

# Amazon-Web-Services

## Exam Questions DAS-C01

AWS Certified Data Analytics - Specialty



**NEW QUESTION 1**

A company is hosting an enterprise reporting solution with Amazon Redshift. The application provides reporting capabilities to three main groups: an executive group to access financial reports, a data analyst group to run long-running ad-hoc queries, and a data engineering group to run stored procedures and ETL processes. The executive team requires queries to run with optimal performance. The data engineering team expects queries to take minutes. Which Amazon Redshift feature meets the requirements for this task?

- A. Concurrency scaling
- B. Short query acceleration (SQA)
- C. Workload management (WLM)
- D. Materialized views

**Answer: D**

**Explanation:**

Materialized views:

**NEW QUESTION 2**

A data analyst is using AWS Glue to organize, cleanse, validate, and format a 200 GB dataset. The data analyst triggered the job to run with the Standard worker type. After 3 hours, the AWS Glue job status is still RUNNING. Logs from the job run show no error codes. The data analyst wants to improve the job execution time without overprovisioning. Which actions should the data analyst take?

- A. Enable job bookmarks in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the executor-cores job parameter.
- B. Enable job metrics in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the maximum capacity job parameter.
- C. Enable job metrics in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the spark.yarn.executor.memoryOverhead job parameter.
- D. Enable job bookmarks in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the num-executors job parameter.

**Answer: B**

**NEW QUESTION 3**

A medical company has a system with sensor devices that read metrics and send them in real time to an Amazon Kinesis data stream. The Kinesis data stream has multiple shards. The company needs to calculate the average value of a numeric metric every second and set an alarm for whenever the value is above one threshold or below another threshold. The alarm must be sent to Amazon Simple Notification Service (Amazon SNS) in less than 30 seconds. Which architecture meets these requirements?

- A. Use an Amazon Kinesis Data Firehose delivery stream to read the data from the Kinesis data stream with an AWS Lambda transformation function that calculates the average per second and sends the alarm to Amazon SNS.
- B. Use an AWS Lambda function to read from the Kinesis data stream to calculate the average per second and sent the alarm to Amazon SNS.
- C. Use an Amazon Kinesis Data Firehose deliver stream to read the data from the Kinesis data stream and store it on Amazon S3. Have Amazon S3 trigger an AWS Lambda function that calculates the average per second and sends the alarm to Amazon SNS.
- D. Use an Amazon Kinesis Data Analytics application to read from the Kinesis data stream and calculatethe average per secon
- E. Send the results to an AWS Lambda function that sends the alarm to Amazon SNS.

**Answer: D**

**NEW QUESTION 4**

A power utility company is deploying thousands of smart meters to obtain real-time updates about power consumption. The company is using Amazon Kinesis Data Streams to collect the data streams from smart meters. The consumer application uses the Kinesis Client Library (KCL) to retrieve the stream data. The company has only one consumer application. The company observes an average of 1 second of latency from the moment that a record is written to the stream until the record is read by a consumer application. The company must reduce this latency to 500 milliseconds. Which solution meets these requirements?

- A. Use enhanced fan-out in Kinesis Data Streams.
- B. Increase the number of shards for the Kinesis data stream.
- C. Reduce the propagation delay by overriding the KCL default settings.
- D. Develop consumers by using Amazon Kinesis Data Firehose.

**Answer: C**

**Explanation:**

The KCL defaults are set to follow the best practice of polling every 1 second. This default results in average propagation delays that are typically below 1 second.

**NEW QUESTION 5**

A company needs to store objects containing log data in JSON format. The objects are generated by eight applications running in AWS. Six of the applications generate a total of 500 KiB of data per second, and two of the applications can generate up to 2 MiB of data per second. A data engineer wants to implement a scalable solution to capture and store usage data in an Amazon S3 bucket. The usage data objects need to be reformatted, converted to .csv format, and then compressed before they are stored in Amazon S3. The company requires the solution to include the least custom code possible and has authorized the data engineer to request a service quota increase if needed. Which solution meets these requirements?

- A. Configure an Amazon Kinesis Data Firehose delivery stream for each applicatio

- B. Write AWS Lambda functions to read log data objects from the stream for each applicatio
- C. Have the function perform reformatting and .csv conversio
- D. Enable compression on all the delivery streams.
- E. Configure an Amazon Kinesis data stream with one shard per applicatio
- F. Write an AWS Lambda function to read usage data objects from the shard
- G. Have the function perform .csv conversion, reformatting, and compression of the dat
- H. Have the function store the output in Amazon S3.
- I. Configure an Amazon Kinesis data stream for each applicatio
- J. Write an AWS Lambda function to read usage data objects from the stream for each applicatio
- K. Have the function perform .csv conversion, reformatting, and compression of the dat
- L. Have the function store the output in Amazon S3.
- M. Store usage data objects in an Amazon DynamoDB tabl
- N. Configure a DynamoDB stream to copy the objects to an S3 bucke
- O. Configure an AWS Lambda function to be triggered when objects are written to the S3 bucke
- P. Have the function convert the objects into .csv format.

**Answer:** A

#### NEW QUESTION 6

A company's marketing team has asked for help in identifying a high performing long-term storage service for their data based on the following requirements:

- The data size is approximately 32 TB uncompressed.
- There is a low volume of single-row inserts each day.
- There is a high volume of aggregation queries each day.
- Multiple complex joins are performed.
- The queries typically involve a small subset of the columns in a table. Which storage service will provide the MOST performant solution?

- A. Amazon Aurora MySQL
- B. Amazon Redshift
- C. Amazon Neptune
- D. Amazon Elasticsearch

**Answer:** B

#### NEW QUESTION 7

A company has a data lake on AWS that ingests sources of data from multiple business units and uses Amazon Athena for queries. The storage layer is Amazon S3 using the AWS Glue Data Catalog. The company wants to make the data available to its data scientists and business analysts. However, the company first needs to manage data access for Athena based on user roles and responsibilities. What should the company do to apply these access controls with the LEAST operational overhead?

- A. Define security policy-based rules for the users and applications by role in AWS Lake Formation.
- B. Define security policy-based rules for the users and applications by role in AWS Identity and Access Management (IAM).
- C. Define security policy-based rules for the tables and columns by role in AWS Glue.
- D. Define security policy-based rules for the tables and columns by role in AWS Identity and Access Management (IAM).

**Answer:** D

#### NEW QUESTION 8

A company has 1 million scanned documents stored as image files in Amazon S3. The documents contain typewritten application forms with information including the applicant first name, applicant last name, application date, application type, and application text. The company has developed a machine learning algorithm to extract the metadata values from the scanned documents. The company wants to allow internal data analysts to analyze and find applications using the applicant name, application date, or application text. The original images should also be downloadable. Cost control is secondary to query performance. Which solution organizes the images and metadata to drive insights while meeting the requirements?

- A. For each image, use object tags to add the metadat
- B. Use Amazon S3 Select to retrieve the files based on the applicant name and application date.
- C. Index the metadata and the Amazon S3 location of the image file in Amazon Elasticsearch Service. Allow the data analysts to use Kibana to submit queries to the Elasticsearch cluster.
- D. Store the metadata and the Amazon S3 location of the image file in an Amazon Redshift tabl
- E. Allow the data analysts to run ad-hoc queries on the table.
- F. Store the metadata and the Amazon S3 location of the image files in an Apache Parquet file in Amazon S3, and define a table in the AWS Glue Data Catalo
- G. Allow data analysts to use Amazon Athena to submit custom queries.

**Answer:** B

**Explanation:**

<https://aws.amazon.com/blogs/machine-learning/automatically-extract-text-and-structured-data-from-documents>

**NEW QUESTION 9**

A retail company's data analytics team recently created multiple product sales analysis dashboards for the average selling price per product using Amazon QuickSight. The dashboards were created from .csv files uploaded to Amazon S3. The team is now planning to share the dashboards with the respective external product owners by creating individual users in Amazon QuickSight. For compliance and governance reasons, restricting access is a key requirement. The product owners should view only their respective product analysis in the dashboard reports.

Which approach should the data analytics team take to allow product owners to view only their products in the dashboard?

- A. Separate the data by product and use S3 bucket policies for authorization.
- B. Separate the data by product and use IAM policies for authorization.
- C. Create a manifest file with row-level security.
- D. Create dataset rules with row-level security.

**Answer:** D

**Explanation:**

<https://docs.aws.amazon.com/quicksight/latest/user/restrict-access-to-a-data-set-using-row-level-security.html>

**NEW QUESTION 10**

A data analyst is using Amazon QuickSight for data visualization across multiple datasets generated by applications. Each application stores files within a separate Amazon S3 bucket. AWS Glue Data Catalog is used as a central catalog across all application data in Amazon S3. A new application stores its data within a separate S3 bucket. After updating the catalog to include the new application data source, the data analyst created a new Amazon QuickSight data source from an Amazon Athena table, but the import into SPICE failed.

How should the data analyst resolve the issue?

- A. Edit the permissions for the AWS Glue Data Catalog from within the Amazon QuickSight console.
- B. Edit the permissions for the new S3 bucket from within the Amazon QuickSight console.
- C. Edit the permissions for the AWS Glue Data Catalog from within the AWS Glue console.
- D. Edit the permissions for the new S3 bucket from within the S3 console.

**Answer:** B

**NEW QUESTION 10**

An ecommerce company stores customer purchase data in Amazon RDS. The company wants a solution to store and analyze historical data. The most recent 6 months of data will be queried frequently for analytics workloads. This data is several terabytes large. Once a month, historical data for the last 5 years must be accessible and will be joined with the more recent data. The company wants to optimize performance and cost.

Which storage solution will meet these requirements?

- A. Create a read replica of the RDS database to store the most recent 6 months of data.
- B. Copy the historical data into Amazon S3. Create an AWS Glue Data Catalog of the data in Amazon S3 and Amazon RDS.
- C. Run historical queries using Amazon Athena.
- D. Use an ETL tool to incrementally load the most recent 6 months of data into an Amazon Redshift cluster.
- E. Run more frequent queries against this cluster.
- F. Create a read replica of the RDS database to run queries on the historical data.
- G. Incrementally copy data from Amazon RDS to Amazon S3. Create an AWS Glue Data Catalog of the data in Amazon S3. Use Amazon Athena to query the data.
- H. Incrementally copy data from Amazon RDS to Amazon S3. Load and store the most recent 6 months of data in Amazon Redshift.
- I. Configure an Amazon Redshift Spectrum table to connect to all historical data.

**Answer:** D

**NEW QUESTION 15**

An insurance company has raw data in JSON format that is sent without a predefined schedule through an Amazon Kinesis Data Firehose delivery stream to an Amazon S3 bucket. An AWS Glue crawler is scheduled to run every 8 hours to update the schema in the data catalog of the tables stored in the S3 bucket. Data analysts analyze the data using Apache Spark SQL on Amazon EMR set up with AWS Glue Data Catalog as the metastore. Data analysts say that, occasionally, the data they receive is stale. A data engineer needs to provide access to the most up-to-date data.

Which solution meets these requirements?

- A. Create an external schema based on the AWS Glue Data Catalog on the existing Amazon Redshift cluster to query new data in Amazon S3 with Amazon Redshift Spectrum.
- B. Use Amazon CloudWatch Events with the rate (1 hour) expression to execute the AWS Glue crawler every hour.
- C. Using the AWS CLI, modify the execution schedule of the AWS Glue crawler from 8 hours to 1 minute.
- D. Run the AWS Glue crawler from an AWS Lambda function triggered by an S3:ObjectCreated:\* event notification on the S3 bucket.

**Answer:** D

**Explanation:**

<https://docs.aws.amazon.com/AmazonS3/latest/dev/NotificationHowTo.html> "you can use a wildcard (for example, s3:ObjectCreated:\*) to request notification when an object is created regardless of the API used" "AWS Lambda can run custom code in response to Amazon S3 bucket events. You upload your custom code to AWS Lambda and create what is called a Lambda function. When Amazon S3 detects an event of a specific type (for example, an object created event), it can publish the event to AWS Lambda and invoke your function in Lambda. In response, AWS Lambda runs your function."

**NEW QUESTION 19**

A technology company is creating a dashboard that will visualize and analyze time-sensitive data. The data will come in through Amazon Kinesis Data Firehose with the buffer interval set to 60 seconds. The dashboard must support near-real-time data.

Which visualization solution will meet these requirements?



- A. Select Amazon Elasticsearch Service (Amazon ES) as the endpoint for Kinesis Data Firehose
- B. Set up a Kibana dashboard using the data in Amazon ES with the desired analyses and visualizations.
- C. Select Amazon S3 as the endpoint for Kinesis Data Firehose
- D. Read data into an Amazon SageMaker Jupyter notebook and carry out the desired analyses and visualizations.
- E. Select Amazon Redshift as the endpoint for Kinesis Data Firehose
- F. Connect Amazon QuickSight with SPICE to Amazon Redshift to create the desired analyses and visualizations.
- G. Select Amazon S3 as the endpoint for Kinesis Data Firehose
- H. Use AWS Glue to catalog the data and Amazon Athena to query it
- I. Connect Amazon QuickSight with SPICE to Athena to create the desired analyses and visualizations.

**Answer:** A

#### NEW QUESTION 23

An IoT company wants to release a new device that will collect data to track sleep overnight on an intelligent mattress. Sensors will send data that will be uploaded to an Amazon S3 bucket. About 2 MB of data is generated each night for each bed. Data must be processed and summarized for each user, and the results need to be available as soon as possible. Part of the process consists of time windowing and other functions. Based on tests with a Python script, every run will require about 1 GB of memory and will complete within a couple of minutes. Which solution will run the script in the MOST cost-effective way?

- A. AWS Lambda with a Python script
- B. AWS Glue with a Scala job
- C. Amazon EMR with an Apache Spark script
- D. AWS Glue with a PySpark job

**Answer:** A

#### NEW QUESTION 27

An online retail company uses Amazon Redshift to store historical sales transactions. The company is required to encrypt data at rest in the clusters to comply with the Payment Card Industry Data Security Standard (PCI DSS). A corporate governance policy mandates management of encryption keys using an on-premises hardware security module (HSM). Which solution meets these requirements?

- A. Create and manage encryption keys using AWS CloudHSM Classic
- B. Launch an Amazon Redshift cluster in a VPC with the option to use CloudHSM Classic for key management.
- C. Create a VPC and establish a VPN connection between the VPC and the on-premises network
- D. Create an HSM connection and client certificate for the on-premises HSM
- E. Launch a cluster in the VPC with the option to use the on-premises HSM to store keys.
- F. Create an HSM connection and client certificate for the on-premises HSM
- G. Enable HSM encryption on the existing unencrypted cluster by modifying the cluster
- H. Connect to the VPC where the Amazon Redshift cluster resides from the on-premises network using a VPN.
- I. Create a replica of the on-premises HSM in AWS CloudHSM
- J. Launch a cluster in a VPC with the option to use CloudHSM to store keys.

**Answer:** B

#### NEW QUESTION 28

A financial company uses Apache Hive on Amazon EMR for ad-hoc queries. Users are complaining of sluggish performance. A data analyst notes the following:  
Approximately 90% of queries are submitted 1 hour after the market opens.  
Hadoop Distributed File System (HDFS) utilization never exceeds 10%.  
Which solution would help address the performance issues?

- A. Create instance fleet configurations for core and task nodes
- B. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch CapacityRemainingGB metric
- C. Create an automatic scaling policy to scale in the instance fleet based on the CloudWatch CapacityRemainingGB metric.
- D. Create instance fleet configurations for core and task nodes
- E. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch YARNMemoryAvailablePercentage metric
- F. Create an automatic scaling policy to scale in the instance fleet based on the CloudWatch YARNMemoryAvailablePercentage metric.
- G. Create instance group configurations for core and task nodes
- H. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch CapacityRemainingGB metric
- I. Create an automatic scaling policy to scale in the instance groups based on the CloudWatch CapacityRemainingGB metric.
- J. Create instance group configurations for core and task nodes
- K. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch YARNMemoryAvailablePercentage metric
- L. Create an automatic scaling policy to scale in the instance groups based on the CloudWatch YARNMemoryAvailablePercentage metric.

**Answer:** D

#### Explanation:

<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html>

#### NEW QUESTION 33

A data engineering team within a shared workspace company wants to build a centralized logging system for all weblogs generated by the space reservation system. The company has a fleet of Amazon EC2 instances that process requests for shared space reservations on its website. The data engineering team wants to ingest all weblogs into a service that will provide a near-real-time search engine. The team does not want to manage the maintenance and operation of the logging system. Which solution allows the data engineering team to efficiently set up the web logging system within AWS?

- A. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis data stream to CloudWatch Logs
- B. Choose Amazon Elasticsearch Service as the end destination of the weblogs.

- C. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis Data Firehose delivery stream to CloudWatc
- D. Choose Amazon Elasticsearch Service as the end destination of the weblogs.
- E. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis data stream to CloudWatc
- F. Configure Splunk as the end destination of the weblogs.
- G. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis Firehose delivery stream to CloudWatc
- H. Configure Amazon DynamoDB as the end destinationof the weblog

**Answer:** B

**Explanation:**

[https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/CWL\\_ES\\_Stream.html](https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/CWL_ES_Stream.html)

#### NEW QUESTION 37

A real estate company has a mission-critical application using Apache HBase in Amazon EMR. Amazon EMR is configured with a single master node. The company has over 5 TB of data stored on an Hadoop Distributed File System (HDFS). The company wants a cost-effective solution to make its HBase data highly available.

Which architectural pattern meets company's requirements?

- A. Use Spot Instances for core and task nodes and a Reserved Instance for the EMR master node.Configurethe EMR cluster with multiple master node
- B. Schedule automated snapshots using AmazonEventBridge.
- C. Store the data on an EMR File System (EMRFS) instead of HDF
- D. Enable EMRFS consistent view.Create an EMR HBase cluster with multiple master node
- E. Point the HBase root directory to an Amazon S3 bucket.
- F. Store the data on an EMR File System (EMRFS) instead of HDFS and enable EMRFS consistent view.Run two separate EMR clusters in two different Availability Zone
- G. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.
- H. Store the data on an EMR File System (EMRFS) instead of HDFS and enable EMRFS consistent view.Create a primary EMR HBase cluster with multiple master node
- I. Create a secondary EMR HBase read- replica cluster in a separate Availability Zon
- J. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.

**Answer:** D

#### NEW QUESTION 39

A telecommunications company is looking for an anomaly-detection solution to identify fraudulent calls. The company currently uses Amazon Kinesis to stream voice call records in a JSON format from its on-premises database to Amazon S3. The existing dataset contains voice call records with 200 columns. To detect fraudulent calls, the solution would need to look at 5 of these columns only.

The company is interested in a cost-effective solution using AWS that requires minimal effort and experience in anomaly-detection algorithms.

Which solution meets these requirements?

- A. Use an AWS Glue job to transform the data from JSON to Apache Parquet
- B. Use AWS Glue crawlers to discover the schema and build the AWS Glue Data Catalog
- C. Use Amazon Athena to create a table with a subset of column
- D. Use Amazon QuickSight to visualize the data and then use Amazon QuickSight machine learning-powered anomaly detection.
- E. Use Kinesis Data Firehose to detect anomalies on a data stream from Kinesis by running SQL queries, which compute an anomaly score for all calls and store the output in Amazon RD
- F. Use Amazon Athena to build a dataset and Amazon QuickSight to visualize the results.
- G. Use an AWS Glue job to transform the data from JSON to Apache Parquet
- H. Use AWS Glue crawlers to discover the schema and build the AWS Glue Data Catalog
- I. Use Amazon SageMaker to build an anomaly detection model that can detect fraudulent calls by ingesting data from Amazon S3.
- J. Use Kinesis Data Analytics to detect anomalies on a data stream from Kinesis by running SQL queries, which compute an anomaly score for all call
- K. Connect Amazon QuickSight to Kinesis Data Analytics to visualize the anomaly scores.

**Answer:** A

#### NEW QUESTION 42

A company stores its sales and marketing data that includes personally identifiable information (PII) in Amazon S3. The company allows its analysts to launch their own Amazon EMR cluster and run analytics reports with the data. To meet compliance requirements, the company must ensure the data is not publicly accessible throughout this process. A data engineer has secured Amazon S3 but must ensure the individual EMR clusters created by the analysts are not exposed to the public internet.

Which solution should the data engineer to meet this compliance requirement with LEAST amount of effort?

- A. Create an EMR security configuration and ensure the security configuration is associated with the EMR clusters when they are created.
- B. Check the security group of the EMR clusters regularly to ensure it does not allow inbound traffic from IPv4 0.0.0.0/0 or IPv6 ::/0.
- C. Enable the block public access setting for Amazon EMR at the account level before any EMR cluster is created.
- D. Use AWS WAF to block public internet access to the EMR clusters across the board.

**Answer:** C

**Explanation:**

<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-block-public-access.html>

#### NEW QUESTION 43

A company developed a new elections reporting website that uses Amazon Kinesis Data Firehose to deliver full logs from AWS WAF to an Amazon S3 bucket.

The company is now seeking a low-cost option to perform this infrequent data analysis with visualizations of logs in a way that requires minimal development effort.

Which solution meets these requirements?

- A. Use an AWS Glue crawler to create and update a table in the Glue data catalog from the log
- B. Use Athena to perform ad-hoc analyses and use Amazon QuickSight to develop data visualizations.

- C. Create a second Kinesis Data Firehose delivery stream to deliver the log files to Amazon Elasticsearch Service (Amazon ES). Use Amazon ES to perform text-based searches of the logs for ad-hoc analyses and use Kibana for data visualizations.
- D. Create an AWS Lambda function to convert the logs into .csv format.
- E. Then add the function to the Kinesis Data Firehose transformation configuration.
- F. Use Amazon Redshift to perform ad-hoc analyses of the logs using SQL queries and use Amazon QuickSight to develop data visualizations.
- G. Create an Amazon EMR cluster and use Amazon S3 as the data source.
- H. Create an Apache Spark job to perform ad-hoc analyses and use Amazon QuickSight to develop data visualizations.

**Answer:** A

**Explanation:**

<https://aws.amazon.com/blogs/big-data/analyzing-aws-waf-logs-with-amazon-es-amazon-athena-and-amazon-qu>

**NEW QUESTION 47**

A company has a data warehouse in Amazon Redshift that is approximately 500 TB in size. New data is imported every few hours and read-only queries are run throughout the day and evening. There is a particularly heavy load with no writes for several hours each morning on business days. During those hours, some queries are queued and take a long time to execute. The company needs to optimize query execution and avoid any downtime. What is the MOST cost-effective solution?

- A. Enable concurrency scaling in the workload management (WLM) queue.
- B. Add more nodes using the AWS Management Console during peak hour.
- C. Set the distribution style to ALL.
- D. Use elastic resize to quickly add nodes during peak time.
- E. Remove the nodes when they are not needed.
- F. Use a snapshot, restore, and resize operation.
- G. Switch to the new target cluster.

**Answer:** A

**Explanation:**

<https://docs.aws.amazon.com/redshift/latest/dg/cm-c-implementing-workload-management.html>

**NEW QUESTION 51**

A data analyst is designing an Amazon QuickSight dashboard using centralized sales data that resides in Amazon Redshift. The dashboard must be restricted so that a salesperson in Sydney, Australia, can see only the Australia view and that a salesperson in New York can see only United States (US) data. What should the data analyst do to ensure the appropriate data security is in place?

- A. Place the data sources for Australia and the US into separate SPICE capacity pools.
- B. Set up an Amazon Redshift VPC security group for Australia and the US.
- C. Deploy QuickSight Enterprise edition to implement row-level security (RLS) to the sales table.
- D. Deploy QuickSight Enterprise edition and set up different VPC security groups for Australia and the US.

**Answer:** D

**NEW QUESTION 56**

A company wants to collect and process events data from different departments in near-real time. Before storing the data in Amazon S3, the company needs to clean the data by standardizing the format of the address and timestamp columns. The data varies in size based on the overall load at each particular point in time. A single data record can be 100 KB-10 MB. How should a data analytics specialist design the solution for data ingestion?

- A. Use Amazon Kinesis Data Stream.
- B. Configure a stream for the raw data.
- C. Use a Kinesis Agent to write data to the stream.
- D. Create an Amazon Kinesis Data Analytics application that reads data from the raw stream, cleanses it, and stores the output to Amazon S3.
- E. Use Amazon Kinesis Data Firehose.
- F. Configure a Firehose delivery stream with a preprocessing AWS Lambda function for data cleansing.
- G. Use a Kinesis Agent to write data to the delivery stream.
- H. Configure Kinesis Data Firehose to deliver the data to Amazon S3.
- I. Use Amazon Managed Streaming for Apache Kafka.
- J. Configure a topic for the raw data.
- K. Use a Kafka producer to write data to the topic.
- L. Create an application on Amazon EC2 that reads data from the topic by using the Apache Kafka consumer API, cleanses the data, and writes to Amazon S3.
- M. Use Amazon Simple Queue Service (Amazon SQS). Configure an AWS Lambda function to read events from the SQS queue and upload the events to Amazon S3.

**Answer:** B

**NEW QUESTION 59**

A company launched a service that produces millions of messages every day and uses Amazon Kinesis Data Streams as the streaming service. The company uses the Kinesis SDK to write data to Kinesis Data Streams. A few months after launch, a data analyst found that write performance is significantly reduced. The data analyst investigated the metrics and determined that Kinesis is throttling the write requests. The data analyst wants to address this issue without significant changes to the architecture. Which actions should the data analyst take to resolve this issue? (Choose two.)

- A. Increase the Kinesis Data Streams retention period to reduce throttling.
- B. Replace the Kinesis API-based data ingestion mechanism with Kinesis Agent.
- C. Increase the number of shards in the stream using the UpdateShardCount API.
- D. Choose partition keys in a way that results in a uniform record distribution across shards.
- E. Customize the application code to include retry logic to improve performance.



**Answer:** CD

**Explanation:**

<https://aws.amazon.com/blogs/big-data/under-the-hood-scaling-your-kinesis-data-streams/>

**NEW QUESTION 62**

A global pharmaceutical company receives test results for new drugs from various testing facilities worldwide. The results are sent in millions of 1 KB-sized JSON objects to an Amazon S3 bucket owned by the company. The data engineering team needs to process those files, convert them into Apache Parquet format, and load them into Amazon Redshift for data analysts to perform dashboard reporting. The engineering team uses AWS Glue to process the objects, AWS Step Functions for process orchestration, and Amazon CloudWatch for job scheduling.

More testing facilities were recently added, and the time to process files is increasing. What will MOST efficiently decrease the data processing time?

- A. Use AWS Lambda to group the small files into larger file
- B. Write the files back to Amazon S3. Process the files using AWS Glue and load them into Amazon Redshift tables.
- C. Use the AWS Glue dynamic frame file grouping option while ingesting the raw input file
- D. Process the files and load them into Amazon Redshift tables.
- E. Use the Amazon Redshift COPY command to move the files from Amazon S3 into Amazon Redshift tables directly
- F. Process the files in Amazon Redshift.
- G. Use Amazon EMR instead of AWS Glue to group the small input file
- H. Process the files in Amazon EMR and load them into Amazon Redshift tables.

**Answer:** A

**NEW QUESTION 63**

A company operates toll services for highways across the country and collects data that is used to understand usage patterns. Analysts have requested the ability to run traffic reports in near-real time. The company is interested in building an ingestion pipeline that loads all the data into an Amazon Redshift cluster and alerts operations personnel when toll traffic for a particular toll station does not meet a specified threshold. Station data and the corresponding threshold values are stored in Amazon S3.

Which approach is the MOST efficient way to meet these requirements?

- A. Use Amazon Kinesis Data Firehose to collect data and deliver it to Amazon Redshift and Amazon Kinesis Data Analytics simultaneously
- B. Create a reference data source in Kinesis Data Analytics to temporarily store the threshold values from Amazon S3 and compare the count of vehicles for a particular toll station against its corresponding threshold value
- C. Use AWS Lambda to publish an Amazon Simple Notification Service (Amazon SNS) notification if the threshold is not met.
- D. Use Amazon Kinesis Data Streams to collect all the data from toll station
- E. Create a stream in Kinesis Data Streams to temporarily store the threshold values from Amazon S3. Send both streams to Amazon Kinesis Data Analytics to compare the count of vehicles for a particular toll station against its corresponding threshold value
- F. Use AWS Lambda to publish an Amazon Simple Notification Service (Amazon SNS) notification if the threshold is not met
- G. Connect Amazon Kinesis Data Firehose to Kinesis Data Streams to deliver the data to Amazon Redshift.
- H. Use Amazon Kinesis Data Firehose to collect data and deliver it to Amazon Redshift
- I. Then, automatically trigger an AWS Lambda function that queries the data in Amazon Redshift, compares the count of vehicles for a particular toll station against its corresponding threshold values read from Amazon S3, and publishes an Amazon Simple Notification Service (Amazon SNS) notification if the threshold is not met.
- J. Use Amazon Kinesis Data Firehose to collect data and deliver it to Amazon Redshift and Amazon Kinesis Data Analytics simultaneously
- K. Use Kinesis Data Analytics to compare the count of vehicles against the threshold value for the station stored in a table as an in-application stream based on information stored in Amazon S3. Configure an AWS Lambda function as an output for the application that will publish an Amazon Simple Queue Service (Amazon SQS) notification to alert operations personnel if the threshold is not met.

**Answer:** D

**NEW QUESTION 67**

An operations team notices that a few AWS Glue jobs for a given ETL application are failing. The AWS Glue jobs read a large number of small JSON files from an Amazon S3 bucket and write the data to a different S3 bucket in Apache Parquet format with no major transformations. Upon initial investigation, a data engineer notices the following error message in the History tab on the AWS Glue console: "Command Failed with Exit Code 1."

Upon further investigation, the data engineer notices that the driver memory profile of the failed jobs crosses the safe threshold of 50% usage quickly and reaches 90–95% soon after. The average memory usage across all executors continues to be less than 4%.

The data engineer also notices the following error while examining the related Amazon CloudWatch Logs. What should the data engineer do to solve the failure in the MOST cost-effective way?

- A. Change the worker type from Standard to G.2X.
- B. Modify the AWS Glue ETL code to use the 'groupFiles': 'inPartition' feature.
- C. Increase the fetch size setting by using AWS Glue dynamics frame.
- D. Modify maximum capacity to increase the total maximum data processing units (DPUs) used.

**Answer:** B

**Explanation:**

<https://docs.aws.amazon.com/glue/latest/dg/monitor-profile-debug-oom-abnormalities.html#monitor-debug-oom>

**NEW QUESTION 71**

A company is building a service to monitor fleets of vehicles. The company collects IoT data from a device in each vehicle and loads the data into Amazon Redshift in near-real time. Fleet owners upload .csv files containing vehicle reference data into Amazon S3 at different times throughout the day. A nightly process loads the vehicle reference data from Amazon S3 into Amazon Redshift. The company joins the IoT data from the device and the vehicle reference data to power reporting and dashboards. Fleet owners are frustrated by waiting a day for the dashboards to update.

Which solution would provide the SHORTEST delay between uploading reference data to Amazon S3 and the change showing up in the owners' dashboards?

- A. Use S3 event notifications to trigger an AWS Lambda function to copy the vehicle reference data into Amazon Redshift immediately when the reference data is uploaded to Amazon S3.
- B. Create and schedule an AWS Glue Spark job to run every 5 minute



- C. The job inserts reference data into Amazon Redshift.
- D. Send reference data to Amazon Kinesis Data Stream
- E. Configure the Kinesis data stream to directly load the reference data into Amazon Redshift in real time.
- F. Send the reference data to an Amazon Kinesis Data Firehose delivery stream
- G. Configure Kinesis with a buffer interval of 60 seconds and to directly load the data into Amazon Redshift.

**Answer:** A

#### NEW QUESTION 74

A data analytics specialist is building an automated ETL ingestion pipeline using AWS Glue to ingest compressed files that have been uploaded to an Amazon S3 bucket. The ingestion pipeline should support incremental data processing.

Which AWS Glue feature should the data analytics specialist use to meet this requirement?

- A. Workflows
- B. Triggers
- C. Job bookmarks
- D. Classifiers

**Answer:** C

#### NEW QUESTION 75

A company uses Amazon Redshift for its data warehousing needs. ETL jobs run every night to load data, apply business rules, and create aggregate tables for reporting. The company's data analysis, data science, and business intelligence teams use the data warehouse during regular business hours. The workload management is set to auto, and separate queues exist for each team with the priority set to NORMAL.

Recently, a sudden spike of read queries from the data analysis team has occurred at least twice daily, and queries wait in line for cluster resources. The company needs a solution that enables the data analysis team to avoid query queuing without impacting latency and the query times of other teams.

Which solution meets these requirements?

- A. Increase the query priority to HIGHEST for the data analysis queue.
- B. Configure the data analysis queue to enable concurrency scaling.
- C. Create a query monitoring rule to add more cluster capacity for the data analysis queue when queries are waiting for resources.
- D. Use workload management query queue hopping to route the query to the next matching queue.

**Answer:** D

#### NEW QUESTION 78

A manufacturing company wants to create an operational analytics dashboard to visualize metrics from equipment in near-real time. The company uses Amazon Kinesis Data Streams to stream the data to other applications. The dashboard must automatically refresh every 5 seconds. A data analytics specialist must design a solution that requires the least possible implementation effort.

Which solution meets these requirements?

- A. Use Amazon Kinesis Data Firehose to store the data in Amazon S3. Use Amazon QuickSight to build the dashboard.
- B. Use Apache Spark Streaming on Amazon EMR to read the data in near-real time
- C. Develop a custom application for the dashboard by using D3.js.
- D. Use Amazon Kinesis Data Firehose to push the data into an Amazon Elasticsearch Service (Amazon ES) cluster
- E. Visualize the data by using a Kibana dashboard.
- F. Use AWS Glue streaming ETL to store the data in Amazon S3. Use Amazon QuickSight to build the dashboard.

**Answer:** B

#### NEW QUESTION 79

An education provider's learning management system (LMS) is hosted in a 100 TB data lake that is built on Amazon S3. The provider's LMS supports hundreds of schools. The provider wants to build an advanced analytics reporting platform using Amazon Redshift to handle complex queries with optimal performance.

System users will query the most recent 4 months of data 95% of the time while 5% of the queries will leverage data from the previous 12 months.

Which solution meets these requirements in the MOST cost-effective way?

- A. Store the most recent 4 months of data in the Amazon Redshift cluster
- B. Use Amazon Redshift Spectrum to query data in the data lake
- C. Use S3 lifecycle management rules to store data from the previous 12 months in Amazon S3 Glacier storage.
- D. Leverage DS2 nodes for the Amazon Redshift cluster
- E. Migrate all data from Amazon S3 to Amazon Redshift
- F. Decommission the data lake.
- G. Store the most recent 4 months of data in the Amazon Redshift cluster
- H. Use Amazon Redshift Spectrum to query data in the data lake
- I. Ensure the S3 Standard storage class is in use with objects in the data lake.
- J. Store the most recent 4 months of data in the Amazon Redshift cluster
- K. Use Amazon Redshift federated queries to join cluster data with the data lake to reduce cost
- L. Ensure the S3 Standard storage class is in use with objects in the data lake.

**Answer:** C

#### NEW QUESTION 84

An ecommerce company is migrating its business intelligence environment from on premises to the AWS Cloud. The company will use Amazon Redshift in a public subnet and Amazon QuickSight. The tables already are loaded into Amazon Redshift and can be accessed by a SQL tool.

The company starts QuickSight for the first time. During the creation of the data source, a data analytics specialist enters all the information and tries to validate the connection. An error with the following message occurs: "Creating a connection to your data source timed out."

How should the data analytics specialist resolve this error?

- A. Grant the SELECT permission on Amazon Redshift tables.
- B. Add the QuickSight IP address range into the Amazon Redshift security group.
- C. Create an IAM role for QuickSight to access Amazon Redshift.
- D. Use a QuickSight admin user for creating the dataset.

**Answer:** A

**Explanation:**

Connection to the database times out

Your client connection to the database appears to hang or time out when running long queries, such as a COPY command. In this case, you might observe that the Amazon Redshift console displays that the query has completed, but the client tool itself still appears to be running the query. The results of the query might be missing or incomplete depending on when the connection stopped.

**NEW QUESTION 85**

A retail company wants to use Amazon QuickSight to generate dashboards for web and in-store sales. A group of 50 business intelligence professionals will develop and use the dashboards. Once ready, the dashboards will be shared with a group of 1,000 users.

The sales data comes from different stores and is uploaded to Amazon S3 every 24 hours. The data is partitioned by year and month, and is stored in Apache Parquet format. The company is using the AWS Glue Data Catalog as its main data catalog and Amazon Athena for querying. The total size of the uncompressed data that the dashboards query from at any point is 200 GB.

Which configuration will provide the MOST cost-effective solution that meets these requirements?

- A. Load the data into an Amazon Redshift cluster by using the COPY command.
- B. Configure 50 author users and 1,000 reader user
- C. Use QuickSight Enterprise edition
- D. Configure an Amazon Redshift data source with a direct query option.
- E. Use QuickSight Standard edition
- F. Configure 50 author users and 1,000 reader user
- G. Configure an Athena data source with a direct query option.
- H. Use QuickSight Enterprise edition
- I. Configure 50 author users and 1,000 reader user
- J. Configure an Athena data source and import the data into SPICE
- K. Automatically refresh every 24 hours.
- L. Use QuickSight Enterprise edition
- M. Configure 1 administrator and 1,000 reader user
- N. Configure an S3 data source and import the data into SPICE
- O. Automatically refresh every 24 hours.

**Answer:** C

**NEW QUESTION 89**

A company uses Amazon Redshift as its data warehouse. A new table has columns that contain sensitive data. The data in the table will eventually be referenced by several existing queries that run many times a day.

A data analyst needs to load 100 billion rows of data into the new table. Before doing so, the data analyst must ensure that only members of the auditing group can read the columns containing sensitive data.

How can the data analyst meet these requirements with the lowest maintenance overhead?

- A. Load all the data into the new table and grant the auditing group permission to read from the table.
- B. Load all the data except for the columns containing sensitive data into a second table.
- C. Grant the appropriate users read-only permissions to the second table.
- D. Load all the data into the new table and grant the auditing group permission to read from the table.
- E. Use the GRANT SQL command to allow read-only access to a subset of columns to the appropriate users.
- F. Load all the data into the new table and grant all users read-only permissions to non-sensitive columns. Attach an IAM policy to the auditing group with explicit ALLOW access to the sensitive data columns.
- G. Load all the data into the new table and grant the auditing group permission to read from the table. Create a view of the new table that contains all the columns, except for those considered sensitive, and grant the appropriate users read-only permissions to the table.

**Answer:** B

**Explanation:**

<https://aws.amazon.com/blogs/big-data/achieve-finer-grained-data-security-with-column-level-access-control-in>

**NEW QUESTION 91**

A company is migrating from an on-premises Apache Hadoop cluster to an Amazon EMR cluster. The cluster runs only during business hours. Due to a company requirement to avoid intraday cluster failures, the EMR cluster must be highly available. When the cluster is terminated at the end of each business day, the data must persist.

Which configurations would enable the EMR cluster to meet these requirements? (Choose three.)

- A. EMR File System (EMRFS) for storage
- B. Hadoop Distributed File System (HDFS) for storage
- C. AWS Glue Data Catalog as the metastore for Apache Hive
- D. MySQL database on the master node as the metastore for Apache Hive
- E. Multiple master nodes in a single Availability Zone
- F. Multiple master nodes in multiple Availability Zones

**Answer:** ACE

**Explanation:**

<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-ha.html> "Note : The cluster can reside only in one Availability Zone or subnet."

**NEW QUESTION 95**

A financial company hosts a data lake in Amazon S3 and a data warehouse on an Amazon Redshift cluster. The company uses Amazon QuickSight to build dashboards and wants to secure access from its on-premises Active Directory to Amazon QuickSight. How should the data be secured?

- A. Use an Active Directory connector and single sign-on (SSO) in a corporate network environment.
- B. Use a VPC endpoint to connect to Amazon S3 from Amazon QuickSight and an IAM role to authenticate Amazon Redshift.
- C. Establish a secure connection by creating an S3 endpoint to connect Amazon QuickSight and a VPC endpoint to connect to Amazon Redshift.
- D. Place Amazon QuickSight and Amazon Redshift in the security group and use an Amazon S3 endpoint to connect Amazon QuickSight to Amazon S3.

**Answer:** A

**Explanation:**

<https://docs.aws.amazon.com/quicksight/latest/user/directory-integration.html>

**NEW QUESTION 97**

A smart home automation company must efficiently ingest and process messages from various connected devices and sensors. The majority of these messages are comprised of a large number of small files. These messages are ingested using Amazon Kinesis Data Streams and sent to Amazon S3 using a Kinesis data stream consumer application. The Amazon S3 message data is then passed through a processing pipeline built on Amazon EMR running scheduled PySpark jobs. The data platform team manages data processing and is concerned about the efficiency and cost of downstream data processing. They want to continue to use PySpark.

Which solution improves the efficiency of the data processing jobs and is well architected?

- A. Send the sensor and devices data directly to a Kinesis Data Firehose delivery stream to send the data to Amazon S3 with Apache Parquet record format conversion enable
- B. Use Amazon EMR running PySpark to process the data in Amazon S3.
- C. Set up an AWS Lambda function with a Python runtime environmen
- D. Process individual Kinesis data stream messages from the connected devices and sensors using Lambda.
- E. Launch an Amazon Redshift cluste
- F. Copy the collected data from Amazon S3 to Amazon Redshift and move the data processing jobs from Amazon EMR to Amazon Redshift.
- G. Set up AWS Glue Python jobs to merge the small data files in Amazon S3 into larger files and transform them to Apache Parquet forma
- H. Migrate the downstream PySpark jobs from Amazon EMR to AWS Glue.

**Answer:** D

**Explanation:**

<https://aws.amazon.com/it/about-aws/whats-new/2020/04/aws-glue-now-supports-serverless-streaming-etl/>

**NEW QUESTION 98**

A bank operates in a regulated environment. The compliance requirements for the country in which the bank operates say that customer data for each state should only be accessible by the bank's employees located in the same state. Bank employees in one state should NOT be able to access data for customers who have provided a home address in a different state.

The bank's marketing team has hired a data analyst to gather insights from customer data for a new campaign being launched in certain states. Currently, data linking each customer account to its home state is stored in a tabular .csv file within a single Amazon S3 folder in a private S3 bucket. The total size of the S3 folder is 2 GB uncompressed. Due to the country's compliance requirements, the marketing team is not able to access this folder.

The data analyst is responsible for ensuring that the marketing team gets one-time access to customer data for their campaign analytics project, while being subject to all the compliance requirements and controls.

Which solution should the data analyst implement to meet the desired requirements with the LEAST amount of setup effort?

- A. Re-arrange data in Amazon S3 to store customer data about each state in a different S3 folder within the same bucke
- B. Set up S3 bucket policies to provide marketing employees with appropriate data access under compliance control
- C. Delete the bucket policies after the project.
- D. Load tabular data from Amazon S3 to an Amazon EMR cluster using s3DistC
- E. Implement a customHadoop-based row-level security solution on the Hadoop Distributed File System (HDFS) to provide marketing employees with appropriate data access under compliance control
- F. Terminate the EMR cluster after the project.
- G. Load tabular data from Amazon S3 to Amazon Redshift with the COPY comman
- H. Use the built-in row- level security feature in Amazon Redshift to provide marketing employees with appropriate data access under compliance control
- I. Delete the Amazon Redshift tables after the project.
- J. Load tabular data from Amazon S3 to Amazon QuickSight Enterprise edition by directly importing it as a data sourc
- K. Use the built-in row-level security feature in Amazon QuickSight to provide marketing employees with appropriate data access under compliance control
- L. Delete Amazon QuickSight data sources after the project is complete.

**Answer:** C

**NEW QUESTION 102**

A streaming application is reading data from Amazon Kinesis Data Streams and immediately writing the data to an Amazon S3 bucket every 10 seconds. The application is reading data from hundreds of shards. The batch interval cannot be changed due to a separate requirement. The data is being accessed by Amazon Athena. Users are seeing degradation in query performance as time progresses.

Which action can help improve query performance?

- A. Merge the files in Amazon S3 to form larger files.
- B. Increase the number of shards in Kinesis Data Streams.
- C. Add more memory and CPU capacity to the streaming application.
- D. Write the files to multiple S3 buckets.

**Answer:** A

**Explanation:**

<https://aws.amazon.com/blogs/big-data/top-10-performance-tuning-tips-for-amazon-athena/>



**NEW QUESTION 103**

An online retail company is migrating its reporting system to AWS. The company's legacy system runs data processing on online transactions using a complex series of nested Apache Hive queries. Transactional data is exported from the online system to the reporting system several times a day. Schemas in the files are stable between updates.

A data analyst wants to quickly migrate the data processing to AWS, so any code changes should be minimized. To keep storage costs low, the data analyst decides to store the data in Amazon S3. It is vital that the data from the reports and associated analytics is completely up to date based on the data in Amazon S3. Which solution meets these requirements?

- A. Create an AWS Glue Data Catalog to manage the Hive metadata
- B. Create an AWS Glue crawler over Amazon S3 that runs when data is refreshed to ensure that data changes are update
- C. Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
- D. Create an AWS Glue Data Catalog to manage the Hive metadata
- E. Create an Amazon EMR cluster with consistent view enable
- F. Run emrfs sync before each analytics step to ensure data changes are update
- G. Create an EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
- H. Create an Amazon Athena table with CREATE TABLE AS SELECT (CTAS) to ensure data is refreshed from underlying queries against the raw dataset
- I. Create an AWS Glue Data Catalog to manage the Hive metadata over the CTAS table
- J. Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
- K. Use an S3 Select query to ensure that the data is properly update
- L. Create an AWS Glue Data Catalog to manage the Hive metadata over the S3 Select table
- M. Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.

**Answer:** A

**NEW QUESTION 104**

A banking company is currently using an Amazon Redshift cluster with dense storage (DS) nodes to store sensitive data. An audit found that the cluster is unencrypted. Compliance requirements state that a database with sensitive data must be encrypted through a hardware security module (HSM) with automated key rotation.

Which combination of steps is required to achieve compliance? (Choose two.)

- A. Set up a trusted connection with HSM using a client and server certificate with automatic key rotation.
- B. Modify the cluster with an HSM encryption option and automatic key rotation.
- C. Create a new HSM-encrypted Amazon Redshift cluster and migrate the data to the new cluster.
- D. Enable HSM with key rotation through the AWS CLI.
- E. Enable Elliptic Curve Diffie-Hellman Ephemeral (ECDHE) encryption in the HSM.

**Answer:** BD

**NEW QUESTION 109**

A company is streaming its high-volume billing data (100 MBps) to Amazon Kinesis Data Streams. A data analyst partitioned the data on account\_id to ensure that all records belonging to an account go to the same Kinesis shard and order is maintained. While building a custom consumer using the Kinesis Java SDK, the data analyst notices that, sometimes, the messages arrive out of order for account\_id. Upon further investigation, the data analyst discovers the messages that are out of order seem to be arriving from different shards for the same account\_id and are seen when a stream resize runs.

What is an explanation for this behavior and what is the solution?

- A. There are multiple shards in a stream and order needs to be maintained in the shard
- B. The data analyst needs to make sure there is only a single shard in the stream and no stream resize runs.
- C. The hash key generation process for the records is not working correctly
- D. The data analyst should generate an explicit hash key on the producer side so the records are directed to the appropriate shard accurately.
- E. The records are not being received by Kinesis Data Streams in order
- F. The producer should use the PutRecords API call instead of the PutRecord API call with the SequenceNumberForOrdering parameter.
- G. The consumer is not processing the parent shard completely before processing the child shards after a stream resize
- H. The data analyst should process the parent shard completely first before processing the child shards.

**Answer:** D

**Explanation:**

<https://docs.aws.amazon.com/streams/latest/dev/kinesis-using-sdk-java-after-resharding.html> the parent shards that remain after the reshard could still contain data that you haven't read yet that was added to the stream before the reshard. If you read data from the child shards before having read all data from the parent shards, you could read data for a particular hash key out of the order given by the data records' sequence numbers. Therefore, assuming that the order of the data is important, you should, after a reshard, always continue to read data from the parent shards until it is exhausted. Only then should you begin reading data from the child shards.

**NEW QUESTION 111**

A marketing company is storing its campaign response data in Amazon S3. A consistent set of sources has generated the data for each campaign. The data is saved into Amazon S3 as .csv files. A business analyst will use Amazon Athena to analyze each campaign's data. The company needs the cost of ongoing data analysis with Athena to be minimized.

Which combination of actions should a data analytics specialist take to meet these requirements? (Choose two.)

- A. Convert the .csv files to Apache Parquet.
- B. Convert the .csv files to Apache Avro.
- C. Partition the data by campaign.
- D. Partition the data by source.
- E. Compress the .csv files.

**Answer:** AC

**Explanation:**

<https://aws.amazon.com/blogs/big-data/top-10-performance-tuning-tips-for-amazon-athena/>



**NEW QUESTION 114**

A company uses Amazon Elasticsearch Service (Amazon ES) to store and analyze its website clickstream data. The company ingests 1 TB of data daily using Amazon Kinesis Data Firehose and stores one day's worth of data in an Amazon ES cluster. The company has very slow query performance on the Amazon ES index and occasionally sees errors from Kinesis Data Firehose when attempting to write to the index. The Amazon ES cluster has 10 nodes running a single index and 3 dedicated master nodes. Each data node has 1.5 TB of Amazon EBS storage attached and the cluster is configured with 1,000 shards. Occasionally, JVMMemoryPressure errors are found in the cluster logs. Which solution will improve the performance of Amazon ES?

- A. Increase the memory of the Amazon ES master nodes.
- B. Decrease the number of Amazon ES data nodes.
- C. Decrease the number of Amazon ES shards for the index.
- D. Increase the number of Amazon ES shards for the index.

**Answer: C**

**Explanation:**

<https://aws.amazon.com/premiumsupport/knowledge-center/high-jvm-memory-pressure-elasticsearch/>

**NEW QUESTION 117**

A company that produces network devices has millions of users. Data is collected from the devices on an hourly basis and stored in an Amazon S3 data lake. The company runs analyses on the last 24 hours of data flow logs for abnormality detection and to troubleshoot and resolve user issues. The company also analyzes historical logs dating back 2 years to discover patterns and look for improvement opportunities. The data flow logs contain many metrics, such as date, timestamp, source IP, and target IP. There are about 10 billion events every day. How should this data be stored for optimal performance?

- A. In Apache ORC partitioned by date and sorted by source IP
- B. In compressed .csv partitioned by date and sorted by source IP
- C. In Apache Parquet partitioned by source IP and sorted by date
- D. In compressed nested JSON partitioned by source IP and sorted by date

**Answer: A**

**NEW QUESTION 120**

A manufacturing company has been collecting IoT sensor data from devices on its factory floor for a year and is storing the data in Amazon Redshift for daily analysis. A data analyst has determined that, at an expected ingestion rate of about 2 TB per day, the cluster will be undersized in less than 4 months. A long-term solution is needed. The data analyst has indicated that most queries only reference the most recent 13 months of data, yet there are also quarterly reports that need to query all the data generated from the past 7 years. The chief technology officer (CTO) is concerned about the costs, administrative effort, and performance of a long-term solution. Which solution should the data analyst use to meet these requirements?

- A. Create a daily job in AWS Glue to UNLOAD records older than 13 months to Amazon S3 and delete those records from Amazon Redshift
- B. Create an external table in Amazon Redshift to point to the S3 location
- C. Use Amazon Redshift Spectrum to join to data that is older than 13 months.
- D. Take a snapshot of the Amazon Redshift cluster
- E. Restore the cluster to a new cluster using dense storage nodes with additional storage capacity.
- F. Execute a CREATE TABLE AS SELECT (CTAS) statement to move records that are older than 13 months to quarterly partitioned data in Amazon Redshift Spectrum backed by Amazon S3.
- G. Unload all the tables in Amazon Redshift to an Amazon S3 bucket using S3 Intelligent-Tiering
- H. Use AWS Glue to crawl the S3 bucket location to create external tables in an AWS Glue Data Catalog
- I. Create an Amazon EMR cluster using Auto Scaling for any daily analytics needs, and use Amazon Athena for the quarterly reports, with both using the same AWS Glue Data Catalog.

**Answer: A**

**NEW QUESTION 123**

A company has an encrypted Amazon Redshift cluster. The company recently enabled Amazon Redshift audit logs and needs to ensure that the audit logs are also encrypted at rest. The logs are retained for 1 year. The auditor queries the logs once a month. What is the MOST cost-effective way to meet these requirements?

- A. Encrypt the Amazon S3 bucket where the logs are stored by using AWS Key Management Service (AWS KMS). Copy the data into the Amazon Redshift cluster from Amazon S3 on a daily basis
- B. Query the data as required.
- C. Disable encryption on the Amazon Redshift cluster, configure audit logging, and encrypt the Amazon Redshift cluster
- D. Use Amazon Redshift Spectrum to query the data as required.
- E. Enable default encryption on the Amazon S3 bucket where the logs are stored by using AES-256 encryption
- F. Copy the data into the Amazon Redshift cluster from Amazon S3 on a daily basis
- G. Query the data as required.
- H. Enable default encryption on the Amazon S3 bucket where the logs are stored by using AES-256 encryption
- I. Use Amazon Redshift Spectrum to query the data as required.

**Answer: A**

**NEW QUESTION 124**

A healthcare company uses AWS data and analytics tools to collect, ingest, and store electronic health record (EHR) data about its patients. The raw EHR data is stored in Amazon S3 in JSON format partitioned by hour, day, and year and is updated every hour. The company wants to maintain the data catalog and metadata in an AWS Glue Data Catalog to be able to access the data using Amazon Athena or Amazon Redshift Spectrum for analytics. When defining tables in the Data Catalog, the company has the following requirements: Choose the catalog table name and do not rely on the catalog table naming algorithm. Keep the table updated with new partitions loaded in the respective S3

bucket prefixes.

Which solution meets these requirements with minimal effort?

- A. Run an AWS Glue crawler that connects to one or more data stores, determines the data structures, and writes tables in the Data Catalog.
- B. Use the AWS Glue console to manually create a table in the Data Catalog and schedule an AWS Lambda function to update the table partitions hourly.
- C. Use the AWS Glue API CreateTable operation to create a table in the Data Catalog.
- D. Create an AWS Glue crawler and specify the table as the source.
- E. Create an Apache Hive catalog in Amazon EMR with the table schema definition in Amazon S3, and update the table partition with a scheduled job.
- F. Migrate the Hive catalog to the Data Catalog.

**Answer: C**

**Explanation:**

Updating Manually Created Data Catalog Tables Using Crawlers: To do this, when you define a crawler, instead of specifying one or more data stores as the source of a crawl, you specify one or more existing Data Catalog tables. The crawler then crawls the data stores specified by the catalog tables. In this case, no new tables are created; instead, your manually created tables are updated.

**NEW QUESTION 126**

A US-based sneaker retail company launched its global website. All the transaction data is stored in Amazon RDS and curated historic transaction data is stored in Amazon Redshift in the us-east-1 Region. The business intelligence (BI) team wants to enhance the user experience by providing a dashboard for sneaker trends. The BI team decides to use Amazon QuickSight to render the website dashboards. During development, a team in Japan provisioned Amazon QuickSight in ap-northeast-1. The team is having difficulty connecting Amazon QuickSight from ap-northeast-1 to Amazon Redshift in us-east-1.

Which solution will solve this issue and meet the requirements?

- A. In the Amazon Redshift console, choose to configure cross-Region snapshots and set the destination Region as ap-northeast-1. Restore the Amazon Redshift Cluster from the snapshot and connect to Amazon QuickSight launched in ap-northeast-1.
- B. Create a VPC endpoint from the Amazon QuickSight VPC to the Amazon Redshift VPC so Amazon QuickSight can access data from Amazon Redshift.
- C. Create an Amazon Redshift endpoint connection string with Region information in the string and use this connection string in Amazon QuickSight to connect to Amazon Redshift.
- D. Create a new security group for Amazon Redshift in us-east-1 with an inbound rule authorizing access from the appropriate IP address range for the Amazon QuickSight servers in ap-northeast-1.

**Answer: B**

**NEW QUESTION 127**

A company has a marketing department and a finance department. The departments are storing data in Amazon S3 in their own AWS accounts in AWS Organizations. Both departments use AWS Lake Formation to catalog and secure their data. The departments have some databases and tables that share common names.

The marketing department needs to securely access some tables from the finance department. Which two steps are required for this process? (Choose two.)

- A. The finance department grants Lake Formation permissions for the tables to the external account for the marketing department.
- B. The finance department creates cross-account IAM permissions to the table for the marketing department role.
- C. The marketing department creates an IAM role that has permissions to the Lake Formation tables.

**Answer: AB**

**Explanation:**

Granting Lake Formation Permissions Creating an IAM role (AWS CLI)

**NEW QUESTION 129**

A media company has been performing analytics on log data generated by its applications. There has been a recent increase in the number of concurrent analytics jobs running, and the overall performance of existing jobs is decreasing as the number of new jobs is increasing. The partitioned data is stored in Amazon S3 One Zone-Infrequent Access (S3 One Zone-IA) and the analytic processing is performed on Amazon EMR clusters using the EMR File System (EMRFS) with consistent view enabled. A data analyst has determined that it is taking longer for the EMR task nodes to list objects in Amazon S3.

Which action would MOST likely increase the performance of accessing log data in Amazon S3?

- A. Use a hash function to create a random string and add that to the beginning of the object prefixes when storing the log data in Amazon S3.
- B. Use a lifecycle policy to change the S3 storage class to S3 Standard for the log data.
- C. Increase the read capacity units (RCUs) for the shared Amazon DynamoDB table.
- D. Redeploy the EMR clusters that are running slowly to a different Availability Zone.

**Answer: C**

**Explanation:**

<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emrfs-metadata.html>

**NEW QUESTION 132**

An online retail company with millions of users around the globe wants to improve its ecommerce analytics capabilities. Currently, clickstream data is uploaded directly to Amazon S3 as compressed files. Several times each day, an application running on Amazon EC2 processes the data and makes search options and reports available for visualization by editors and marketers. The company wants to make website clicks and aggregated data available to editors and marketers in minutes to enable them to connect with users more effectively.

Which options will help meet these requirements in the MOST efficient way? (Choose two.)

- A. Use Amazon Kinesis Data Firehose to upload compressed and batched clickstream records to Amazon Elasticsearch Service.
- B. Upload clickstream records to Amazon S3 as compressed file
- C. Then use AWS Lambda to send data to Amazon Elasticsearch Service from Amazon S3.
- D. Use Amazon Elasticsearch Service deployed on Amazon EC2 to aggregate, filter, and process the data.Refresh content performance dashboards in near-real time.
- E. Use Kibana to aggregate, filter, and visualize the data stored in Amazon Elasticsearch Service

F. Refresh content performance dashboards in near-real time.

G. Upload clickstream records from Amazon S3 to Amazon Kinesis Data Streams and use a Kinesis Data Streams consumer to send records to Amazon Elasticsearch Service.

**Answer:** AD

#### NEW QUESTION 137

A company analyzes historical data and needs to query data that is stored in Amazon S3. New data is generated daily as .csv files that are stored in Amazon S3. The company's analysts are using Amazon Athena to perform SQL queries against a recent subset of the overall data. The amount of data that is ingested into Amazon S3 has increased substantially over time, and the query latency also has increased. Which solutions could the company implement to improve query performance? (Choose two.)

- A. Use MySQL Workbench on an Amazon EC2 instance, and connect to Athena by using a JDBC or ODBC connecto
- B. Run the query from MySQL Workbench instead of Athena directly.
- C. Use Athena to extract the data and store it in Apache Parquet format on a daily basi
- D. Query the extracted data.
- E. Run a daily AWS Glue ETL job to convert the data files to Apache Parquet and to partition the converted file
- F. Create a periodic AWS Glue crawler to automatically crawl the partitioned data on a daily basis.
- G. Run a daily AWS Glue ETL job to compress the data files by using the .gzip forma
- H. Query the compressed data.
- I. Run a daily AWS Glue ETL job to compress the data files by using the .lzo forma
- J. Query the compressed data.

**Answer:** BC

#### NEW QUESTION 141

A data analyst is designing a solution to interactively query datasets with SQL using a JDBC connection. Users will join data stored in Amazon S3 in Apache ORC format with data stored in Amazon Elasticsearch Service (Amazon ES) and Amazon Aurora MySQL. Which solution will provide the MOST up-to-date results?

- A. Use AWS Glue jobs to ETL data from Amazon ES and Aurora MySQL to Amazon S3. Query the data with Amazon Athena.
- B. Use Amazon DMS to stream data from Amazon ES and Aurora MySQL to Amazon Redshif
- C. Query the data with Amazon Redshift.
- D. Query all the datasets in place with Apache Spark SQL running on an AWS Glue developer endpoint.
- E. Query all the datasets in place with Apache Presto running on Amazon EMR.

**Answer:** C

#### NEW QUESTION 143

.....

## Thank You for Trying Our Product

### We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questions and Answers in PDF Format

### DAS-C01 Practice Exam Features:

- \* DAS-C01 Questions and Answers Updated Frequently
- \* DAS-C01 Practice Questions Verified by Expert Senior Certified Staff
- \* DAS-C01 Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- \* DAS-C01 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

**100% Actual & Verified — Instant Download, Please Click**  
**[Order The DAS-C01 Practice Test Here](#)**