100% Money Back Guarantee

Vendor: Amazon

Exam Code:DAS-C01

Exam Name:AWS Certified Data Analytics - Specialty (DAS-C01)

Version: Demo

QUESTION 1

A retail company stores order invoices in an Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster Indices on the cluster are created monthly. Once a new month begins, no new writes are made to any of the indices from the previous months. The company has been expanding the storage on the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster to avoid running out of space, but the company wants to reduce costs. Most searches on the cluster are on the most recent 3 months of data, while the audit team requires infrequent access to older data to generate periodic reports. The most recent 3 months of data must be quickly available for queries, but the audit team can tolerate slower queries if the solution saves on cluster costs

Which of the following is the MOST operationally efficient solution to meet these requirements?

A. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to store the indices in Amazon S3 Glacier. When the audit team requires the archived data, restore the archived indices back to the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster.

B. Archive indices that are older than 3 months by taking manual snapshots and storing the snapshots in Amazon S3. When the audit team requires the archived data, restore the archived indices back to the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster.

C. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to migrate the indices to Amazon OpenSearch Service (Amazon Elasticsearch Service) UltraWarm storage.

D. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to migrate the indices to Amazon OpenSearch Service (Amazon Elasticsearch Service) UltraWarm storage. When the audit team requires the older data, migrate the indices in UltraWarm storage back to hot storage.

Correct Answer: D

Reference: https://docs.aws.amazon.com/da_pv/opensearch-service/latest/developerguide/opensearch-service-dg.pdf

---

QUESTION 2

A retail company stores order invoices in an Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster. Indices on the cluster are created monthly. Once a new month begins, no new writes are made to any of the indices from the previous months. The company has been expanding the storage on the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster to avoid running out of space, but the company wants to reduce costs. Most searches on the cluster are on the most recent 3 months of data, while the audit team requires infrequent access to older data to generate periodic reports. The most recent 3 months of data must be quickly available for queries, but the audit team can tolerate slower queries if the solution saves on cluster costs.

Which of the following is the MOST operationally efficient solution to meet these requirements?

A. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to store the indices in Amazon S3 Glacier Instant Retrieval. When the audit team requires the archived data, restore the archived indices back to the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster.

B. Archive indices that are older than 3 months by taking manual snapshots and storing the snapshots in Amazon S3. When the audit team requires the archived data, restore the archived indices back to the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster.

C. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to migrate the indices to Amazon OpenSearch Service (Amazon Elasticsearch Service) UltraWarm storage.

D. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to migrate the indices to Amazon OpenSearch Service (Amazon Elasticsearch Service) UltraWarm storage. When the audit team requires the older data, migrate the indices in UltraWarm storage back to hot storage.

Correct Answer: C

QUESTION 3

A data architect is building an Amazon S3 data lake for a bank. The goal is to provide a single data repository for customer data needs, such as personalized recommendations. The bank uses Amazon Kinesis Data Firehose to ingest customers\\' personal information bank accounts, and transactions in near-real time from a transactional relational database. The bank requires all personally identifiable information (PII) that is stored in the AWS Cloud to be masked.

Which solution will meet these requirements?

A. Invoke an AWS Lambda function from Kinesis Data Firehose to mask PII before delivering the data into Amazon S3.

B. Use Amazon Made, and configure it to discover and mask PII.

C. Enable server-side encryption (SSE) in Amazon S3.

D. Invoke Amazon Comprehend from Kinesis Data Firehose to detect and mask PII before delivering the data into Amazon S3.

Correct Answer: C

Reference: https://docs.aws.amazon.com/AmazonS3/latest/userguide/UsingServerSideEncryption.html

---

QUESTION 4

A technology company has an application with millions of active users every day. The company queries daily usage data with Amazon Athena to understand how users interact with the application. The data includes the date and time, the location ID, and the services used. The company wants to use Athena to run queries to analyze the data with the lowest latency possible.

Which solution meets these requirements?

A. Store the data in Apache Avro format with the date and time as the partition, with the data sorted by the location ID.

B. Store the data in Apache Parquet format with the date and time as the partition, with the data sorted by the location ID.

C. Store the data in Apache ORC format with the location ID as the partition, with the data sorted by the date and time.

D. Store the data in .csv format with the location ID as the partition, with the data sorted by the date and time.

Correct Answer: B

Reference: https://cwiki.apache.org/confluence/display/hive/languagemanual+orc

---

QUESTION 5

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 data 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 to 5 PB over time. The query latency also has increased. The company needs to segment the data to reduce the amount of data that is scanned.

Which solutions will improve query performance? (Choose two.)

A. Use MySQL Workbench on an Amazon EC2 instance. Connect to Athena by using a JDBC connector. Run the query from MySQL Workbench instead of Athena directly.

B. Configure Athena to use S3 Select to load only the files of the data subset.

C. Create the data subset in Apache Parquet format each day by using the Athena CREATE TABLE AS SELECT (CTAS) statement. Query the Parquet data.

D. Run a daily AWS Glue ETL job to convert the data files to Apache Parquet format and to partition the converted files. Create a periodic AWS Glue crawler to automatically crawl the partitioned data each day.

E. Create an S3 gateway endpoint. Configure VPC routing to access Amazon S3 through the gateway endpoint.

Correct Answer: CE

QUESTION 6

A data engineer is using AWS Glue ETL jobs to process data at frequent intervals. The processed data is then copied into Amazon S3. The ETL jobs run every 15 minutes. The AWS Glue Data Catalog partitions need to be updated automatically after the completion of each job.

Which solution will meet these requirements MOST cost-effectively?

A. Use the AWS Glue Data Catalog to manage the data catalog. Define an AWS Glue workflow for the ETL process. Define a trigger within the workflow that can start the crawler when an ETL job run is complete.

B. Use the AWS Glue Data Catalog to manage the data catalog. Use AWS Glue Studio to manage ETL jobs. Use the AWS Glue Studio feature that supports updates to the AWS Glue Data Catalog during job runs.

C. Use an Apache Hive metastore to manage the data catalog. Update the AWS Glue ETL code to include the enableUpdateCatalog and partitionKeys arguments.

D. Use the AWS Glue Data Catalog to manage the data catalog. Update the AWS Glue ETL code to include the enableUpdateCatalog and partitionKeys arguments.

Correct Answer: A

Upon successful completion of both jobs, an event trigger, Fix/De-dupe succeeded, starts a crawler, Update schema.
Reference: https://docs.aws.amazon.com/glue/latest/dg/workflows_overview.html

---

QUESTION 7

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 Regions. Once the data is crawled, run Athena queries in us-west-2.

C. 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.

D. 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.

Correct Answer: B

Reference: https://docs.aws.amazon.com/athena/latest/ug/other-notable-limitations.html
https://docs.aws.amazon.com/glue/latest/dg/glue-resource-policies.html

---

QUESTION 8

An airline has been collecting metrics on flight activities for analytics. A recently completed proof of concept demonstrates how the company provides insights to data analysts to improve on-time departures. The proof of concept used objects in Amazon S3, which contained the metrics in .csv format, and used Amazon Athena for querying the data. As the amount of data increases, the data analyst wants to optimize the storage solution to improve query performance.

Which options should the data analyst use to improve performance as the data lake grows? (Choose three.)

A. Add a randomized string to the beginning of the keys in S3 to get more throughput across partitions.

B. Use an S3 bucket in the same account as Athena.

C. Compress the objects to reduce the data transfer I/O.

D. Use an S3 bucket in the same Region as Athena.

E. Preprocess the .csv data to JSON to reduce I/O by fetching only the document keys needed by the query.

F. Preprocess the .csv data to Apache Parquet to reduce I/O by fetching only the data blocks needed for predicates.

Correct Answer: ACE

QUESTION 9

A banking company wants to collect large volumes of transactional data using Amazon Kinesis Data Streams for real-time analytics. The company uses PutRecord to send data to Amazon Kinesis, and has observed network outages during certain times of the day. The company wants to obtain exactly once semantics for the entire processing pipeline.

What should the company do to obtain these characteristics?

A. Design the application so it can remove duplicates during processing be embedding a unique ID in each record.

B. Rely on the processing semantics of Amazon Kinesis Data Analytics to avoid duplicate processing of events.

C. Design the data producer so events are not ingested into Kinesis Data Streams multiple times.

D. Rely on the exactly one processing semantics of Apache Flink and Apache Spark Streaming included in Amazon EMR.

Correct Answer: A

Reference: https://docs.aws.amazon.com/streams/latest/dev/kinesis-record-processor-duplicates.html

---

QUESTION 10

A large company has a central data lake to run analytics across different departments. Each department uses a separate AWS account and stores its data in an Amazon S3 bucket in that account. Each AWS account uses the AWS Glue Data Catalog as its data catalog. There are different data lake access requirements based on roles. Associate analysts should only have read access to their departmental data. Senior data analysts can have access in multiple departments including theirs, but for a subset of columns only.

Which solution achieves these required access patterns to minimize costs and administrative tasks?

A. Consolidate all AWS accounts into one account. Create different S3 buckets for each department and move all the data from every account to the central data lake account. Migrate the individual data catalogs into a central data catalog and apply fine-grained permissions to give to each user the required access to tables and databases in AWS Glue and Amazon S3.

B. Keep the account structure and the individual AWS Glue catalogs on each account. Add a central data lake account and use AWS Glue to catalog data from various accounts. Configure cross-account access for AWS Glue crawlers to scan the data in each departmental S3 bucket to identify the schema and populate the catalog. Add the senior data analysts into the central account and apply highly detailed access controls in the Data Catalog and Amazon S3.

C. Set up an individual AWS account for the central data lake. Use AWS Lake Formation to catalog the cross-account locations. On each individual S3 bucket, modify the bucket policy to grant S3 permissions to the Lake Formation service-linked role. Use Lake Formation permissions to add fine-grained access controls to allow senior analysts to view specific tables and columns.

D. Set up an individual AWS account for the central data lake and configure a central S3 bucket. Use an AWS Lake Formation blueprint to move the data from the various buckets into the central S3 bucket. On each individual bucket, modify the bucket policy to grant S3 permissions to the Lake Formation service-linked role. Use Lake Formation permissions to add fine-grained access controls for both associate and senior analysts to view specific tables and columns.

Correct Answer: B

QUESTION 11

A financial institution is building an Amazon QuickSight business intelligence (BI) dashboard to show financial performance and analyze trends. The development team is using an Amazon Redshift database in the development environment and is having difficulty with validating the accuracy of the metrics calculation algorithm due to the lack of quality data. The Redshift production environment database is 500 TB and is in a different AWS account in the same AWS Region as the development environment account. The company needs to use up-to-date production environment data for development purposes.

Which solution MOST cost-effectively meets these requirements?

A. Setup data streaming with Amazon Kinesis Data Streams from the production environment Redshift database to replicate the data to the development environment Redshift database.

B. Create a Redshift datashare to share the production environment data with the development team.

C. Upload the data from Amazon Redshift to Amazon S3. Then load the data directly from Amazon S3 to the development environment Redshift cluster using the COPY command.

D. Create Redshift views that are configured to share all the data between the production and development clusters.

Correct Answer: D

QUESTION 12

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. Load all the data except for the columns containing sensitive data into a second table. Grant the appropriate users read-only permissions to the second table.

B. Load all the data into the new table and grant the auditing group permission to read from the table. Use the GRANT SQL command to allow read-only access to a subset of columns to the appropriate users.

C. 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.

D. 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.

Correct Answer: D