



# Amazon-Web-Services

## Exam Questions DAS-C01

AWS Certified Data Analytics - Specialty

#### NEW QUESTION 1

A media analytics company consumes a stream of social media posts. The posts are sent to an Amazon Kinesis data stream partitioned on user\_id. An AWS Lambda function retrieves the records and validates the content before loading the posts into an Amazon Elasticsearch cluster. The validation process needs to receive the posts for a given user in the order they were received. A data analyst has noticed that, during peak hours, the social media platform posts take more than an hour to appear in the Elasticsearch cluster.

What should the data analyst do reduce this latency?

- A. Migrate the validation process to Amazon Kinesis Data Firehose.
- B. Migrate the Lambda consumers from standard data stream iterators to an HTTP/2 stream consumer.
- C. Increase the number of shards in the stream.
- D. Configure multiple Lambda functions to process the stream.

**Answer: D**

#### NEW QUESTION 2

A company that monitors weather conditions from remote construction sites is setting up a solution to collect temperature data from the following two weather stations.

- > Station A, which has 10 sensors
- > Station B, which has five sensors

These weather stations were placed by onsite subject-matter experts.

Each sensor has a unique ID. The data collected from each sensor will be collected using Amazon Kinesis Data Streams.

Based on the total incoming and outgoing data throughput, a single Amazon Kinesis data stream with two shards is created. Two partition keys are created based on the station names. During testing, there is a bottleneck on data coming from Station A, but not from Station B. Upon review, it is confirmed that the total stream throughput is still less than the allocated Kinesis Data Streams throughput.

How can this bottleneck be resolved without increasing the overall cost and complexity of the solution, while retaining the data collection quality requirements?

- A. Increase the number of shards in Kinesis Data Streams to increase the level of parallelism.
- B. Create a separate Kinesis data stream for Station A with two shards, and stream Station A sensor data to the new stream.
- C. Modify the partition key to use the sensor ID instead of the station name.
- D. Reduce the number of sensors in Station A from 10 to 5 sensors.

**Answer: C**

#### Explanation:

<https://docs.aws.amazon.com/streams/latest/dev/kinesis-using-sdk-java-resharding.html>

"Splitting increases the number of shards in your stream and therefore increases the data capacity of the stream. Because you are charged on a per-shard basis, splitting increases the cost of your stream"

#### NEW QUESTION 3

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 4

A large telecommunications company is planning to set up a data catalog and metadata management for multiple data sources running on AWS. The catalog will be used to maintain the metadata of all the objects stored in the data stores. The data stores are composed of structured sources like Amazon RDS and Amazon Redshift, and semistructured sources like JSON and XML files stored in Amazon S3. The catalog must be updated on a regular basis, be able to detect the changes to object metadata, and require the least possible administration.

Which solution meets these requirements?

- A. Use Amazon Aurora as the data catalog
- B. Create AWS Lambda functions that will connect and gather the metadata information from multiple sources and update the data catalog in Aurora
- C. Schedule the Lambda functions periodically.
- D. Use the AWS Glue Data Catalog as the central metadata repository
- E. Use AWS Glue crawlers to connect to multiple data stores and update the Data Catalog with metadata change
- F. Schedule the crawlers periodically to update the metadata catalog.
- G. Use Amazon DynamoDB as the data catalog
- H. Create AWS Lambda functions that will connect and gather the metadata information from multiple sources and update the DynamoDB catalog
- I. Schedule the Lambda functions periodically.
- J. Use the AWS Glue Data Catalog as the central metadata repository
- K. Extract the schema for RDS and Amazon Redshift sources and build the Data Catalog
- L. Use AWS crawlers for data stored in Amazon S3 to infer the schema and automatically update the Data Catalog.

**Answer: D**

#### 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 has developed several AWS Glue jobs to validate and transform its data from Amazon S3 and load it into Amazon RDS for MySQL in batches once every day. The ETL jobs read the S3 data using a DynamicFrame. Currently, the ETL developers are experiencing challenges in processing only the incremental data on every run, as the AWS Glue job processes all the S3 input data on each run. Which approach would allow the developers to solve the issue with minimal coding effort?

- A. Have the ETL jobs read the data from Amazon S3 using a DataFrame.
- B. Enable job bookmarks on the AWS Glue jobs.
- C. Create custom logic on the ETL jobs to track the processed S3 objects.
- D. Have the ETL jobs delete the processed objects or data from Amazon S3 after each run.

**Answer:** B

#### NEW QUESTION 7

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 8

A global company has different sub-organizations, and each sub-organization sells its products and services in various countries. The company's senior leadership wants to quickly identify which sub-organization is the strongest performer in each country. All sales data is stored in Amazon S3 in Parquet format. Which approach can provide the visuals that senior leadership requested with the least amount of effort?

- A. Use Amazon QuickSight with Amazon Athena as the data sourc
- B. Use heat maps as the visual type.
- C. Use Amazon QuickSight with Amazon S3 as the data sourc
- D. Use heat maps as the visual type.
- E. Use Amazon QuickSight with Amazon Athena as the data sourc
- F. Use pivot tables as the visual type.
- G. Use Amazon QuickSight with Amazon S3 as the data sourc
- H. Use pivot tables as the visual type.

**Answer:** A

#### NEW QUESTION 9

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 10**

A large company receives files from external parties in Amazon EC2 throughout the day. At the end of the day, the files are combined into a single file, compressed into a gzip file, and uploaded to Amazon S3. The total size of all the files is close to 100 GB daily. Once the files are uploaded to Amazon S3, an AWS Batch program executes a COPY command to load the files into an Amazon Redshift cluster. Which program modification will accelerate the COPY process?

- A. Upload the individual files to Amazon S3 and run the COPY command as soon as the files become available.
- B. Split the number of files so they are equal to a multiple of the number of slices in the Amazon Redshift cluster.
- C. Gzip and upload the files to Amazon S3. Run the COPY command on the files.
- D. Split the number of files so they are equal to a multiple of the number of compute nodes in the Amazon Redshift cluster.
- E. Gzip and upload the files to Amazon S3. Run the COPY command on the files.
- F. Apply sharding by breaking up the files so the distkey columns with the same values go to the same file. Gzip and upload the sharded files to Amazon S3. Run the COPY command on the files.

**Answer:** B

**NEW QUESTION 10**

A company is planning to create a data lake in Amazon S3. The company wants to create tiered storage based on access patterns and cost objectives. The solution must include support for JDBC connections from legacy clients, metadata management that allows federation for access control, and batch-based ETL using PySpark and Scala. Operational management should be limited. Which combination of components can meet these requirements? (Choose three.)

- A. AWS Glue Data Catalog for metadata management
- B. Amazon EMR with Apache Spark for ETL
- C. AWS Glue for Scala-based ETL
- D. Amazon EMR with Apache Hive for JDBC clients
- E. Amazon Athena for querying data in Amazon S3 using JDBC drivers
- F. Amazon EMR with Apache Hive, using an Amazon RDS with MySQL-compatible backed metastore

**Answer:** BEF

**NEW QUESTION 13**

A marketing company has data in Salesforce, MySQL, and Amazon S3. The company wants to use data from these three locations and create mobile dashboards for its users. The company is unsure how it should create the dashboards and needs a solution with the least possible customization and coding. Which solution meets these requirements?

- A. Use Amazon Athena federated queries to join the data source
- B. Use Amazon QuickSight to generate the mobile dashboards.
- C. Use AWS Lake Formation to migrate the data sources into Amazon S3. Use Amazon QuickSight to generate the mobile dashboards.
- D. Use Amazon Redshift federated queries to join the data source
- E. Use Amazon QuickSight to generate the mobile dashboards.
- F. Use Amazon QuickSight to connect to the data sources and generate the mobile dashboards.

**Answer:** C

**NEW QUESTION 16**

A company is building a data lake and needs to ingest data from a relational database that has time-series data. The company wants to use managed services to accomplish this. The process needs to be scheduled daily and bring incremental data only from the source into Amazon S3. What is the MOST cost-effective approach to meet these requirements?

- A. Use AWS Glue to connect to the data source using JDBC Driver
- B. Ingest incremental records only using job bookmarks.
- C. Use AWS Glue to connect to the data source using JDBC Driver
- D. Store the last updated key in an Amazon DynamoDB table and ingest the data using the updated key as a filter.
- E. Use AWS Glue to connect to the data source using JDBC Drivers and ingest the entire dataset
- F. Use appropriate Apache Spark libraries to compare the dataset, and find the delta.
- G. Use AWS Glue to connect to the data source using JDBC Drivers and ingest the full dataset
- H. Use AWS DataSync to ensure the delta only is written into Amazon S3.

**Answer:** A

**Explanation:**

<https://docs.aws.amazon.com/glue/latest/dg/monitor-continuations.html>

**NEW QUESTION 20**

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 22

A company has an application that uses the Amazon Kinesis Client Library (KCL) to read records from a Kinesis data stream. After a successful marketing campaign, the application experienced a significant increase in usage. As a result, a data analyst had to split some shards in the data stream. When the shards were split, the application started throwing an ExpiredIteratorExceptions error sporadically. What should the data analyst do to resolve this?

- A. Increase the number of threads that process the stream records.
- B. Increase the provisioned read capacity units assigned to the stream's Amazon DynamoDB table.
- C. Increase the provisioned write capacity units assigned to the stream's Amazon DynamoDB table.
- D. Decrease the provisioned write capacity units assigned to the stream's Amazon DynamoDB table.

**Answer:** C

#### NEW QUESTION 27

A company owns facilities with IoT devices installed across the world. The company is using Amazon Kinesis Data Streams to stream data from the devices to Amazon S3. The company's operations team wants to get insights from the IoT data to monitor data quality at ingestion. The insights need to be derived in near-real time, and the output must be logged to Amazon DynamoDB for further analysis. Which solution meets these requirements?

- A. Connect Amazon Kinesis Data Analytics to analyze the stream data
- B. Save the output to DynamoDB by using the default output from Kinesis Data Analytics.
- C. Connect Amazon Kinesis Data Analytics to analyze the stream data
- D. Save the output to DynamoDB by using an AWS Lambda function.
- E. Connect Amazon Kinesis Data Firehose to analyze the stream data by using an AWS Lambda function. Save the output to DynamoDB by using the default output from Kinesis Data Firehose.
- F. Connect Amazon Kinesis Data Firehose to analyze the stream data by using an AWS Lambda function. Save the data to Amazon S3. Then run an AWS Glue job on schedule to ingest the data into DynamoDB.

**Answer:** C

#### NEW QUESTION 28

A company currently uses Amazon Athena to query its global datasets. The regional data is stored in Amazon S3 in the us-east-1 and us-west-2 Regions. The data is not encrypted. To simplify the query process and manage it centrally, the company wants to use Athena in us-west-2 to query data from Amazon S3 in both Regions. The solution should be as low-cost as possible. What should the company do to achieve this goal?

- A. Use AWS DMS to migrate the AWS Glue Data Catalog from us-east-1 to us-west-2. Run Athena queries in us-west-2.
- B. Run the AWS Glue crawler in us-west-2 to catalog datasets in all Region
- C. Once the data is crawled, run Athena queries in us-west-2.
- D. Enable cross-Region replication for the S3 buckets in us-east-1 to replicate data in us-west-2. Once the data is replicated in us-west-2, run the AWS Glue crawler there to update the AWS Glue Data Catalog in us-west-2 and run Athena queries.
- E. Update AWS Glue resource policies to provide us-east-1 AWS Glue Data Catalog access to us-west-2. Once the catalog in us-west-2 has access to the catalog in us-east-1, run Athena queries in us-west-2.

**Answer:** B

#### NEW QUESTION 30

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 35

A large ride-sharing company has thousands of drivers globally serving millions of unique customers every day. The company has decided to migrate an existing data mart to Amazon Redshift. The existing schema includes the following tables.

A trips fact table for information on completed rides. A drivers dimension table for driver profiles. A customers fact table holding customer profile information. The company analyzes trip details by date and destination to examine profitability by region. The drivers data rarely changes. The customers data frequently changes.

What table design provides optimal query performance?

- A. Use DISTSTYLE KEY (destination) for the trips table and sort by date
- B. Use DISTSTYLE ALL for the drivers and customers tables.
- C. Use DISTSTYLE EVEN for the trips table and sort by date
- D. Use DISTSTYLE ALL for the drivers table. Use DISTSTYLE EVEN for the customers table.
- E. Use DISTSTYLE KEY (destination) for the trips table and sort by date
- F. Use DISTSTYLE ALL for the drivers table
- G. Use DISTSTYLE EVEN for the customers table.
- H. Use DISTSTYLE EVEN for the drivers table and sort by date
- I. Use DISTSTYLE ALL for both fact tables.

**Answer: C**

### Explanation:

[https://www.matillion.com/resources/blog/aws-redshift-performance-choosing-the-right-distribution-styles/#:~:text="](https://www.matillion.com/resources/blog/aws-redshift-performance-choosing-the-right-distribution-styles/#:~:text=)

[https://docs.aws.amazon.com/redshift/latest/dg/c\\_best-practices-best-dist-key.html](https://docs.aws.amazon.com/redshift/latest/dg/c_best-practices-best-dist-key.html)

### NEW QUESTION 37

Three teams of data analysts use Apache Hive on an Amazon EMR cluster with the EMR File System (EMRFS) to query data stored within each team's Amazon S3 bucket. The EMR cluster has Kerberos enabled and is configured to authenticate users from the corporate Active Directory. The data is highly sensitive, so access must be limited to the members of each team.

Which steps will satisfy the security requirements?

- A. For the EMR cluster Amazon EC2 instances, create a service role that grants no access to Amazon S3. Create three additional IAM roles, each granting access to each team's specific bucket
- B. Add the additional IAM roles to the cluster's EMR role for the EC2 trust policy
- C. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.
- D. For the EMR cluster Amazon EC2 instances, create a service role that grants no access to Amazon S3. Create three additional IAM roles, each granting access to each team's specific bucket
- E. Add the service role for the EMR cluster EC2 instances to the trust policies for the additional IAM role
- F. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.
- G. For the EMR cluster Amazon EC2 instances, create a service role that grants full access to Amazon S3. Create three additional IAM roles, each granting access to each team's specific bucket
- H. Add the service role for the EMR cluster EC2 instances to the trust policies for the additional IAM role
- I. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.
- J. For the EMR cluster Amazon EC2 instances, create a service role that grants full access to Amazon S3. Create three additional IAM roles, each granting access to each team's specific bucket
- K. Add the service role for the EMR cluster EC2 instances to the trust policies for the base IAM role
- L. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.

**Answer: C**

### NEW QUESTION 40

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 42

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 44**

A manufacturing company uses Amazon S3 to store its data. The company wants to use AWS Lake Formation to provide granular-level security on those data assets. The data is in Apache Parquet format. The company has set a deadline for a consultant to build a data lake. How should the consultant create the MOST cost-effective solution that meets these requirements?

- A. Run Lake Formation blueprints to move the data to Lake Formation
- B. Once Lake Formation has the data, apply permissions on Lake Formation.
- C. To create the data catalog, run an AWS Glue crawler on the existing Parquet data
- D. Register the Amazon S3 path and then apply permissions through Lake Formation to provide granular-level security.
- E. Install Apache Ranger on an Amazon EC2 instance and integrate with Amazon EM
- F. Using Ranger policies, create role-based access control for the existing data assets in Amazon S3.
- G. Create multiple IAM roles for different users and group
- H. Assign IAM roles to different data assets in Amazon S3 to create table-based and column-based access controls.

**Answer:** A

**Explanation:**

<https://aws.amazon.com/blogs/big-data/building-securing-and-managing-data-lakes-with-aws-lake-formation/>

**NEW QUESTION 47**

A company has an application that ingests streaming data. The company needs to analyze this stream over a 5-minute timeframe to evaluate the stream for anomalies with Random Cut Forest (RCF) and summarize the current count of status codes. The source and summarized data should be persisted for future use. Which approach would enable the desired outcome while keeping data persistence costs low?

- A. Ingest the data stream with Amazon Kinesis Data Stream
- B. Have an AWS Lambda consumer evaluate the stream, collect the number status codes, and evaluate the data against a previously trained RCF model
- C. Persist the source and results as a time series to Amazon DynamoDB.
- D. Ingest the data stream with Amazon Kinesis Data Stream
- E. Have a Kinesis Data Analytics application evaluate the stream over a 5-minute window using the RCF function and summarize the count of status code
- F. Persist the source and results to Amazon S3 through output delivery to Kinesis Data Firehouse.
- G. Ingest the data stream with Amazon Kinesis Data Firehose with a delivery frequency of 1 minute or 1 MB in Amazon S3. Ensure Amazon S3 triggers an event to invoke an AWS Lambda consumer that evaluates the batch data, collects the number status codes, and evaluates the data against a previously trained RCF model
- H. Persist the source and results as a time series to Amazon DynamoDB.
- I. Ingest the data stream with Amazon Kinesis Data Firehose with a delivery frequency of 5 minutes or 1 MB into Amazon S3. Have a Kinesis Data Analytics application evaluate the stream over a 1-minute window using the RCF function and summarize the count of status code
- J. Persist the results to Amazon S3 through a Kinesis Data Analytics output to an AWS Lambda integration.

**Answer:** B

**NEW QUESTION 48**

A company wants to use an automatic machine learning (ML) Random Cut Forest (RCF) algorithm to visualize complex real-world scenarios, such as detecting seasonality and trends, excluding outliers, and imputing missing values. The team working on this project is non-technical and is looking for an out-of-the-box solution that will require the LEAST amount of management overhead. Which solution will meet these requirements?

- A. Use an AWS Glue ML transform to create a forecast and then use Amazon QuickSight to visualize the data.
- B. Use Amazon QuickSight to visualize the data and then use ML-powered forecasting to forecast the key business metrics.
- C. Use a pre-build ML AMI from the AWS Marketplace to create forecasts and then use Amazon QuickSight to visualize the data.
- D. Use calculated fields to create a new forecast and then use Amazon QuickSight to visualize the data.

**Answer:** A

**NEW QUESTION 50**

An airline has .csv-formatted data stored in Amazon S3 with an AWS Glue Data Catalog. Data analysts want to join this data with call center data stored in Amazon Redshift as part of a daily batch process. The Amazon Redshift cluster is already under a heavy load. The solution must be managed, serverless, well-functioning, and minimize the load on the existing Amazon Redshift cluster. The solution should also require minimal effort and development activity. Which solution meets these requirements?

- A. Unload the call center data from Amazon Redshift to Amazon S3 using an AWS Lambda function. Perform the join with AWS Glue ETL scripts.
- B. Export the call center data from Amazon Redshift using a Python shell in AWS Glue
- C. Perform the join with AWS Glue ETL scripts.
- D. Create an external table using Amazon Redshift Spectrum for the call center data and perform the join with Amazon Redshift.
- E. Export the call center data from Amazon Redshift to Amazon EMR using Apache Sqoop
- F. Perform the join with Apache Hive.

**Answer:** C

**Explanation:**

<https://docs.aws.amazon.com/redshift/latest/dg/c-spectrum-external-tables.html>

**NEW QUESTION 54**

A retail company has 15 stores across 6 cities in the United States. Once a month, the sales team requests a visualization in Amazon QuickSight that provides the ability to easily identify revenue trends across cities and stores. The visualization also helps identify outliers that need to be examined with further analysis. Which visual type in QuickSight meets the sales team's requirements?

- A. Geospatial chart
- B. Line chart
- C. Heat map
- D. Tree map

**Answer:** A

**NEW QUESTION 59**

A company is sending historical datasets to Amazon S3 for storage. A data engineer at the company wants to make these datasets available for analysis using Amazon Athena. The engineer also wants to encrypt the Athena query results in an S3 results location by using AWS solutions for encryption. The requirements for encrypting the query results are as follows:

Use custom keys for encryption of the primary dataset query results.

Use generic encryption for all other query results.

Provide an audit trail for the primary dataset queries that shows when the keys were used and by whom. Which solution meets these requirements?

- A. Use server-side encryption with S3 managed encryption keys (SSE-S3) for the primary dataset
- B. Use SSE-S3 for the other datasets.
- C. Use server-side encryption with customer-provided encryption keys (SSE-C) for the primary dataset. Use server-side encryption with S3 managed encryption keys (SSE-S3) for the other datasets.
- D. Use server-side encryption with AWS KMS managed customer master keys (SSE-KMS CMKs) for the primary dataset
- E. Use server-side encryption with S3 managed encryption keys (SSE-S3) for the other datasets.
- F. Use client-side encryption with AWS Key Management Service (AWS KMS) customer managed keys for the primary dataset
- G. Use S3 client-side encryption with client-side keys for the other datasets.

**Answer:** A

**NEW QUESTION 63**

A company wants to optimize the cost of its data and analytics platform. The company is ingesting a number of .csv and JSON files in Amazon S3 from various data sources. Incoming data is expected to be 50 GB each day. The company is using Amazon Athena to query the raw data in Amazon S3 directly. Most queries aggregate data from the past 12 months, and data that is older than 5 years is infrequently queried. The typical query scans about 500 MB of data and is expected to return results in less than 1 minute. The raw data must be retained indefinitely for compliance requirements.

Which solution meets the company's requirements?

- A. Use an AWS Glue ETL job to compress, partition, and convert the data into a columnar data format
- B. Use Athena to query the processed dataset
- C. Configure a lifecycle policy to move the processed data into the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class 5 years after object creation
- D. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after object creation.
- E. Use an AWS Glue ETL job to partition and convert the data into a row-based data format
- F. Use Athena to query the processed dataset
- G. Configure a lifecycle policy to move the data into the Amazon S3 Standard- Infrequent Access (S3 Standard-IA) storage class 5 years after object creation
- H. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after object creation.
- I. Use an AWS Glue ETL job to compress, partition, and convert the data into a columnar data format
- J. Use Athena to query the processed dataset
- K. Configure a lifecycle policy to move the processed data into the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class 5 years after the object was last accessed
- L. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after the last date the object was accessed.
- M. Use an AWS Glue ETL job to partition and convert the data into a row-based data format
- N. Use Athena to query the processed dataset
- O. Configure a lifecycle policy to move the data into the Amazon S3 Standard- Infrequent Access (S3 Standard-IA) storage class 5 years after the object was last accessed
- P. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after the last date the object was accessed.

**Answer:** A

**NEW QUESTION 67**

A retail company is building its data warehouse solution using Amazon Redshift. As a part of that effort, the company is loading hundreds of files into the fact table created in its Amazon Redshift cluster. The company wants the solution to achieve the highest throughput and optimally use cluster resources when loading data into the company's fact table.

How should the company meet these requirements?

- A. Use multiple COPY commands to load the data into the Amazon Redshift cluster.
- B. Use S3DistCp to load multiple files into the Hadoop Distributed File System (HDFS) and use an HDFS connector to ingest the data into the Amazon Redshift cluster.
- C. Use LOAD commands equal to the number of Amazon Redshift cluster nodes and load the data in parallel into each node.
- D. Use a single COPY command to load the data into the Amazon Redshift cluster.

**Answer:** D

**Explanation:**

[https://docs.aws.amazon.com/redshift/latest/dg/c\\_best-practices-single-copy-command.html](https://docs.aws.amazon.com/redshift/latest/dg/c_best-practices-single-copy-command.html)

#### NEW QUESTION 69

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 73

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 75

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 connector
- 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 basis
- 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 format
- H. Query the compressed data.
- I. Run a daily AWS Glue ETL job to compress the data files by using the .lzo format
- J. Query the compressed data.

**Answer:** BC

#### NEW QUESTION 76

A financial services company needs to aggregate daily stock trade data from the exchanges into a data store. The company requires that data be streamed directly into the data store, but also occasionally allows data to be modified using SQL. The solution should integrate complex, analytic queries running with minimal latency. The solution must provide a business intelligence dashboard that enables viewing of the top contributors to anomalies in stock prices.

Which solution meets the company's requirements?

- A. Use Amazon Kinesis Data Firehose to stream data to Amazon S3. Use Amazon Athena as a data source for Amazon QuickSight to create a business intelligence dashboard.
- B. Use Amazon Kinesis Data Streams to stream data to Amazon Redshift
- C. Use Amazon Redshift as a data source for Amazon QuickSight to create a business intelligence dashboard.
- D. Use Amazon Kinesis Data Firehose to stream data to Amazon Redshift
- E. Use Amazon Redshift as a data source for Amazon QuickSight to create a business intelligence dashboard.
- F. Use Amazon Kinesis Data Streams to stream data to Amazon S3. Use Amazon Athena as a data source for Amazon QuickSight to create a business intelligence dashboard.

**Answer:** C

**NEW QUESTION 78**

.....

## 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 First Try
- \* DAS-C01 Practice Test Questions in Multiple Choice Formats and Updates for 1 Year

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