DAS-C01^{Q&As}

DAS-C01^{Q&As}

AWS Certified Data Analytics - Specialty (DAS-C01)

Pass Amazon DAS-C01 Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.passapply.com/das-c01.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Amazon Official Exam Center

⚙ **Instant Download** After Purchase

⚙ **100% Money Back** Guarantee

⚙ **365 Days** Free Update

⚙ **800,000+** Satisfied Customers

QUESTION 1

A real estate company maintains data about all properties listed in a market. The company receives data about new property listings from vendors who upload the data daily as compressed files into Amazon S3. The company\\'s leadership team wants to see the most up-to-date listings as soon as the data is uploaded to Amazon S3. The data analytics team must automate and orchestrate the data processing workflow of the listings to feed a dashboard. The team also must provide the ability to perform one-time queries and analytical reporting in a scalable manner.

Which solution meets these requirements MOST cost-effectively?

A. Use Amazon EMR for processing incoming data. Use AWS Step Functions for workflow orchestration. Use Apache Hive for one-time queries and analytical reporting. Bulk ingest the data in Amazon OpenSearch Service (Amazon Elasticsearch Service). Use OpenSearch Dashboards (Kibana) on Amazon OpenSearch Service (Amazon Elasticsearch Service) for the dashboard.

B. Use Amazon EMR for processing incoming data. Use AWS Step Functions for workflow orchestration. Use Amazon Athena for one-time queries and analytical reporting. Use Amazon QuickSight for the dashboard.

C. Use AWS Glue for processing incoming data. Use AWS Step Functions for workflow orchestration. Use Amazon Redshift Spectrum for one-time queries and analytical reporting. Use OpenSearch Dashboards (Kibana) on Amazon OpenSearch Service (Amazon Elasticsearch Service) for the dashboard.

D. Use AWS Glue for processing incoming data. Use AWS Lambda and S3 Event Notifications for workflow orchestration. Use Amazon Athena for one-time queries and analytical reporting. Use Amazon QuickSight for the dashboard.

Correct Answer: B

Reference: https://aws.amazon.com/blogs/compute/visualizing-aws-step-functions-workflows-from-the-amazon-athena-console/

QUESTION 2

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

Correct Answer: C

QUESTION 3

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

QUESTION 4

A company ingests a large set of sensor data in nested JSON format from different sources and stores it in an Amazon S3 bucket. The sensor data must be joined with performance data currently stored in an Amazon Redshift cluster.

A business analyst with basic SQL skills must build dashboards and analyze this data in Amazon QuickSight. A data engineer needs to build a solution to prepare the data for use by the business analyst. The data engineer does not know the structure of the JSON file. The company requires a solution with the least possible implementation effort.

Which combination of steps will create a solution that meets these requirements? (Choose three.)

A. Use an AWS Glue ETL job to convert the data into Apache Parquet format and write to Amazon S3.

B. Use an AWS Glue crawler to catalog the data.

C. Use an AWS Glue ETL job with the ApplyMapping class to un-nest the data and write to Amazon Redshift tables.

D. Use an AWS Glue ETL job with the Regionalize class to un-nest the data and write to Amazon Redshift tables.

E. Use QuickSight to create an Amazon Athena data source to read the Apache Parquet files in Amazon S3.

F. Use QuickSight to create an Amazon Redshift data source to read the native Amazon Redshift tables.

Correct Answer: ABD

QUESTION 5

A software company wants to use instrumentation data to detect and resolve errors to improve application recovery time. The company requires API usage anomalies, like error rate and response time spikes, to be detected in near-real time (NRT) The company also requires that data analysts have access to dashboards for log analysis in NRT.

Which solution meets these requirements\\'?

A. Use Amazon Kinesis Data Firehose as the data transport layer for logging data Use Amazon Kinesis Data Analytics to uncover the NRT API usage anomalies Use Kinesis Data Firehose to deliver log data to Amazon OpenSearch Service (Amazon Elasticsearch Service) for search, log analytics, and application monitoring Use OpenSearch Dashboards (Kibana) in Amazon OpenSearch Service (Amazon Elasticsearch Service) for the dashboards.

B. Use Amazon Kinesis Data Analytics as the data transport layer for logging data. Use Amazon Kinesis Data Streams to uncover NRT monitoring metrics. Use Amazon Kinesis Data Firehose to deliver log data to Amazon OpenSearch Service (Amazon Elasticsearch Service) for search, log analytics, and application monitoring Use Amazon QuickSight for the dashboards

C. Use Amazon Kinesis Data Analytics as the data transport layer for logging data and to uncover NRT monitoring metrics Use Amazon Kinesis Data Firehose to deliver log data to Amazon OpenSearch Service (Amazon Elasticsearch Service) for search, log analytics, and application monitoring Use OpenSearch Dashboards (Kibana) in Amazon OpenSearch Service (Amazon Elasticsearch Service) for the dashboards

D. Use Amazon Kinesis Data Firehose as the data transport layer for logging data Use Amazon Kinesis Data Analytics to uncover NRT monitoring metrics Use Amazon Kinesis Data Streams to deliver log data to Amazon OpenSearch Service (Amazon Elasticsearch Service) for search, log analytics, and application monitoring Use Amazon QuickSight for the dashboards.

Correct Answer: C

Reference: https://docs.aws.amazon.com/opensearch-service/latest/developerguide/integrations.html

QUESTION 6

A data analytics specialist has a 50 GB data file in .csv format and wants to perform a data transformation task. The data analytics specialist is using the Amazon Athena CREATE TABLE AS SELECT (CTAS) statement to perform the transformation. The resulting output will be used to query the data from Amazon Redshift Spectrum.

Which CTAS statement should the data analytics specialist use to provide the MOST efficient performance?

A. CREATE TABLE new_Table

WITH (

 format = \\'TEXTFILE\\',

 orc_compression = \\'SNAPPY\\')

AS SELECT *

FROM old_table;

B. CREATE TABLE new_Table

WITH (

 format = \\'TEXTFILE\\',

)

AS SELECT *

FROM old_table;

C. CREATE TABLE new_Table

WITH (

 format = \\'PARQUET\\',

 parquet_compression = \\'SNAPPY\\')

AS SELECT *

FROM old_table;

D. CREATE TABLE new_Table

WITH (

 format = JSON,

)

AS SELECT *

FROM old_table;

Correct Answer: C

Reference: https://docs.aws.amazon.com/athena/latest/ug/performance-tuning.html#performance-tuning-file-formats

---

QUESTION 7

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 8

A company has developed an Apache Hive script to batch process data stared in Amazon S3. The script needs to run once every day and store the output in Amazon S3. The company tested the script, and it completes within 30 minutes on a small local three-node cluster.

Which solution is the MOST cost-effective for scheduling and executing the script?

A. Create an AWS Lambda function to spin up an Amazon EMR cluster with a Hive execution step. Set KeepJobFlowAliveWhenNoSteps to false and disable the termination protection flag. Use Amazon CloudWatch Events to schedule the Lambda function to run daily.

B. Use the AWS Management Console to spin up an Amazon EMR cluster with Python Hue. Hive, and Apache Oozie. Set the termination protection flag to true and use Spot Instances for the core nodes of the cluster. Configure an Oozie workflow in the cluster to invoke the Hive script daily.

C. Create an AWS Glue job with the Hive script to perform the batch operation. Configure the job to run once a day using a time-based schedule.

D. Use AWS Lambda layers and load the Hive runtime to AWS Lambda and copy the Hive script. Schedule the Lambda function to run daily by creating a workflow using AWS Step Functions.

Correct Answer: C

QUESTION 9

A company wants to research user turnover by analyzing the past 3 months of user activities. With millions of users, 1.5 TB of uncompressed data is generated each day. A 30-node Amazon Redshift cluster with 2.56 TB of solid state drive (SSD) storage for each node is required to meet the query performance goals.

The company wants to run an additional analysis on a year\\'s worth of historical data to examine trends indicating which features are most popular. This analysis will be done once a week.

What is the MOST cost-effective solution?

A. Increase the size of the Amazon Redshift cluster to 120 nodes so it has enough storage capacity to hold 1 year of data. Then use Amazon Redshift for the additional analysis.

B. Keep the data from the last 90 days in Amazon Redshift. Move data older than 90 days to Amazon S3 and store it in Apache Parquet format partitioned by date. Then use Amazon Redshift Spectrum for the additional analysis.

C. Keep the data from the last 90 days in Amazon Redshift. Move data older than 90 days to Amazon S3 and store it in Apache Parquet format partitioned by date. Then provision a persistent Amazon EMR cluster and use Apache Presto for the additional analysis.

D. Resize the cluster node type to the dense storage node type (DS2) for an additional 16 TB storage capacity on each individual node in the Amazon Redshift cluster. Then use Amazon Redshift for the additional analysis.

Correct Answer: B

QUESTION 10

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:

1.

Choose the catalog table name and do not rely on the catalog table naming algorithm.

2.

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. Create an AWS Glue crawler and specify the table as the source.

D. 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. Migrate the Hive catalog to the Data Catalog.

Correct Answer: C

Reference: https://docs.aws.amazon.com/glue/latest/dg/tables-described.html

---

QUESTION 11

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

Which solution organizes the images and metadata to drive insights while meeting the requirements?

A. For each image, use object tags to add the metadata. Use Amazon S3 Select to retrieve the files based on the applicant name and application date.

B. Index the metadata and the Amazon S3 location of the image file in Amazon OpenSearch Service (Amazon Elasticsearch Service). Allow the data analysts to use OpenSearch Dashboards (Kibana) to submit queries to the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster.

C. Store the metadata and the Amazon S3 location of the image file in an Amazon Redshift table. Allow the data analysts to run ad-hoc queries on the table.

D. Store the metadata and the Amazon S3 location of the image files in an Apache Parquet file in Amazon S3, and define a table in the AWS Glue Data Catalog. Allow data analysts to use Amazon Athena to submit custom queries.

Correct Answer: A

QUESTION 12

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 13

A company\'s data analytics specialist must build a solution to implement quality checks on a dataset before the company uses the data in a sales report. The dataset is stored in an Amazon S3 bucket and is in CSV format.

The data quality checks must include identification of duplicate rows, removal of duplicate rows, and validation of date formats. The solution must run daily and must produce output data in Apache Parquet format in Amazon S3.

Which solution will meet these requirements with the LEAST development effort?

A. Create an AWS Glue ETL job that includes transformation steps to implement data quality checks. Configure the job to write to Amazon S3. Create a schedule-based job within an AWS Glue workflow to run the job daily.

B. Create an AWS Glue DataBrew job that includes data quality recipe steps to implement data quality checks. Configure the job to write to Amazon S3. Create a schedule within the DataBrew job to run the job daily.

C. Create an Amazon EMR cluster. Use an Apache Spark ETL job that includes data processing steps to implement data quality checks. Configure the job to write to Amazon S3. Create an Apache Oozie workflow to run the job daily.

D. Create an AWS Lambda function. Use custom code to implement data quality checks and to write to Amazon S3. Create an Amazon EventBridge rule to run the Lambda function daily.

Correct Answer: C

QUESTION 14

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 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. Decrypt the sensitive data using the same KMS key. Save the output to a restricted S3 bucket for the finance team. Create a finance table in Amazon Redshift that is accessible to the finance team only. Use the COPY command to load the data from Amazon S3 to the finance table.

B. Create an AWS Lambda function to process the DynamoDB stream. Save the output to a restricted S3 bucket for the finance team. Create a finance table in Amazon Redshift that is accessible to the finance team only. 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.

C. 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. In Hive, select the data from DynamoDB and then insert the output to the finance table in Amazon Redshift.

D. Create an Amazon EMR cluster. Create Apache Hive tables that reference the data stored in DynamoDB. Insert the output to the restricted Amazon S3 bucket for the finance team. 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.

Correct Answer: B