DAS-C01$^{\text{Q\&As}}$

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.leads4pass.com/das-c01.html

100% Passing Guarantee
100% Money Back Assurance

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

QUESTION 1

A company uses Amazon EC2 instances to receive files from external vendors throughout each day. At the end of each day, the EC2 instances combine the files into a single file, perform gzip compression, and upload the single file to an

Amazon S3 bucket. The total size of all the files is approximately 100 GB each day.

When the files are uploaded to Amazon S3, an AWS Batch job runs a COPY command to load the files into an Amazon Redshift cluster.

Which solution will MOST accelerate the COPY process?

A. Upload the individual files to Amazon S3. Run the COPY command as soon as the files become available.

B. Split the files so that the number of files is equal to a multiple of the number of slices in the Redshift cluster. Compress and upload the files to Amazon S3. Run the COPY command on the files.

C. Split the files so that each file uses 50% of the free storage on each compute node in the Redshift cluster. Compress and upload the files to Amazon S3. Run the COPY command on the files.

D. Apply sharding by breaking up the files so that the DISTKEY columns with the same values go to the same file. Compress and upload the sharded files to Amazon S3. Run the COPY command on the files.

Correct Answer: B

QUESTION 2

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 3

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.

Correct Answer: B

Reference: https://aws.amazon.com/blogs/big-data/harmonize-query-and-visualize-data-from-various-providers-using-aws-glue-amazon-athena-and-amazon-quicksight/

QUESTION 4

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 5

An event ticketing website has a data lake on Amazon S3 and a data warehouse on Amazon Redshift. Two datasets exist: events data and sales data. Each dataset has millions of records.

The entire events dataset is frequently accessed and is stored in Amazon Redshift. However, only the last 6 months of sales data is frequently accessed and is stored in Amazon Redshift. The rest of the sales data is available only in

Amazon

S3.

A data analytics specialist must create a report that shows the total revenue that each event has generated in the last 12 months. The report will be accessed thousands of times each week.

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

A. Create an AWS Glue job to access sales data that is older than 6 months from Amazon S3 and to access event and sales data from Amazon Redshift. Load the results into a new table in Amazon Redshift.

B. Create a stored procedure to copy sales data that is older than 6 months and newer than 12 months from Amazon S3 to Amazon Redshift. Create a materialized view with the autorefresh option.

C. Create an AWS Lambda function to copy sales data that is older than 6 months and newer than 12 months to an Amazon Kinesis Data Firehose delivery stream. Specify Amazon Redshift as the destination of the delivery stream. Create a materialized view with the autorefresh option.

D. Create a materialized view in Amazon Redshift with the autorefresh option. Use Amazon Redshift Spectrum to include sales data that is older than 6 months.

Correct Answer: A

QUESTION 6

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 7

A financial services firm is processing a stream of real-time data from an application by using Apache Kafka and Kafka MirrorMaker. These tools are running on premises to Amazon Managed Streaming for Apache Kafka (Amazon MSK) in the us-east-1 Region. An Apache Flink consumer running on Amazon EMR enriches the data in real time and transfers the output files to an Amazon S3 bucket. The company wants to ensure that the streaming application is highly available across AWS Regions with an RTO of less than 2 minutes.

Which solution meets these requirements?

A. Launch another Amazon MSK and Apache Flink cluster in the us-west-1 Region that is the same size as the original cluster in the us-east-1 Region Simultaneously publish and process the data in both Regions. In the event of a disaster that impacts one of the Regions, switch to the other Region.

B. Set up Cross-Region Replication from the Amazon s3 bucket in the us-east-1 Region to the us-west-1 Region. In the event of a disaster, immediately create Amazon MSK and Apache Flink clusters in the us-west-1 Region and start publishing data to this Region.

C. Add an AWS Lambda function in the us-east-1 Region to read from Amazon MSK and write to a global Amazon DynamoDB table in on-demand capacity mode. Export the data from DynamoDB to Amazon S3 in the us-west-1 Region. In the event of a disaster that impacts the us-east-1 Region, immediately create Amazon MSK and Apache Flink clusters in the us-west-1 Region and start publishing data to this Region.

D. Set up Cross-Region Replication from the Amazon S3 bucket in the us-east-1 Region to the us-west-1 Region. In the event of a disaster, immediately create Amazon MSK and Apache Flink clusters in the us-west-1 Region and start publishing data to this Region. Store 7 days of data in on-premises Kafka clusters and recover the data missed during the recovery time from the on-premises cluster.

Correct Answer: C

QUESTION 8

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

Correct Answer: BCF

QUESTION 9

A startup company runs its data processing and machine learning workloads on Amazon EMR. To increase productivity, the company granted permissions to engineers and analysts to create EMR clusters. A recent security review showed

that some EMR clusters had ports open to the public internet.

A data analytics specialist must make changes so that any new EMR clusters have public access blocked when they are created.

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

A. Create an AWS Lambda function that checks whether the Amazon EMR security group is open to the public internet. Invoke a Lambda function when an EMR cluster is created. Integrate the function with Amazon Simple Notification Service (Amazon SNS) to send notification email.

B. Turn on the block public access feature by using the Amazon EMR console.

C. Use AWS Config to track the security on the EMR cluster. Use Amazon EventBridge to send notifications when an open cluster is detected. Create an AWS Lambda function that the notification invokes to block public access.

D. Update security groups to remove inbound traffic from IPv4 0.0.0.0/0 or IPv6 ::/0.

Correct Answer: A

QUESTION 10

A financial services company is building a data lake solution on Amazon S3. The company plans to use analytics offerings from AWS to meet user needs for one-time querying and business intelligence reports. A portion of the columns will contain personally identifiable information (PII) Only authorized users should be able to see plaintext PII data.

What is the MOST operationally efficient solution that meets these requirements?

A. Define a bucket policy for each S3 bucket of the data lake to allow access to users who have authorization to see PII data. Catalog the data by using AWS Glue. Create two IAM roles. Attach a permissions policy with access to PII columns to one role. Attach a policy without these permissions to the other role.

B. Register the S3 locations with AWS Lake Formation. Create two IAM roles. Use Lake Formation data permissions to grant Select permissions to all of the columns for one role. Grant Select permissions to only columns that contain non-PII data for the other role.

C. Register the S3 locations with AWS Lake Formation. Create an AWS Glue job to create an ETL workflow that removes the PII columns from the data and creates a separate copy of the data in another data lake S3 bucket. Register the new S3 locations with Lake Formation. Grant users the permissions to each data lake data based on whether the users are authorized to see PII data.

D. Register the S3 locations with AWS Lake Formation. Create two IAM roles. Attach a permissions policy with access to PII columns to one role. Attach a policy without these permissions to the other role. For each downstream analytics service, use its native security functionality and the IAM roles to secure the PII data.

Correct Answer: C

Reference: https://docs.aws.amazon.com/lake-formation/latest/dg/lake-formation-dg.pdf

QUESTION 11

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.

Correct Answer: B

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

QUESTION 12

A US-based sneaker retail company launched its global website. All the transaction data is stored in Amazon RDS and curated historic transaction data is stored in Amazon Redshift in the us-east-1 Region. The business intelligence (BI) team wants to enhance the user experience by providing a dashboard for sneaker trends.

The BI team decides to use Amazon QuickSight to render the website dashboards. During development, a team in Japan provisioned Amazon QuickSight in ap-northeast-1. The team is having difficulty connecting Amazon QuickSight from ap-northeast-1 to Amazon Redshift in us-east-1.

Which solution will solve this issue and meet the requirements?

A. In the Amazon Redshift console, choose to configure cross-Region snapshots and set the destination Region as ap-northeast-1. Restore the Amazon Redshift Cluster from the snapshot and connect to Amazon QuickSight launched in apnortheast-1.

B. Create a VPC endpoint from the Amazon QuickSight VPC to the Amazon Redshift VPC so Amazon QuickSight can access data from Amazon Redshift.

C. Create an Amazon Redshift endpoint connection string with Region information in the string and use this connection string in Amazon QuickSight to connect to Amazon Redshift.

D. Create a new security group for Amazon Redshift in us-east-1 with an inbound rule authorizing access from the appropriate IP address range for the Amazon QuickSight servers in ap-northeast-1.

Correct Answer: D

QUESTION 13

An online gaming company is using an Amazon Kinesis Data Analytics SQL application with a Kinesis data stream as its source. The source sends three non-null fields to the application: player_id, score, and us_5_digit_zip_code.

A data analyst has a .csv mapping file that maps a small number of us_5_digit_zip_code values to a territory code. The data analyst needs to include the territory code, if one exists, as an additional output of the Kinesis Data Analytics application.

How should the data analyst meet this requirement while minimizing costs?

A. Store the contents of the mapping file in an Amazon DynamoDB table. Preprocess the records as they arrive in the Kinesis Data Analytics application with an AWS Lambda function that fetches the mapping and supplements each record to include the territory code, if one exists. Change the SQL query in the application to include the new field in the SELECT statement.

B. Store the mapping file in an Amazon S3 bucket and configure the reference data column headers for the .csv file in the Kinesis Data Analytics application. Change the SQL query in the application to include a join to the file\\'s S3 Amazon Resource Name (ARN), and add the territory code field to the SELECT columns.

C. Store the mapping file in an Amazon S3 bucket and configure it as a reference data source for the Kinesis Data Analytics application. Change the SQL query in the application to include a join to the reference table and add the territory code field to the SELECT columns.

D. Store the contents of the mapping file in an Amazon DynamoDB table. Change the Kinesis Data Analytics application to send its output to an AWS Lambda function that fetches the mapping and supplements each record to include the territory code, if one exists. Forward the record from the Lambda function to the original application destination.

Correct Answer: C

QUESTION 14

An energy company collects voltage data in real time from sensors that are attached to buildings. The company wants to receive notifications when a sequence of two voltage drops is detected within 10 minutes of a sudden voltage increase at the same building. All notifications must be delivered as quickly as possible. The system must be highly available. The company needs a solution that will automatically scale when this monitoring feature is implemented in other cities. The notification system is subscribed to an Amazon Simple Notification Service (Amazon SNS) topic for remediation.

Which solution will meet these requirements?

A. Create an Amazon Managed Streaming for Apache Kafka cluster to ingest the data. Use an Apache Spark Streaming with Apache Kafka consumer API in an automatically scaled Amazon EMR cluster to process the incoming data. Use the Spark Streaming application to detect the known event sequence and send the SNS message.

B. Create a REST-based web service by using Amazon API Gateway in front of an AWS Lambda function. Create an Amazon RDS for PostgreSQL database with sufficient Provisioned IOPS to meet current demand. Configure the Lambda function to store incoming events in the RDS for PostgreSQL database, query the latest data to detect the known event sequence, and send the SNS message.

C. Create an Amazon Kinesis Data Firehose delivery stream to capture the incoming sensor data. Use an AWS Lambda transformation function to detect the known event sequence and send the SNS message.

D. Create an Amazon Kinesis data stream to capture the incoming sensor data. Create another stream for notifications. Set up AWS Application Auto Scaling on both streams. Create an Amazon Kinesis Data Analytics for Java application to detect the known event sequence, and add a message to the message stream Configure an AWS Lambda function to poll the message stream and publish to the SNS topic.

Correct Answer: D

Reference: https://aws.amazon.com/kinesis/data-streams/faqs/

QUESTION 15

A media content company has a streaming playback application. The company wants to collect and analyze the data to provide near-real-time feedback on playback issues. The company needs to consume this data and return results within 30 seconds according to the service-level agreement (SLA). The company needs the consumer to identify playback issues, such as quality during a specified timeframe. The data will be emitted as JSON and may change schemas over time.

Which solution will allow the company to collect data for processing while meeting these requirements?

A. Send the data to Amazon Kinesis Data Firehose with delivery to Amazon S3. Configure an S3 event trigger an AWS Lambda function to process the data. The Lambda function will consume the data and process it to identify potential playback issues. Persist the raw data to Amazon S3.

B. Send the data to Amazon Managed Streaming for Kafka and configure an Amazon Kinesis Analytics for Java application as the consumer. The application will consume the data and process it to identify potential playback issues. Persist the raw data to Amazon DynamoDB.

C. Send the data to Amazon Kinesis Data Firehose with delivery to Amazon S3. Configure Amazon S3 to trigger an event for AWS Lambda to process. The Lambda function will consume the data and process it to identify potential playback issues. Persist the raw data to Amazon DynamoDB.

D. Send the data to Amazon Kinesis Data Streams and configure an Amazon Kinesis Analytics for Java application as the consumer. The application will consume the data and process it to identify potential playback issues. Persist the raw data to Amazon S3.

Correct Answer: D

https://aws.amazon.com/blogs/aws/new-amazon-kinesis-data-analytics-for- java/

DAS-C01 VCE Dumps DAS-C01 Study Guide DAS-C01 Exam Questions