

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 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 company wants to enrich application logs in near-real-time and use the enriched dataset for further analysis. The application is running on Amazon EC2 instances across multiple Availability Zones and storing its logs using Amazon CloudWatch Logs. The enrichment source is stored in an Amazon DynamoDB table. Which solution meets the requirements for the event collection and enrichment?

- A. Use a CloudWatch Logs subscription to send the data to Amazon Kinesis Data Firehose
- B. Use AWS Lambda to transform the data in the Kinesis Data Firehose delivery stream and enrich it with the data in the DynamoDB tabl
- C. Configure Amazon S3 as the Kinesis Data Firehose delivery destination.
- D. Export the raw logs to Amazon S3 on an hourly basis using the AWS CL
- E. Use AWS Glue crawlers to catalog the log
- F. Set up an AWS Glue connection for the DynamoDB table and set up an AWS Glue ETL job to enrich the dat
- G. Store the enriched data in Amazon S3.
- H. Configure the application to write the logs locally and use Amazon Kinesis Agent to send the data to Amazon Kinesis Data Stream
- I. Configure a Kinesis Data Analytics SQL application with the Kinesis data stream as the sourc
- J. Join the SQL application input stream with DynamoDB records, and then store the enriched output stream in Amazon S3 using Amazon Kinesis Data Firehose.
- K. Export the raw logs to Amazon S3 on an hourly basis using the AWS CL
- L. Use Apache Spark SQL on Amazon EMR to read the logs from Amazon S3 and enrich the records with the data from DynamoD
- M. Store the enriched data in Amazon S3.

Answer: A

Explanation:

<https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/SubscriptionFilters.html#FirehoseExample>

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 catalo
- B. Create AWS Lambda functions that will connect and gather themetadata information from multiple sources and update the data catalog in Auror
- C. Schedule the Lambda functions periodically.
- D. Use the AWS Glue Data Catalog as the central metadata repositor
- 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 catalo
- H. Create AWS Lambda functions that will connect and gather the metadata information from multiple sources and update the DynamoDB catalo
- I. Schedule the Lambda functions periodically.
- J. Use the AWS Glue Data Catalog as the central metadata repositor
- K. Extract the schema for RDS and Amazon Redshift sources and build the Data Catalo
- 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 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 6

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 7

A financial company uses Amazon S3 as its data lake and has set up a data warehouse using a multi-node Amazon Redshift cluster. The data files in the data lake are organized in folders based on the data source of each data file. All the data files are loaded to one table in the Amazon Redshift cluster using a separate COPY command for each data file location. With this approach, loading all the data files into Amazon Redshift takes a long time to complete. Users want a faster solution with little or no increase in cost while maintaining the segregation of the data files in the S3 data lake.

Which solution meets these requirements?

- A. Use Amazon EMR to copy all the data files into one folder and issue a COPY command to load the data into Amazon Redshift.
- B. Load all the data files in parallel to Amazon Aurora, and run an AWS Glue job to load the data into Amazon Redshift.
- C. Use an AWS Glue job to copy all the data files into one folder and issue a COPY command to load the data into Amazon Redshift.
- D. Create a manifest file that contains the data file locations and issue a COPY command to load the data into Amazon Redshift.

Answer: D

Explanation:

<https://docs.aws.amazon.com/redshift/latest/dg/loading-data-files-using-manifest.html> "You can use a manifest to ensure that the COPY command loads all of the required files, and only the required files, for a data load"

NEW QUESTION 8

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 9

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 10

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 10

A hospital uses wearable medical sensor devices to collect data from patients. The hospital is architecting a near-real-time solution that can ingest the data securely at scale. The solution should also be able to remove the patient's protected health information (PHI) from the streaming data and store the data in durable storage.

Which solution meets these requirements with the least operational overhead?

- A. Ingest the data using Amazon Kinesis Data Streams, which invokes an AWS Lambda function using Kinesis Client Library (KCL) to remove all PHI
- B. Write the data in Amazon S3.
- C. Ingest the data using Amazon Kinesis Data Firehose to write the data to Amazon S3. Have Amazon S3 trigger an AWS Lambda function that parses the sensor data to remove all PHI in Amazon S3.
- D. Ingest the data using Amazon Kinesis Data Streams to write the data to Amazon S3. Have the data stream launch an AWS Lambda function that parses the sensor data and removes all PHI in Amazon S3.
- E. Ingest the data using Amazon Kinesis Data Firehose to write the data to Amazon S3. Implement a transformation AWS Lambda function that parses the sensor data to remove all PHI.

Answer: D

Explanation:

<https://aws.amazon.com/blogs/big-data/persist-streaming-data-to-amazon-s3-using-amazon-kinesis-firehose-and>

NEW QUESTION 11

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 14

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 15

A mortgage company has a microservice for accepting payments. This microservice uses the Amazon DynamoDB encryption client with AWS KMS managed keys to encrypt the sensitive data before writing the data to DynamoDB. The finance team should be able to load this data into Amazon Redshift and aggregate the values within the sensitive fields. The Amazon Redshift cluster is shared with other data analysts from different business units.

Which steps should a data analyst take to accomplish this task efficiently and securely?

- A. Create an AWS Lambda function to process the DynamoDB stream
- B. Decrypt the sensitive data using the same KMS key
- C. Save the output to a restricted S3 bucket for the finance team
- D. Create a finance table in Amazon Redshift that is accessible to the finance team only

- E. Use the COPY command to load the data from Amazon S3 to the finance table.
- F. Create an AWS Lambda function to process the DynamoDB stream.
- G. Save the output to a restricted S3 bucket for the finance team.
- H. Create a finance table in Amazon Redshift that is accessible to the finance team only.
- I. Use the COPY command with the IAM role that has access to the KMS key to load the data from S3 to the finance table.
- J. Create an Amazon EMR cluster with an EMR_EC2_DefaultRole role that has access to the KMS key. Create Apache Hive tables that reference the data stored in DynamoDB and the finance table in Amazon Redshift.
- K. In Hive, select the data from DynamoDB and then insert the output to the finance table in Amazon Redshift.
- L. Create an Amazon EMR cluster.
- M. Create Apache Hive tables that reference the data stored in DynamoDB.
- N. Insert the output to the restricted Amazon S3 bucket for the finance team.
- O. Use the COPY command with the IAM role that has access to the KMS key to load the data from Amazon S3 to the finance table in Amazon Redshift.

Answer: B

NEW QUESTION 20

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 23

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=The%20fact%20tables%20are%20fact%20tables,>

https://docs.aws.amazon.com/redshift/latest/dg/c_best-practices-best-dist-key.html

NEW QUESTION 28

A data analytics specialist is setting up workload management in manual mode for an Amazon Redshift environment. The data analytics specialist is defining query monitoring rules to manage system performance and user experience of an Amazon Redshift cluster.

Which elements must each query monitoring rule include?

- A. A unique rule name, a query runtime condition, and an AWS Lambda function to resubmit any failed queries in off hours
- B. A queue name, a unique rule name, and a predicate-based stop condition
- C. A unique rule name, one to three predicates, and an action
- D. A workload name, a unique rule name, and a query runtime-based condition

Answer: C

NEW QUESTION 31

Once a month, a company receives a 100 MB .csv file compressed with gzip. The file contains 50,000 property listing records and is stored in Amazon S3 Glacier. The company needs its data analyst to query a subset of the data for a specific vendor.

What is the most cost-effective solution?

- A. Load the data into Amazon S3 and query it with Amazon S3 Select.
- B. Query the data from Amazon S3 Glacier directly with Amazon Glacier Select.
- C. Load the data to Amazon S3 and query it with Amazon Athena.

D. Load the data to Amazon S3 and query it with Amazon Redshift Spectrum.

Answer: A

NEW QUESTION 35

A human resources company maintains a 10-node Amazon Redshift cluster to run analytics queries on the company's data. The Amazon Redshift cluster contains a product table and a transactions table, and both tables have a product_sku column. The tables are over 100 GB in size. The majority of queries run on both tables.

Which distribution style should the company use for the two tables to achieve optimal query performance?

- A. An EVEN distribution style for both tables
- B. A KEY distribution style for both tables
- C. An ALL distribution style for the product table and an EVEN distribution style for the transactions table
- D. An EVEN distribution style for the product table and an KEY distribution style for the transactions table

Answer: B

NEW QUESTION 40

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 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 45

A company wants to provide its data analysts with uninterrupted access to the data in its Amazon Redshift cluster. All data is streamed to an Amazon S3 bucket with Amazon Kinesis Data Firehose. An AWS Glue job that is scheduled to run every 5 minutes issues a COPY command to move the data into Amazon Redshift. The amount of data delivered is uneven throughout the day, and cluster utilization is high during certain periods. The COPY command usually completes within a couple of seconds. However, when load spike occurs, locks can exist and data can be missed. Currently, the AWS Glue job is configured to run without retries, with timeout at 5 minutes and concurrency at 1.

How should a data analytics specialist configure the AWS Glue job to optimize fault tolerance and improve data availability in the Amazon Redshift cluster?

- A. Increase the number of retries
- B. Decrease the timeout value
- C. Increase the job concurrency.
- D. Keep the number of retries at 0. Decrease the timeout value
- E. Increase the job concurrency.
- F. Keep the number of retries at 0. Decrease the timeout value
- G. Keep the job concurrency at 1.
- H. Keep the number of retries at 0. Increase the timeout value
- I. Keep the job concurrency at 1.

Answer: B

NEW QUESTION 48

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 bucket
- 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 command
- 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 source
- 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 53

A company has a business unit uploading .csv files to an Amazon S3 bucket. The company's data platform team has set up an AWS Glue crawler to do discovery, and create tables and schemas. An AWS Glue job writes processed data from the created tables to an Amazon Redshift database. The AWS Glue job handles column mapping and creating the Amazon Redshift table appropriately. When the AWS Glue job is rerun for any reason in a day, duplicate records are introduced into the Amazon Redshift table.

Which solution will update the Redshift table without duplicates when jobs are rerun?

- A. Modify the AWS Glue job to copy the rows into a staging table
- B. Add SQL commands to replace the existing rows in the main table as postactions in the DynamicFrameWriter class.
- C. Load the previously inserted data into a MySQL database in the AWS Glue job
- D. Perform an upsert operation in MySQL, and copy the results to the Amazon Redshift table.
- E. Use Apache Spark's DataFrame dropDuplicates() API to eliminate duplicates and then write the data to Amazon Redshift.
- F. Use the AWS Glue ResolveChoice built-in transform to select the most recent value of the column.

Answer: A

Explanation:

<https://aws.amazon.com/premiumsupport/knowledge-center/sql-commands-redshift-glue-job/> See the section Merge an Amazon Redshift table in AWS Glue (upsert)

NEW QUESTION 57

A university intends to use Amazon Kinesis Data Firehose to collect JSON-formatted batches of water quality readings in Amazon S3. The readings are from 50 sensors scattered across a local lake. Students will query the stored data using Amazon Athena to observe changes in a captured metric over time, such as water temperature or acidity. Interest has grown in the study, prompting the university to reconsider how data will be stored.

Which data format and partitioning choices will MOST significantly reduce costs? (Choose two.)

- A. Store the data in Apache Avro format using Snappy compression.
- B. Partition the data by year, month, and day.
- C. Store the data in Apache ORC format using no compression.
- D. Store the data in Apache Parquet format using Snappy compression.
- E. Partition the data by sensor, year, month, and day.

Answer: CD

NEW QUESTION 59

A company is planning to do a proof of concept for a machine learning (ML) project using Amazon SageMaker with a subset of existing on-premises data hosted in the company's 3 TB data warehouse. For part of the project, AWS Direct Connect is established and tested. To prepare the data for ML, data analysts are performing data curation. The data analysts want to perform multiple steps, including mapping, dropping null fields, resolving choice, and splitting fields. The company needs the fastest solution to curate the data for this project.

Which solution meets these requirements?

- A. Ingest data into Amazon S3 using AWS DataSync and use Apache Spark scripts to curate the data in an Amazon EMR cluster
- B. Store the curated data in Amazon S3 for ML processing.
- C. Create custom ETL jobs on-premises to curate the data
- D. Use AWS DMS to ingest data into Amazon S3 for ML processing.
- E. Ingest data into Amazon S3 using AWS DMS
- F. Use AWS Glue to perform data curation and store the data in Amazon S3 for ML processing.
- G. Take a full backup of the data store and ship the backup files using AWS Snowball
- H. Upload Snowball data into Amazon S3 and schedule data curation jobs using AWS Batch to prepare the data for ML.

Answer: C

NEW QUESTION 61

A company wants to improve the data load time of a sales data dashboard. Data has been collected as .csv files and stored within an Amazon S3 bucket that is partitioned by date. The data is then loaded to an Amazon Redshift data warehouse for frequent analysis. The data volume is up to 500 GB per day. Which solution will improve the data loading performance?

- A. Compress .csv files and use an INSERT statement to ingest data into Amazon Redshift.
- B. Split large .csv files, then use a COPY command to load data into Amazon Redshift.
- C. Use Amazon Kinesis Data Firehose to ingest data into Amazon Redshift.
- D. Load the .csv files in an unsorted key order and vacuum the table in Amazon Redshift.

Answer: B

Explanation:

https://docs.aws.amazon.com/redshift/latest/dg/c_loading-data-best-practices.html

NEW QUESTION 64

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 67

A large university has adopted a strategic goal of increasing diversity among enrolled students. The data analytics team is creating a dashboard with data visualizations to enable stakeholders to view historical trends. All access must be authenticated using Microsoft Active Directory. All data in transit and at rest must be encrypted.

Which solution meets these requirements?

- A. Amazon QuickSight Standard edition configured to perform identity federation using SAML 2.0. and the default encryption settings.
- B. Amazon QuickSight Enterprise edition configured to perform identity federation using SAML 2.0 and the default encryption settings.
- C. Amazon QuickSight Standard edition using AD Connector to authenticate using Active Directory. Configure Amazon QuickSight to use customer-provided keys imported into AWS KMS.
- D. Amazon QuickSight Enterprise edition using AD Connector to authenticate using Active Directory. Configure Amazon QuickSight to use customer-provided keys imported into AWS KMS.

Answer: D

NEW QUESTION 72

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 74

.....

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](#)