**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

**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.

**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. Oozie to schedule and run the Spark jobs
2. Airflow to schedule and run the Spark jobs
3. Step Functions to schedule and run the Spark jobs
4. Lambda 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.

**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/>), and the **Apache Livy overview page** (<https://livy.apache.org/>)