edge is used to move bulk data from your data center to S3. You would not use Snowball@Edge to ingest real-time streaming data. Also, you use the UPDATE Redshift command to update values in table columns, not to move new data into your Redshift cluster.

Option C is incorrect. Kinesis Data Firehose is the correct choice to ingest your sports data into your S3 data lake. However, using the INSERT Redshift command will be much slower than using the Redshift COPY command.

Option D is correct. Kinesis Data Firehose is the correct choice to ingest your sports data into your S3 data lake. Also, the Redshift COPY command is the most performant way to load your data into your Redshift cluster.

**Reference:**

Please see the Amazon EMR management guide titled **How to Get Data Into Amazon EMR** (<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-get-data-in.html>), the AWS Whitepaper titled Building Big Data Storage Solutions (Data Lakes) for Maximum Flexibility, specifically the sectiontitled **Data Ingestion Methods** (<https://docs.aws.amazon.com/whitepapers/latest/building-data-lakes/data-ingestion-methods.html>), the Amazon Redshift database developer guide titled **Step 6: Run the COPY command to load the data** (<https://docs.aws.amazon.com/redshift/latest/dg/load-from-emr-steps-run-copy.html>), the Amazon Redshift database developer guide titled **SQL Command**s (<https://docs.aws.amazon.com/redshift/latest/dg/c_SQL_commands.html>), and the AWS Snowball developer guide titled What is an **AWS Snowball Edge?** (<https://docs.amazonaws.cn/en_us/snowball/latest/developer-guide/whatisedge.html>)

Question: 41

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Collection

**Question text**:

You work as a data scientist for a government agency contracting firm that collects real-time polling data for various elections and public opinion items. You have built a streaming data collection architecture using Kinesis Data Streams and its Kinesis Producer Library (KPL). Your producer code is using the addUserRecord API call to add records which are eventually flushed to your Kinesis Data stream using the PutRecords API call. You have used the default settings for your PutRecords API KPL calls. Your Kinesis Data Stream PutRecords API call is occasionally experiencing partial and sometimes full failures. You have noticed that your data collection system sometimes experiences excessive retries, sometimes referred to as “retry spamming.”

What is the best approach to mitigate the request spamming resulting from your PutRecords retries?

1. Implement rate limiting for the offending producer by setting the token limit to 50% higher than the shard limit.
2. Expand the capacity of your Kinesis Data Stream while also implementing a suitable partition key strategy
3. Implement rate limiting for the offending producer by setting the threshold to 50% higher than the shard limit.
4. Implement rate limiting for the offending producer by setting the threshold to 30% higher than the shard limit.

**Answer:** B

**Explanation:**

Option A is incorrect. The KPL rate limiting feature limits the shard throughput for a producer. Rate limiting uses a token algorithm, but it doesn’t have the concept of a token limit. It uses a threshold limit, which by default is set to 50% higher than the shard limit.

Option B is correct. Lowering the rate limiting threshold is an approach you could use to reduce retry spamming, but the recommended approach is to expand the capacity of your Kinesis Data Stream while also implementing a suitable partition key strategy.

Option C is incorrect. You are using the default settings for your PutRecords KPL calls. The threshold limit default is 50% higher than the shard limit. So if you set the threshold to 50% you haven’t really changed anything. Also, the recommended approach to reduce retry spamming is to expand the capacity of your Kinesis Data Stream while also implementing a suitable partition key strategy.

Option D is incorrect. You could lower the rate limiting threshold from the default 50% to 30%, but the recommended approach to reduce retry spamming is to expand the capacity of your Kinesis Data Stream while also implementing a suitable partition key strategy.

**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 AWS Big Data blog titled **Implementing Efficient and Reliable Producers with the Amazon Kinesis Producer Library** (<https://aws.amazon.com/blogs/big-data/implementing-efficient-and-reliable-producers-with-the-amazon-kinesis-producer-library/>), and the Amazon Kinesis Data Streams developer guide titled **KPL Retries and Rate Limiting** (<https://docs.aws.amazon.com/streams/latest/dev/kinesis-producer-adv-retries-rate-limiting.html>)

Question: 42

**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 work as a data scientist for a data analytics firm that collects data for various industries, including the airline industry. Your airline clients wish to have your firm create analytics for use in machine learning models that predict air travel in the global market. To this end, you have created a Kinesis Data Streams data collection system that gathers flight data for use in your analysis.

You are writing a consumer application, using the Kinesis Client Library (KCL), that will consume the flight data stream records and process them before placing the data into your S3 data lake.

You need to handle the condition of when your consumer application fails in the middle of reading a data record from the data stream. What is the most efficient way to handle this condition?

1. Use a Lambda function to monitor the KCL consumer application’s log and restart the consumer application on failure
2. Use the KCL application state tracking feature implemented in the global DynamoDB table
3. Use the KCL application state tracking feature implemented in the DynamoDB table associated with the KCL application that failed when reading
4. Use the KCL application state tracking feature implemented in the DynamoDB table associated with the shard that the KCL consumer application was reading from when it failed

**Answer:** C

**Explanation:**

Option A is incorrect. This approach does not take advantage of the KCL application state tracking feature. Using a Lambda function to handle read failures is redundant and inefficient.

Option B is incorrect. The KCL application state tracking feature is implemented in a unique DynamoDB table that is associated with the KCL consumer application. The table is created using the name of the KCL consumer application. The feature does not use a global DynamoDB table.

Option C is correct. The KCL application state tracking feature is implemented in a unique DynamoDB table that is associated with the KCL consumer application. The table is created using the name of the KCL consumer application.

Option D is incorrect. The KCL application state tracking feature is implemented in a unique DynamoDB table that is associated with the KCL consumer application. The table is created using the name of the KCL consumer application. The DynamoDB table is not associated with the shard.

**Reference:**

Please see the Amazon Kinesis Data Streams developer guide titled **Tracking Amazon Kinesis Data Streams Application State** (<https://docs.aws.amazon.com/streams/latest/dev/kinesis-record-processor-ddb.html>), the Amazon Kinesis Data Streams developer guide titled **Developing Custom Consumers with Shared Throughput Using KCL** (<https://docs.aws.amazon.com/streams/latest/dev/shared-throughput-kcl-consumers.html#shared-throughput-kcl-consumers-overview>), and the Amazon Kinesis Data Streams developer guide titled **Reading Data from Amazon Kinesis Data Streams** (<https://docs.aws.amazon.com/streams/latest/dev/building-consumers.html>)

Question: 43

**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 scientist for a financial services firm that is building an automated trading system using data streamed from market data sources. The market data records the market data sources produce are small in size (512 bytes) and are sent very rapidly (1,500 records per second) to your Kinesis Producer Library based producer application. You have your data collection system configured like this:

Market data source -> KPL producer application -> Kinesis Data stream -> Kinesis Data Firehose stream -> Lambda -> S3

Your Lambda function transforms the market data for use in your automated trading system.

At the size and rate of production of your market data records, your data collection pipeline is constrained. Why is it constrained, and what can you do to remove the constraint?

1. The pipeline is constrained because a Kinesis Data Streams shard has a limit of supporting up to 1,000 records per second, or 1 MB throughput. Increase the throughput by using KPL aggregation to aggregate the 1,500 market data records into 10 Kinesis Data Stream records, bringing the records per second to 10, each of which will hold 76 KB.
2. The pipeline is constrained because you can only have 50 Kinesis Data Firehose delivery streams per region. Increase the throughput by using Service Quota to request a quota increase.
3. The pipeline is constrained because your Lambda function has a limit of only 15 minutes runtime. Increase the Lambda function capacity by leveraging the messaging fanout pattern for Lambda using SNS.
4. The pipeline is constrained because you are attempting to send greater than 1 MB of data per second through your shard. Compress your records in your KPL code before sending your records to your Kinesis Data Streams stream. This will remove the message volume constraint.

**Answer:** A

**Explanation:**

Option A is correct. Using the KPL aggregation feature allows you to overcome the 1,000 records per second limitation by combining your market data records in your KPL code before you write them to your Kinesis Data Streams stream. Aggregating them into 10 chunks removes the 1,000 records per second constraint. Kinesis Data Firehose de-aggregates the records before delivering them to your Lambda function, which transforms the data before saving it to S3.

Option B is incorrect. You are only using one Kinesis Data Firehose stream, so you would not run into the quota limit per region.

Option C is incorrect. The messaging fanout pattern for Lambda using SNS is useful for designing loosely coupled interaction between your Lambda functions. This pattern will not help you solve the throughput limitation of your data collection pipeline.

Option D is incorrect. Even if you compress your 1,500 records to get their combined message size per second under 1 MB, you will still be attempting to send more than the 1,000 records per second limit through your Kinesis Data Streams shard.

**Reference:**

Please see 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>), 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>), the Amazon Kinesis Data Streams developer guide titled **Aggregation** (<https://docs.aws.amazon.com/streams/latest/dev/kinesis-kpl-concepts.html#kinesis-kpl-concepts-aggretation>), 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 Firehose developer guide titled **Amazon Kinesis Data Firehose Quota** (<https://docs.aws.amazon.com/firehose/latest/dev/limits.html>), the Service Quotas user guide titled **What is Service Quotas?** (<https://docs.aws.amazon.com/servicequotas/latest/userguide/intro.html>), and the AWS Compute blog titled **Messaging Fanout Pattern for Serverless Architectures Using Amazon SNS** (<https://aws.amazon.com/blogs/compute/messaging-fanout-pattern-for-serverless-architectures-using-amazon-sns/>)

Question: 44

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Collection

**Question text**:

You work as a data scientist for an oil refining company. Your team is building a data lake in S3 which will be used to do complex analysis of crude oil chemical compounds. Your company will use this analysis to improve their product to make it more cost effective. Your data lake will have many disparate sources of compound data that need to be loaded into your S3 buckets. You have decided to use AWS Glue to crawl your data sources to allow for the use of Glue transform jobs to process the data while loading it into your S3 buckets.

When you run your Glue crawler on one of your RDS instances you are getting a resource unavailable error. What might be the root cause of this problem?

1. The ApplyMapping transform is able to map the source columns from your DynamicFrame to your target columns.
2. You have tried to use a JDBC connection to access your RDS instance. Glue crawlers connect to RDS instances using the RDS native interface.
3. You have not opened up all ports to TCP on your database security group.
4. Your custom classifier used by your Glue crawler is not recognizing the structure of your RDS data.

**Answer:** C

**Explanation:**

Option A is incorrect. If your transform job is accessing your source columns, then you are able to reach your source data repository, RDS in this instance. The ApplyMapping transform is a built-in transform that Glue provides.

Option B is incorrect. Glue crawlers connect to RDS using a JDBC connection, not a native interface.

Option C is correct. Glue crawlers need all TCP ports open on the security group where the data source resides, your RDS instance in this scenario. To protect your database security group from outside access via a TCP port, you also configure a self-referencing inbound rule for all TCP ports.

Option D is incorrect. If your Glue crawler has reached your RDS instance and has attempted to infer its schema, then the crawler has gained access to the resource. Therefore, a resource unavailable message would not be appropriate.

**Reference:**

Please see the AWS Glue developer guide titled **Troubleshooting Errors in AWS Glue** (<https://docs.aws.amazon.com/glue/latest/dg/glue-troubleshooting-errors.html#error-resource-unavailable>), the AWS Glue developer guide titled **Troubleshooting Connection Issues in AWS Glue** (<https://docs.aws.amazon.com/glue/latest/dg/troubleshooting-connection.html>), the AWS Glue developer guide titled **Setting Up a VPC to Connect to JDBC Data Stores** (<https://docs.aws.amazon.com/glue/latest/dg/setup-vpc-for-glue-access.html>), the AWS Glue developer guide titled **Defining Crawlers** (<https://docs.aws.amazon.com/glue/latest/dg/add-crawler.html>), the AWS Glue developer guide titled **Built-in Transform**s (<https://docs.aws.amazon.com/glue/latest/dg/built-in-transforms.html>), and the AWS Glue developer guide titled **Populating the AWS Glue Data Catalog** (<https://docs.aws.amazon.com/glue/latest/dg/populate-data-catalog.html>)

Question: 45

**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 scientist for an alternative energy source company. Your company has several fields of wind turbines located across several continents and in the open Atlantic and Pacific oceans. You are responsible for implementing the data collection system that feeds turbine sensor data to your analytics platform in real-time for use in preventative maintenance analytics applications for the turbines. These analytics applications are used to schedule maintenance in response to changes in the turbine sensor data in real-time. This allows your company to address turbine low output situations, thereby helping to maximize revenue.

Which data collection architecture handles the frequency, volume, and source of your data while also delivering the real-time analytics needed by the turbine preventive maintenance analytics application in the most cost effective manner?

1. Sensor data -> AWS IoT -> S3 -> Lambda -> AWS Batch -> EMR -> Turbine Analytics App
2. Sensor data -> AWS IoT -> S3 -> Lambda -> Kinesis Data Firehose -> Redshift -> Turbine Analytics App
3. Sensor data -> AWS IoT -> Kinesis Data Firehose -> Lambda -> Redshift -> Turbine Analytics App
4. Sensor data -> AWS IoT -> Kinesis Data Firehose -> Lambda -> DynamoDB -> Turbine Analytics App

**Answer:** C

**Explanation:**

Option A is incorrect. This architecture has unnecessary points of latency in the flow. The S3 bucket-to-Lambda interaction coupled with the Lambda-AWS Batch interaction will introduce significant latency in a data collection system that needs to feed real-time analytics.

Option B is incorrect. This option also introduces unnecessary latency into the data collection system. The S3 and Lambda components are unnecessary; you can move your data from AWS IoT directly to Kinesis Data Firehose.

Option C is correct. Using Kinesis Data Firehose to stream your IoT data received from AWS IoT, transforming the data using Lambda, and copying the data into Redshift is the most efficient option.

Option D is incorrect. DynamoDB is not a valid destination data store for Kinesis Data Firehose. Kinesis Data Firehose can stream data to S3, Redshift, Elasticsearch, or Splunk.

**Reference:**

Please see the **AWS Batch overview page** (<https://aws.amazon.com/batch/>), the **AWS IoT overview page** (<https://aws.amazon.com/iot/>), the AWS Kinesis Data Firehose developer guide titled **Select Destination** (<https://docs.aws.amazon.com/firehose/latest/dev/create-destination.html>), the Amazon Kinesis Data Firehose developer guide titled **Writing to Kinesis Data Firehose Using AWS IoT** (<https://docs.aws.amazon.com/firehose/latest/dev/writing-with-iot.html>), and the **Amazon Kinesis Data Firehose FAQs** (<https://aws.amazon.com/kinesis/data-firehose/faqs/>)

Question: 46

**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 scientist for a utility company that is implementing real-time management of its electricity meters at the homes of its customers. These meters have sensors on them that transmit usage and other measurements back to your data collection system using AWS IoT. Your management team wishes to use this IoT data to perform analytics and build Key Performance Indicator (KPI) dashboards to help give better service to their customers.

You do not need to transform the IoT data before feeding it into your Redshift cluster. Which architecture option is the most cost effective and efficient?

1. Sensor data -> AWS IoT -> KPL Application -> Redshift -> BI Tools
2. Sensor data -> AWS IoT -> S3 -> Lambda -> Kinesis Data Firehose -> Redshift -> BI Tools
3. Sensor data -> AWS IoT -> Kinesis Data Firehose -> Lambda -> Redshift -> BI Tools
4. Sensor data -> AWS IoT -> Kinesis Data Firehose -> S3 -> Redshift -> BI Tools

**Answer:** D

**Explanation:**

Option A is incorrect. This option is functionally incorrect. Your Kinesis Producer Library application cannot write directly to Redshift. The KPL app has to write to a Kinesis Data Streams shard, which is not present in the proposed architecture.

Option B is incorrect. This option is overly complicated. Storing your IoT data on S3 and triggering a Lambda function are unnecessary steps. Also, to COPY the data from Firehose to Redshift, and intermediary S3 bucket is needed.

Option C is incorrect. Since you don’t need to transform your IoT data, the Lambda function is not required.

Option D is correct. AWS IoT can be configured to have an action to send the streamed IoT data directly to a Kinesis Data Firehose stream. The Firehose stream can then store the streamed data into an S3 bucket and then issue the Redshift COPY command to load the data into your Redshift cluster.

**Reference:**

Please see the Amazon Kinesis Data Firehose developer guide titled **What is Amazon Kinesis Data Firehose?** (<https://docs.aws.amazon.com/firehose/latest/dev/what-is-this-service.html#data-flow-diagrams>), the Amazon Kinesis Data Firehose developer guide titled **Writing to Kinesis Data Firehose Using AWS IoT** (<https://docs.aws.amazon.com/firehose/latest/dev/writing-with-iot.html>), the **AWS IoT overview page** (<https://aws.amazon.com/iot/>), and the **Amazon Kinesis Data Firehose FAQs** (<https://aws.amazon.com/kinesis/data-firehose/faqs/>)

Question: 47

**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 lead data scientist on a development team in a large consulting firm. Your team is working on a contract for a client that needs to gather key statistics from their application server logs. This data needs to be loaded into their S3 data lake for use in analytics applications.

Your data collection process requires transformation of the streamed data records as they are ingested through the collection process. You also have the requirement to keep an unaltered copy of every source record ingested by your data collection process.

Which option meets all of your requirements in the most efficient manner?

1. Kinesis Data Stream KPL application streams source data into Kinesis Data Firehose. Kinesis Data Firehose invokes a Lambda function which transforms the data record. Kinesis Data Firehose then writes the transformed record to the S3 data lake destination. Kinesis Data Firehose saves the unaltered record to another S3 destination.
2. Amazon Kinesis Agent streams source data into Kinesis Data Firehose. Kinesis Data Firehose invokes a Lambda function which transforms the data record. Kinesis Data Firehose then writes the transformed record to the S3 data lake destination. Kinesis Data Firehose saves the unaltered record to another S3 destination.
3. Amazon Kinesis Agent streams source data into Kinesis Data Firehose. Kinesis Data Firehose invokes a Lambda function which transforms the data record. The Lambda function then writes the transformed record to the S3 data lake destination. Kinesis Data Firehose saves the unaltered record to another S3 destination.
4. Amazon Kinesis Agent streams source data into Kinesis Data Firehose. Kinesis Data Firehose transforms the data record and writes the transformed record to the S3 data lake destination. Kinesis Data Firehose saves the unaltered record to another S3 destination.

**Answer:** B

**Explanation:**

Option A is incorrect. To stream application log data to a Kinesis stream, the most efficient way is to use the Amazon Kinesis Agent. Also, a KPL application can only send data to a Kinesis Data Streams shard, not a Kinesis Data Firehose.

Option B is correct. The Amazon Kinesis Agent is the most efficient way to collect data from application log files and send the data to a Kinesis Data Firehose stream. To transform the log data, Kinesis uses a Lambda function. Once the Lambda function has transformed the data, the Lambda function returns the transformed data record to Kinesis Data Firehose which writes the transformed record to your S3 destination. Kinesis Data Firehose can also be configured to write the original source data record to another S3 bucket.

Option C is incorrect. This option is missing a step. The Lambda function doesn’t write the transformed record to your S3 destination, the Lambda function returns the transformed data record to Kinesis Data Firehose which writes the transformed record to your S3 destination.

Option D is incorrect. This option is missing a step. Kinesis Data Firehose cannot transform your records without the use of a Lambda function. To transform the log data, Kinesis uses a Lambda function. Once the Lambda function has transformed the data, the Lambda function returns the transformed data record to Kinesis Data Firehose which writes the transformed record to your S3 destination.

**Reference:**

Please see the Amazon Kinesis Data Firehose developer guide titled **What is Amazon Kinesis Data Firehose?** (<https://docs.aws.amazon.com/firehose/latest/dev/what-is-this-service.html#data-flow-diagrams>), the Amazon Kinesis Data Firehose developer guide titled **Writing to Kinesis Data Firehose Using Kinesis Agent** (<https://docs.aws.amazon.com/firehose/latest/dev/writing-with-agents.html>), theAmazon Kinesis Data Firehose developer guide titled **Amazon Kinesis Data Firehose Data Transformation** (<https://docs.aws.amazon.com/firehose/latest/dev/data-transformation.html>), and the Medium article titled **Amazon Kinesis Firehose- Send your Apache logs to S3** (<https://medium.com/tensult/amazon-kinesis-firehose-send-your-apache-logs-to-s3-26876f7cac84>)

Question: 48

**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 scientist on a development team in a large bank. The bank has asked your team to move your existing on-prem customer account Oracle database to a PostgreSQL RDS instance running in your company’s AWS account. In the process of moving your on-prem database to your AWS RDS instance you cannot take your Oracle database offline. Also, you need to perform some transformations of your database schema as you move to your new instance, including adding new primary keys to certain columns of the customer table and changing some of the data types of some of the target columns.

How would you use AWS services to accomplish this transformation in the most efficient manner?

1. Create a Glue crawler that crawls the on-prem Oracle instance and catalogs the schema of the customers data. Then write a Glue ETL job that extracts the data from the Oracle instance, transforms the tables and creates the primary key, and finally loads the data into your PostgreSQL RDS instance.
2. Create an AWS Schema Conversion Tool (SCT) project that defines the transformations you need to accomplish. Use the AWS SCT to copy your on-prem Oracle database to your RDS PostgreSQL database.
3. Use the AWS Database Migration Service (DMS) to migrate your Oracle database to your PostgreSQL RDS instance. Use the ongoing replication change data capture (CDC) feature of DMS to allow the migration to capture ongoing changes to your source database as you copy to your target database. Perform your transformations using the table-mapping rules defined in your DMS tasks.
4. Use an Amazon Kinesis Agent to stream your source data into Kinesis Data Firehose. Kinesis Data Firehose transforms the data record and writes the transformed data into your RDS PostgreSQL instance.

**Answer:** C

**Explanation:**

Option A is incorrect. Using Glue to extract, transform, and load your data from your on-prem Oracle database to your PostgreSQL RDS instance is not the most efficient option. This option requires you to code the Glue ETL transformation job. The option using AWS DMS requires much less effort on your part.

Option B is incorrect. When moving from on-prem to RDS, the Amazon Schema Conversion tool only converts the source schema to the target schema. You would also need to use a database migration service together with SCT to complete your task.

Option C is correct. This AWS DMS service is the correct choice since it is the most efficient at migrating on-prem heterogeneous databases to RDS. Since DMS has data transformation replication tasks, you can perform your primary key and data type changes directly within your DMS data migration.

Option D is incorrect. The Kinesis Agent is used to monitor files. You wouldn’t use it to try to monitor and capture changes to your Oracle database instance. Also, you would have to somehow get your foundational data, the data that hasn’t changed, from your Oracle instance to your PostgreSQL RDS instance. This approach would be much less efficient than the DMS approach.

**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>), the AWS Database Migration Service user guide titled **Specifying Table Selection and Transformations by Table Mapping Using JSON** (<https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Tasks.CustomizingTasks.TableMapping.html#CHAP_Tasks.CustomizingTasks.TableMapping.SelectionTransformation>), 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.Transformations>), the AWS Database Migration Service user guide titled **How AWS Database Migration Service Works** (<https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Introduction.html>), the AWS Schema Conversion Tool user guide titled **Converting Database Schemas Using the AWS Schema Conversion Tool** (<https://docs.aws.amazon.com/SchemaConversionTool/latest/userguide/CHAP_Converting.html>), the AWS Database blog titled **How to extract, transform, and load data for analytic processing using AWS Glue (Part 2)** (<https://aws.amazon.com/blogs/database/how-to-extract-transform-and-load-data-for-analytic-processing-using-aws-glue-part-2/>), the AWS Big Data blog titled **How to access and analyze on-premises data stores using AWS Glue** (<https://aws.amazon.com/blogs/big-data/how-to-access-and-analyze-on-premises-data-stores-using-aws-glue/>), the AWS Prescriptive Guidance Oracle database migration guide titled **Heterogeneous database migration** (<https://docs.aws.amazon.com/prescriptive-guidance/latest/migration-oracle-database/heterogenous-migration.html>), and the AWS Kinesis Data Firehose developer guide titled **Writing to Kinesis Data Firehose Using Kinesis Agent** (<https://docs.aws.amazon.com/firehose/latest/dev/writing-with-agents.html>)

Question: 49

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Processing

**Question text**:

You work as a data scientist for a farming collective that gathers data from small farms all across the country. The data is used in analytics applications to help the farmer members of the collective better understand their market. Your team has been assigned the task of migrating the collective’s legacy database to an RDS Aurora database on AWS. Since your collective is expanding into other parts of the globe, you want to take advantage of Aurora’s global database feature.

You are running a full load and ongoing replication change data capture (CDC) task to migrate the data from your legacy database instance to your new RDS instance. However, the migration tasks are running very slowly. What is a probable reason for the slowness and what can you do to correct it? (SELECT TWO)

1. The process of capturing changes from the source legacy database instance is suboptimal. Use CloudWatch metrics to monitor CDCLatencySource to find source activity that is interfering with your DMS performance, such as if a transaction has started, but has not committed.
2. The source legacy database must be of the same engine type as the target database engine type. Change your target database to the same instance type as your legacy database instance type.
3. Your VPC database subnet group is only configured to use two availability zones. Add an additional availability zone to the subnet group.
4. Your DMS replication instance is running on an EC2 instance in your default VPC.
5. The process of applying change events to the target instance is suboptimal. Use CloudWatch metrics to monitor CDCLatencyTarget to find target issues that are interfering with your DMS performance, such as lack of primary keys or indexes on the target.

**Answers:** A, E

**Explanation:**

Option A is correct. One of the ways your migration tasks can slow down is because your source latency is high. To discover the problem and understand how to fix it, you use your CloudWatch log entries to find the delay in capturing from your source. One example is transactions that have started, but have not committed.

Option B is incorrect. AWS DMS does not require that the source database and the target database be of the same instance type. Also, if you change your target instance type from Aurar to some other type you will no longer be able to take advantage of Aurora’s global database feature.

Option C is incorrect. According to the AWS documentation, in order to use your DB subnet group, it needs to be in at least two Availability Zones. Your subnet group is already across two availability zones.

Option D is incorrect. The DMS replication instance can run on an EC2 instance in either your default VPC or another VPC in your account.

Option E is correct. Another way your migration tasks can run slow is when your target latency is high. To discover the problem and understand how to fix it, you use your CloudWatch log entries to find the delay in applying changes to your target. An example is when you have no primary keys or indexes on the target.

**Reference:**

Please see the **Amazon Aurora overview page** (<https://aws.amazon.com/rds/aurora/>), the AWS knowledge center page titled **How can I troubleshoot high target latency on an AWS DMS task?** (<https://aws.amazon.com/premiumsupport/knowledge-center/dms-high-target-latency/>), the AWS Database Migration Service user guide titled **Monitoring AWS DMS Tasks** (<https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Monitoring.html#CHAP_Monitoring.Metrics.Task>), the Amazon Aurora user guide for Aurora titled **Working with a DB Instance in a VPC** (<https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/USER_VPC.WorkingWithRDSInstanceinaVPC.html>), the Amazon Aurora user guide for Aurora titled **Scenarios for Accessing a DB Instance in a VPC** (<https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/USER_VPC.Scenarios.html>), the **AWS Database Migration Service overview page** (<https://aws.amazon.com/dms/>), and the AWS Database Migration Service user guide titled **Setting Up a Network for a Replication Instance** (<https://docs.aws.amazon.com/dms/latest/userguide/CHAP_ReplicationInstance.VPC.html>)

Question: 50

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Processing

**Question text**:

You work as a data scientist for a streaming music service. Your company wishes to catalog and analyze the metadata about the most frequently streamed songs in their catalog. To do this you have created a Glue crawler that you have scheduled to crawl the company song database every hour. You want to load the song play statistics and metadata into your Redshift data warehouse using a Glue ETL job as soon as the crawler completes.

What is the most efficient way to automatically start the Glue ETL job as soon as the crawler completes?

1. Create a Glue Workflow. Create two triggers, one for the crawler and one for the ETL job. The workflow will start the crawler automatically based on a timer trigger. Once the crawler completes, the second trigger, an event trigger, will watch for the event of the crawler completing and start the ETL job. The ETL job will transform and load the data into your Redshift cluster.
2. Create a Glue trigger, a Lambda function, and a CloudWatch events rule. The Glue trigger is a timed based event that triggers every hour and starts the crawler. The CloudWatch event watches for the crawler state change. When the crawler state changes to complete, the CloudWatch event automatically triggers the Lambda function which starts the Glue ETL job. The ETL job will transform and load the data into your Redshift cluster.
3. Create two Glue triggers. The first Glue trigger is a timed based event that triggers every hour and starts the crawler. The second Glue trigger watches for the crawler to reach the COMPLETED state and then starts the ETL job. The ETL job will transform and load the data into your Redshift cluster.
4. Create two Glue triggers. The first Glue trigger is a timer based event that triggers every hour and starts the crawler. The second Glue trigger watches for the crawler to reach the SUCCEEDED state and then starts the ETL job. The ETL job will transform and load the data into your Redshift cluster.

**Answer:** D

**Explanation:**

Option A is incorrect. The one problem with this option is that the Glue Workflow requires that you start the crawler on demand from the Workflows page on the AWS Glue console. This is not an automated step in your workflow so this option does not satisfy your requirements, nor is it the most efficient option.

Option B is incorrect. This option automates all of the steps in your crawl, extract, transform, load process. However, it requires you to write a Lambda function to trigger your ETL job. This is not as efficient as using the Glue triggers approach described in option D.

Option C is incorrect. This option describes the most efficient approach, the use of Glue triggers, but it has the wrong state for watching the Glue crawler. There is no COMPLETED state. The proper state is SUCCEEDED.

Option D is correct. This option describes the most efficient approach, the use of Glue triggers, to start both the crawler and the ETL job automatically.

**Reference:**

Please see the AWS Glue developers guide titled **AWS Glue Concepts** (<https://docs.aws.amazon.com/glue/latest/dg/components-key-concepts.html>), the AWS Knowledge Center article titled **How can I automatically start an AWS Glue job when a crawler run completes?** (<https://aws.amazon.com/premiumsupport/knowledge-center/start-glue-job-run-end/>), the AWS Knowledge Center article titled **How can I use AWS Glue workflows to automatically start a job when a crawler run completes?** (<https://aws.amazon.com/premiumsupport/knowledge-center/start-glue-job-after-crawler-workflow/>), the AWS Knowledge Center article titled **How can I use a Lambda function to automatically start an AWS Glue job when a crawler run completes?** (<https://aws.amazon.com/premiumsupport/knowledge-center/start-glue-job-crawler-completes-lambda/>), the AWS Glue developers guide titled **Starting Jobs and Crawlers Using Triggers** (<https://docs.aws.amazon.com/glue/latest/dg/trigger-job.html>), the AWS Glue developers guide titled **Performing Complex ETL Activities Using Workflows in AWS Glue** (<https://docs.aws.amazon.com/glue/latest/dg/orchestrate-using-workflows.html>), the AWS Glue developers guide titled **Overview of Workflows in AWS Glue** (<https://docs.aws.amazon.com/glue/latest/dg/workflows_overview.html>), and the AWS Glue developers guide titled **AWS Glue Triggers** (<https://docs.aws.amazon.com/glue/latest/dg/about-triggers.html>)

Question: 51

**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 scientist for a mobile gaming company. Your company wants to run analytics on their gaming community’s in-app purchase rates. Specifically, they want to be able to use data collected from their users' gaming activity to feed a machine learning model. The machine learning model will allow them to classify gamers in various categories to be used in a recommendation engine that will recommend in-app purchases based on gamer activity.

You are building a Glue ETL job to extract the gaming data from your player data DynamoDB database and perform machine learning transformations on the data such as feature engineering. When defining your Glue ETL job properties, which worker type should you choose for your Glue data processing units (DPUs) so that you optimize cost while still delivering the correct type of processor for your workload?

1. Standard
2. G.1X
3. G.2X
4. M.2X

**Answer:** C

**Explanation:**

Option A is incorrect. The Standard option for the Glue ETL worker type you only specify the maximum capacity, which is the number of DPUs that can be allocated when you run your job. The Standard option uses 50 GB disk and 2 executors. This option doesn’t give you enough control over your costs or your processing needs, specifically your machine learning transformation needs.

Option B is incorrect. This G.1X option is recommended for memory-intensive jobs. This option uses 1 DPU (4 vCPU, 16 GB of memory, 64 GB disk) per worker. While capable, this option does not give you the worker performance capacity needed for your machine learning transform workload use case.

Option C is correct. This G.2X option is recommended for jobs that run machine learning transforms. This option uses 2 DPU (8 vCPU, 32 GB of memory, 128 GB disk per worker. You get the larger DPU capacity that is best suited to your specific use case.

Option D is incorrect. There is no M.2X worker type for a Glue job.

**Reference:**

Please see the AWS Glue developers guide titled **Jobs** (<https://docs.aws.amazon.com/glue/latest/dg/aws-glue-api-jobs-job.html>), the AWS Glue developers guide titled **Adding Jobs in AWS Glue** (<https://docs.aws.amazon.com/glue/latest/dg/add-job.html>), and the **AWS Databases for Games overview page** (<https://aws.amazon.com/gametech/databases/>)

Question: 52

**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 scientist for a financial services company that uses market data producer services to calibrate their market trading systems. You receive trade data, equity and fixed income security pricing data, and other time series data that requires proper ordering when received into your data lake. You and your data science team use the data lake to build analytics applications used by your trading management team to assess trading risk and overall firm liquidity. As the data you receive from your market data providers is streamed at a message per second rate, your data analytics applications that use your data lake can tolerate occasionally missing a data record.

Which AWS streaming service will meet your needs for ingesting the market data in the most efficient manner? (SELECT TWO)

1. Kinesis Data Streams
2. Kinesis Data Firehose
3. MSK
4. SQS FIFO
5. SQS Standard

**Answers:** A, C

**Explanation:**

Option A is correct. Kinesis Data Streams guarantees the order of your messages is maintained, and it guarantees at-least-once delivery.

Option B is incorrect. Kinesis Data Firehose guarantees at-least-once delivery, but it does not guarantee that the order of your messages is maintained.

Option C is correct. MSK, or Amazon Managed Streaming for Apache Kafka, guarantees the order of your messages is maintained, and it guarantees at-least-once delivery.

Option D is incorrect. SQS FIFO guarantees the order of your messages is maintained, and it guarantees exactly-once delivery, but alone it is not a data stream collection service. You would have to write software to receive the incoming messages from your market data providers and then place the data records onto the SQS queue.

Option E is incorrect. SQS Standard guarantees at-least-once delivery, but it does not guarantee that the order of your messages is maintained. Also, alone it is not a data stream collection service. You would have to write software to receive the incoming messages from your market data providers and then place the data records onto the SQS queue.

**Reference:**

Please see the Amazon Kinesis Data Streams developer guide titled **Handling Duplicate Records** (<https://docs.aws.amazon.com/streams/latest/dev/kinesis-record-processor-duplicates.html>), the **Amazon Managed Streaming for Apache Kafka (Amazon MSK) overview page** (<https://aws.amazon.com/msk/>), the **Amazon SQS FAQs page** (<https://aws.amazon.com/sqs/faqs/>), the Amazon Managed Streaming for Apache Kafka page titled **What is Apache Kafka?** (<https://aws.amazon.com/msk/what-is-kafka/>), and the **Amazon Kinesis overview page** (<https://aws.amazon.com/kinesis/>)

Question: 53

**Main​ ​Topic​ ​:​** Data Analytics

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

**Domain:** Processing

**Question text**:

You work as a data scientist for a sports analytics company that produces statistics for use in sports media production. You receive streaming data from all of the major sports leagues around the world. The data typically is streamed into your data collection pipeline from the providers in either the KPL or GZIP format. Your sports analytics are done in real-time using Kinesis Data Analytics as you ingest the streamed data. On several data sources you need to expand data elements represented as strings into multiple columns. This allows your analytics to present the statistics at a more granular level.

Which data collection architecture best suits your real-time analytics requirements with the least amount of work on your part?

1. Use Kinesis Data Firehose to ingest the data. Use the Kinesis Data Firehose data transformation feature that uses Lambda to convert the data from the KPL or GZIP format into either the XML, TSV, or CSV format. Also, use the Lambda Kinesis Data Firehose data transformation feature to expand the strings into separate columns.
2. Use Kinesis Data Streams to ingest the data. Use the Kinesis Data Streams data transformation feature that uses Lambda to convert the data from the KPL or GZIP format into either the JSON, TSV, or CSV format. Also, use the Lambda Kinesis Data Streams data transformation feature to expand the strings into separate columns.
3. Use Kinesis Data Firehose to ingest the data. Use the Kinesis Data Firehose data transformation feature that uses Lambda to convert the data from the KPL or GZIP format into either the CSV, TSV, or JSON format. Also, use the Lambda Kinesis Data Firehose data transformation feature to expand the strings into separate columns.
4. Use Kinesis Data Firehose to ingest the data. Use the Kinesis Data Firehose data transformation feature that uses Lambda to convert the data from the KPL or GZIP format into either the CSV, TSV, or SAML format. Also, use the Lambda Kinesis Data Firehose data transformation feature to expand the strings into separate columns.

**Answer:** C

**Explanation:**

Option A is incorrect. While Kinesis Data Firehose is a good choice to ingest the data, and using the Firehose Lambda integration to transform your data also makes sense. Kinesis Data Analytics supports the CSV, TSV, and JSON formats. Therefore, this option is incorrect since it mentions converting to the XML format.

Option B is incorrect. While you can use Lambda to read records from your Kinesis data stream, this will require more work on your part when compared to using the Kinesis Data Firehose data transformation feature, which integrates with Lambda in a much simpler manner.

Option C is correct. Using the Kinesis Data Fire data transformation feature that leverages Lambda to convert your file format and to expand your data elements is the solution that requires the least amount of effort on your part. Many of the common tasks, such as record format conversion, and record data element expansion, are already written for you in the Lambda blueprints provided for use with the Kinesis Data Firehose data transformation feature.

Option D is incorrect. While Kinesis Data Firehose is a good choice to ingest the data, and using the Firehose Lambda integration to transform your data also makes sense. Kinesis Data Analytics supports the CSV, TSV, and JSON formats. Therefore, this option is incorrect since it mentions converting to the SAML format.

**Reference:**

Please see the **Amazon Kinesis Data Analytics overview page** (<https://aws.amazon.com/kinesis/data-analytics/>**),** the **Amazon Kinesis Data Analytics FAQs page** (<https://aws.amazon.com/kinesis/data-analytics/faqs/>), the **Getting started with Amazon Kinesis Data Streams page** (<https://aws.amazon.com/kinesis/data-streams/getting-started/>), 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>), and the **AWS Serverless Application Repository page** (<https://serverlessrepo.aws.amazon.com/applications>)

Question: 54

**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 a healthcare company that wishes to create a data lake from its various data stores located throughout their siloed infrastructure. They have several data stores in DynamoDB, Oracle, and flat files on a file share. You have been tasked with loading their customer data from one of their DynamoDB tables into your EMR cluster.

You have written a MapReduce job using Hive for which you wish to use all of your nodes in your EMR cluster to copy your DynamoDB table data into your HDFS. You wish to create your files on HDFS in the CSV format.

Which set of commands will execute this parallel copy for you?

1. INSERT OVERWRITE TABLE hdfs\_customers\_csv  
   SELECT \* FROM ddb\_customers
2. INSERT OVERWRITE DIRECTORY 'hdfs:///user/hadoop/hive-customers'  
   SELECT \* FROM ddb\_customers;
3. CREATE EXTERNAL TABLE hdfs\_customers\_csv(customer\_id, …)  
   LOCATION ‘hdfs://user/hadoop/hive-customers’;  
   INSERT OVERWRITE TABLE hdfs\_customers\_csv  
   SELECT \* FROM ddb\_customers
4. CREATE EXTERNAL TABLE hdfs\_customers\_csv(customer\_id, …)  
   ROW FORMAT DELIMITED   
   FIELDS TERMINATED BY ‘,’  
   LOCATION ‘hdfs://user/hadoop/hive-customers’;  
   INSERT OVERWRITE TABLE hdfs\_customers\_csv  
   SELECT \* FROM ddb\_customers

**Answer:** D

**Explanation:**

Option A is incorrect. This set of commands will direct Hive to copy your ddb\_customers data from your DynamoDB customers table to your HDFS hdfs\_customers\_csv file but with the default delimiter. Also, the correct syntax for the INSERT command is INSERT OVERWRITE DIRECTORY if you are using the default settings.

Option B is incorrect. This option, using the default settings, will create your ddb\_customers data from your DynamoDB customers table to your HDFS hdfs\_customers\_csv file but with the default delimiter, not the comma delimiter needed for a CSV file.

Option C is incorrect. This option is missing the ROW FORMAT DELIMITED and FIELDS TERMINATED BY ‘,’ statements needed to create a CSV formatted file on HDFS.

Option D is correct. This option uses the correct syntax and row delimited statements to properly create a CSV formatted file on your HDFS.

**Reference:**

Please see the Amazon DynamoDB developer guide titled **Copying Data to and from Amazon DynamoDB** (<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/EMRforDynamoDB.CopyingData.html>), the Amazon DynamoDB developer guide titled **Copying Data Between DynamoDB and HDFS** (<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/EMRforDynamoDB.CopyingData.HDFS.html>), and the Amazon EMR release guide titled **Set Up a Hive Table to Run Hive Commands** [**https://docs.aws.amazon.com/emr/latest/ReleaseGuide/EMR\_Interactive\_Hive.html**](https://docs.aws.amazon.com/emr/latest/ReleaseGuide/EMR_Interactive_Hive.html))

Question: 55

**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 architect for a literature publishing firm that publishes literature (novels, non-fiction, poetry, etc.) around the globe in several different languages. Your management team has moved your published format to almost exclusively digital to allow for immediate delivery of their product to their consumers. The platforms used by your customers to read your literature generate many IoT data messages as the customers interact with your literature. This data flows into your data collection system at very high volume levels.

You have been given the requirements that the IoT data must be housed in your corporate data lake and that the data must be highly available. You have also been asked to transform the IoT data and group the data records into batches according to the literature’s published language. Your most important data collection system characteristics are durability of the data and data lake retrieval performance.

You have built a Kinesis Data Stream to collect the IoT data. Which of the following options will meet your requirements in the most cost optimized, durable, and performant manner?

1. Construct multiple Kinesis Data Firehose streams while using custom shards to batch the data. The Kinesis Data Firehose will write the data stream to your data lake.
2. Construct a Kinesis Data Firehose that receives the IoT record data from the Kinesis Data Stream. The Kinesis Data Firehose buffers and converts the data to partitioned ORC files and writes them to your data lake.
3. Write a Lambda function to accept batches of IoT records from the Kinesis Data Stream. The Lambda function converts the data records to CSV and then writes the CSV data to your data lake.
4. Write a Spark streaming job in your EMR cluster that reads your IoT data records from your Kinesis Data Stream. The Spark job converts the data to partitioned ORC files and writes them to your data lake.

**Answer:** B

**Explanation:**

Option A is incorrect. Writing a custom sharding process is not as cost effective as just using Kinesis Data Firehose to receive the IoT data records from your Kinesis Data Stream and batch/convert the IoT data records to partitioned ORC files.

Option B is correct. Kinesis Data Firehose is durable in that it uses multiple availability zones. Kinesis Data Firehose also facilitates the easy conversion of your data to partitioned ORC files for storage in your data lake. The partitioned ORC files make for highly optimized SQL queries, which gives you the best performance when retrieving data from your data lake.

Option C is incorrect. The AWS Lambda infrastructure is not optimized for buffering data. Lambda has runtime duration limitations (15 minutes). Also, the CSV format will not be as performant as partitioned ORC files when retrieving your data from your data lake.

Option D is incorrect. This option will result in a less durable solution than a Kinesis Data Firehose based data collection solution. EMR runs in a single availability zone, whereas Kinesis Data Firehose runs across multiple availability zones.

**Reference:**

Please see the Amazon Kinesis Data Firehose developer guide titled **Resilience in Amazon Kinesis Data Firehose** (<https://docs.aws.amazon.com/firehose/latest/dev/disaster-recovery-resiliency.html>), the AWS Big Data blog titled **Analyze Apache Parquet optimized data using Amazon Kinesis Data Firehose, Amazon Athena, and Amazon Redshift** (<https://aws.amazon.com/blogs/big-data/analyzing-apache-parquet-optimized-data-using-amazon-kinesis-data-firehose-amazon-athena-and-amazon-redshift/>), and the AWS What’s-New article titled **Stream Real-Time Data in Apache Parquet or ORC Format Using Amazon Kinesis Data Firehose** (<https://aws.amazon.com/about-aws/whats-new/2018/05/stream_real_time_data_in_apache_parquet_or_orc_format_using_firehose/>)