Free Questions for DAS-C01

Shared by Colon on 20-10-2022

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

A marketing company collects clickstream data The company sends the data to Amazon Kinesis Data Firehose and stores the data in Amazon S3 The company wants to build a series of dashboards that will be used by hundreds of users across different departments The company will use Amazon QuickSight to develop these dashboards The company has limited resources and wants a solution that could scale and provide daily updates about clickstream activity

Which combination of options will provide the MOST cost-effective solution? (Select TWO)

Options:

A- Use Amazon Redshift to store and query the clickstream data
B- Use QuickSight with a direct SQL query
C- Use Amazon Athena to query the clickstream data in Amazon S3
D- Use S3 analytics to query the clickstream data
E- Use the QuickSight SPICE engine with a daily refresh

Answer:

B, D

Question 2

Question Type: MultipleChoice

A hospital is building a research data lake to ingest data from electronic health records (EHR) systems from multiple hospitals and clinics. The EHR systems are independent of each other and do not have a common patient identifier. The data engineering team is not experienced in machine learning (ML) and has been asked to generate a unique patient identifier for the ingested records.

Which solution will accomplish this task?

Options:

A- An AWS Glue ETL job with the FindMatches transform
B- Amazon Kendra
C- Amazon SageMaker Ground Truth
D- An AWS Glue ETL job with the ResolveChoice transform

Answer:

A

Explanation:

Matching Records with AWS Lake Formation FindMatches

Question 3

Question Type: MultipleChoice

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?

Options:

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.

Question 4

Question Type: MultipleChoice

A large ecommerce company uses Amazon DynamoDB with provisioned read capacity and auto scaled write capacity to store its product catalog. The company uses Apache HiveQL statements on an Amazon EMR cluster to query the DynamoDB table. After the company announced a sale on all of its products, wait times for each query have increased. The data analyst has determined that the longer wait times are being caused by throttling when querying the table.

Which solution will solve this issue?

Options:

A- Increase the size of the EMR nodes that are provisioned.
B- Increase the number of EMR nodes that are in the cluster.
C- Increase the DynamoDB table's provisioned write throughput.
D- Increase the DynamoDB table's provisioned read throughput.

Answer:

D

Question 5

Question Type: MultipleChoice

A manufacturing company has many loT devices in different facilities across the world The company is using Amazon Kinesis Data Streams to collect the data from the devices

The company's operations team has started to observe many WnteThroughputExceeded exceptions The operations team determines that the reason is the number of records that are being written to certain shards The data contains device ID capture date measurement type, measurement value and facility ID The facility ID is used as the partition key

Which action will resolve this issue?

Options:

A- Change the partition key from facility ID to a randomly generated key
B- Increase the number of shards
C- Archive the data on the producers' side
D- Change the partition key from facility ID to capture date

Answer:

B

Question 6

Question Type: MultipleChoice

A company has several Amazon EC2 instances sitting behind an Application Load Balancer (ALB) The company wants its IT Infrastructure team to analyze the IP addresses coming into the company's ALB The ALB is configured to store access logs in Amazon S3 The access logs create about 1 TB of data each day, and access to the data will be infrequent The company needs a solution that is scalable, cost-effective and has minimal maintenance requirements

Which solution meets these requirements?

Options:

A- Copy the data into Amazon Redshift and query the data
B- Use Amazon EMR and Apache Hive to query the S3 data
C- Use Amazon Athena to query the S3 data
D- Use Amazon Redshift Spectrum to query the S3 data

Answer:

D

Question 7

Question Type: MultipleChoice

A company receives datasets from partners at various frequencies. The datasets include baseline data and incremental data. The company needs to merge and store all the datasets without reprocessing the data.

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

Options:

A- Use an AWS Glue job with a temporary table to process the datasets. Store the data in an Amazon RDS table.
B- Use an Apache Spark job in an Amazon EMR cluster to process the datasets. Store the data in EMR File System (EMRFS).
C- Use an AWS Glue job with job bookmarks enabled to process the datasets. Store the data in Amazon S3.
D- Use an AWS Lambda function to process the datasets. Store the data in Amazon S3.

Answer:

C

Explanation:

AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy to prepare and load data for analytics1. It can process datasets from various sources and formats, such as JDBC, Amazon S3, Amazon RDS, etc.

AWS Glue job bookmarks are a feature that helps AWS Glue track data that has already been processed during a previous run of an ETL job.This can prevent the reprocessing of old data and enable the processing of new data when rerunning on a scheduled interval2. Job bookmarks can handle both baseline data and incremental data from different sources.

Amazon S3 is a highly scalable, durable, and secure object storage service that can store any amount and type of data3. It can be used as a data lake to store the merged and processed datasets from AWS Glue. It can also integrate with other AWS services, such as Amazon Athena, Amazon Redshift Spectrum, Amazon EMR, etc., for further analysis and processing.

Question 8

Question Type: MultipleChoice

A company is reading data from various customer databases that run on Amazon RDS. The databases contain many inconsistent fields For example, a customer record field that is place_id in one database is location_id in another database. The company wants to link customer records across different databases, even when many customer record fields do not match exactly

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

Options:

A- Create an Amazon EMR cluster to process and analyze data in the databases Connect to the Apache Zeppelin notebook, and use the FindMatches transform to find duplicate records in the data.
B- Create an AWS Glue crawler to crawl the databases. Use the FindMatches transform to find duplicate records in the data Evaluate and tune the transform by evaluating performance and results of finding matches
C- Create an AWS Glue crawler to crawl the data in the databases Use Amazon SageMaker to construct Apache Spark ML pipelines to find duplicate records in the data
D- Create an Amazon EMR cluster to process and analyze data in the databases. Connect to the Apache Zeppelin notebook, and use Apache Spark ML to find duplicate records in the data. Evaluate and tune the model by evaluating performance and results of finding duplicates

Answer:

B

Question 9

Question Type: MultipleChoice

A global pharmaceutical company receives test results for new drugs from various testing facilities worldwide. The results are sent in millions of 1 KB-sized JSON objects to an Amazon S3 bucket owned by the company. The data engineering team needs to process those files, convert them into Apache Parquet format, and load them into Amazon Redshift for data analysts to perform dashboard reporting. The engineering team uses AWS Glue to process the objects, AWS Step Functions for process orchestration, and Amazon CloudWatch for job scheduling.

More testing facilities were recently added, and the time to process files is increasing.

What will MOST efficiently decrease the data processing time?

Options:

A- Use AWS Lambda to group the small files into larger files. Write the files back to Amazon S3. Process the files using AWS Glue and load them into Amazon Redshift tables.
B- Use the AWS Glue dynamic frame file grouping option while ingesting the raw input files. Process the files and load them into Amazon Redshift tables.
C- Use the Amazon Redshift COPY command to move the files from Amazon S3 into Amazon Redshift tables directly. Process the files in Amazon Redshift.
D- Use Amazon EMR instead of AWS Glue to group the small input files. Process the files in

Amazon EMR and load them into Amazon Redshift tables.

Answer:

A