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.certbus.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 company uses Amazon OpenSearch Service (Amazon Elasticsearch Service) to store and analyze its website clickstream data. The company ingests 1 TB of data daily using Amazon Kinesis Data Firehose and stores one day\\'s worth of data in an Amazon ES cluster.

The company has very slow query performance on the Amazon ES index and occasionally sees errors from Kinesis Data Firehose when attempting to write to the index. The Amazon ES cluster has 10 nodes running a single index and 3 dedicated master nodes. Each data node has 1.5 TB of Amazon EBS storage attached and the cluster is configured with 1,000 shards. Occasionally, JVMMemoryPressure errors are found in the cluster logs.

Which solution will improve the performance of Amazon ES?

A. Increase the memory of the Amazon ES master nodes.

B. Decrease the number of Amazon ES data nodes.

C. Decrease the number of Amazon ES shards for the index.

D. Increase the number of Amazon ES shards for the index.

Correct Answer: C

QUESTION 2

A logistics company has an application that generates status data for order shipments. The company uses Amazon Kinesis Data Firehose to ingest data into Amazon S3 in near-real time. The newly ingested data must be merged with historical data before it can be used for business analytics. The company wants a solution that supports data inserts, updates, and deletes with minimal time delays.

Which solution meets these requirements with the LEAST amount of operational effort?

A. Use AWS Glue Spark jobs to populate an Apache ORC table.

B. Use AWS Glue Spark jobs with an Apache Hudi connector to populate an Apache Hudi table.

C. Use Amazon EMR with an Apache Spark script to populate an Apache ORC table.

D. Use Amazon EMR with an Apache Spark script to populate an Apache Hudi table.

Correct Answer: C

QUESTION 3

A manufacturing company is storing data from its operational systems in Amazon S3. The company\\'s business analysts need to perform one-time queries of the data in Amazon S3 with Amazon Athena. The company needs to access the Athena network from the on-premises network by using a JDBC connection. The company has created a VPC Security policies mandate that requests to AWS services cannot traverse the Internet.

Which combination of steps should a data analytics specialist take to meet these requirements? (Choose two.)

A. Establish an AWS Direct Connect connection between the on-premises network and the VPC.

B. Configure the JDBC connection to connect to Athena through Amazon API Gateway.

C. Configure the JDBC connection to use a gateway VPC endpoint for Amazon S3.

D. Configure the JDBC connection to use an interface VPC endpoint for Athena.

E. Deploy Athena within a private subnet.

Correct Answer: AE

AWS Direct Connect makes it easy to establish a dedicated connection from an on-premises network to one or more VPCs in the same region.

Reference: https://docs.aws.amazon.com/whitepapers/latest/aws-vpc-connectivity-options/aws-direct-connect.html https://stackoverflow.com/questions/68798311/aws-athena-connect-from-lambda

---

QUESTION 4

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 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 substantially over time, and the query latency also has increased.

Which solutions could the company implement to improve query performance? (Choose two.)

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

B. Use Athena to extract the data and store it in Apache Parquet format on a daily basis. Query the extracted data.

C. Run a daily AWS Glue ETL job to convert the data files to Apache Parquet and to partition the converted files. Create a periodic AWS Glue crawler to automatically crawl the partitioned data on a daily basis.

D. Run a daily AWS Glue ETL job to compress the data files by using the .gzip format. Query the compressed data.

E. Run a daily AWS Glue ETL job to compress the data files by using the .lzo format. Query the compressed data.

Correct Answer: BC

Reference: https://www.upsolver.com/blog/apache-parquet-why-use https://aws.amazon.com/blogs/big-data/work-with-partitioned-data-in-aws-glue/

QUESTION 5

A company is creating a data lake by using AWS Lake Formation. The data that will be stored in the data lake contains sensitive customer information and must be encrypted at rest using an AWS Key Management Service (AWS KMS) customer managed key to meet regulatory requirements.

How can the company store the data in the data lake to meet these requirements?

A. Store the data in an encrypted Amazon Elastic Block Store (Amazon EBS) volume. Register the Amazon EBS volume with Lake Formation.

B. Store the data in an Amazon S3 bucket by using server-side encryption with AWS KMS (SSE-KMS). Register the S3 location with Lake Formation.

C. Encrypt the data on the client side and store the encrypted data in an Amazon S3 bucket. Register the S3 location with Lake Formation.

D. Store the data in an Amazon S3 Glacier Flexible Retrieval vault bucket. Register the S3 Glacier Flexible Retrieval vault with Lake Formation.

Correct Answer: D

QUESTION 6

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 7

A data analyst at a fast-growing retail company needs to store data coming in from several dozen marketing campaigns. Each source will write its output to a CSV file that is stored in Amazon S3. The data will later be analyzed by individual campaign managers using Amazon Athena to roughly track the number of daily unique visits to their specific campaign websites over time. The company wants to minimize the cost of data analysis.

Which combination of actions would lead to the MOST efficient one-time analysis of the data? (Choose two.)

A. Use an AWS Glue job to convert all files to Apache ORC format. Use the COUNT(DISTINCT column) function to obtain a count of unique visitors.

B. Create one S3 bucket for all the data. Partition the data by date.

C. Create a separate S3 bucket for each campaign. Partition the data by date.

D. Create a separate S3 bucket for each month. Partition the data by campaign.

E. Use an AWS Glue job to convert all files to Apache Parquet format. Use the approx_distinct() function to obtain a count of unique visitors.

Correct Answer: AE

QUESTION 8

A company is using an AWS Lambda function to run Amazon Athena queries against a cross-account AWS Glue Data Catalog. A query returns the following error:

HIVE_METASTORE_ERROR

The error message states that the response payload size exceeds the maximum allowed size. The queried table is already partitioned, and the data is stored in an Amazon S3 bucket in the Apache Hive partition format.

Which solution will resolve this error?

A. Modify the Lambda function to upload the query response payload as an object into the S3 bucket. Include an S3 object presigned URL as the payload in the Lambda function response.

B. Run the MSCK REPAIR TABLE command on the queried table.

C. Create a separate folder in the S3 bucket. Move the data files that need to be queried into that folder. Create an AWS Glue crawler that points to the folder instead of the S3 bucket.

D. Check the schema of the queried table for any characters that Athena does not support. Replace any unsupported characters with characters that Athena supports.

Correct Answer: C

Reference: https://docs.aws.amazon.com/athena/latest/ug/tables-location-format.html

QUESTION 9

A transport company wants to track vehicular movements by capturing geolocation records. The records are 10 B in size and up to 10,000 records are captured each second. Data transmission delays of a few minutes are acceptable, considering unreliable network conditions. The transport company decided to use Amazon Kinesis Data Streams to ingest the data. The company is looking for a reliable mechanism to send data to Kinesis Data Streams while maximizing the throughput efficiency of the Kinesis shards.

Which solution will meet the company\\'s requirements?

A. Kinesis Agent

B. Kinesis Producer Library (KPL)

C. Kinesis Data Firehose

D. Kinesis SDK

Correct Answer: B

Reference: https://docs.aws.amazon.com/streams/latest/dev/developing-producers-with-sdk.htmls

QUESTION 10

A hospital uses an electronic health records (EHR) system to collect two types of data

1.

Patient information, which includes a patient\\'s name and address

2.

Diagnostic tests conducted and the results of these tests Patient information is expected to change periodically Existing diagnostic test data never changes and only new records are added. The hospital runs an Amazon Redshift cluster with four dc2.large nodes and wants to automate the ingestion of the patient information and diagnostic test data into respective Amazon Redshift tables for analysis The EHR system exports data

as CSV files to an Amazon S3 bucket on a daily basis Two sets of CSV files are generated One set of files is for patient information with updates, deletes, and inserts The other set of files is for new diagnostic test data only. What is the MOST cost-effective solution to meet these requirements?

A. Use Amazon EMR with Apache Hudi. Run daily ETL jobs using Apache Spark and the Amazon Redshift JDBC driver

B. Use an AWS Glue crawler to catalog the data in Amazon S3 Use Amazon Redshift Spectrum to perform scheduled queries of the data in Amazon S3 and ingest the data into the patient information table and the diagnostic tests table.

C. Use an AWS Lambda function to run a COPY command that appends new diagnostic test data to the diagnostic tests table Run another COPY command to load the patient information data into the staging tables Use a stored procedure to handle create update, and delete operations for the patient information table

D. Use AWS Database Migration Service (AWS DMS) to collect and process change data capture (CDC) records Use the COPY command to load patient information data into the staging tables. Use a stored procedure to handle create, update and delete operations for the patient information table

Correct Answer: B

Reference: https://docs.aws.amazon.com/prescriptive-guidance/latest/serverless-etl-aws-glue/aws-glue-data-catalog.html https://docs.aws.amazon.com/redshift/latest/mgmt/query-editor-schedule-query.html

QUESTION 11

A data analytics specialist is building a solution to securely collect and store data from multiple applications for analytics. The solution must store the data across multiple Availability Zones less than 2 minutes after collecting the data. The solution also must provide a secure way to authenticate and authorize the source application users.

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

A. Authenticate by using an Amazon Cognito user pool that is authorized to write to an Amazon API Gateway REST API. Configure the API as a proxy for an Amazon Kinesis Data Firehose delivery stream that has an Amazon S3 destination.

B. Write data to an Amazon S3 bucket by using the AWS SDK. Configure a bucket policy that limits writes to specific IAM roles.

C. Create an IAM access key that is authorized to write to an Amazon Kinesis Data Firehose delivery stream that has an Amazon S3 destination. Embed the access key into the applications.

D. Call a REST-based service such as Amazon API Gateway that uses a custom authentication service to store data on a Kubernetes cluster.

Correct Answer: A

QUESTION 12

A company hosts a large data warehouse on Amazon Redshift. A business intelligence (BI) team requires access to tables in schemas A and B. However, the BI team must not have access to tables in schema C.

Members of the BI team connect to the Redshift cluster through a client that uses a JDBC connector. A data analytics specialist needs to set up access for these users.

Which combination of steps will meet these requirements? (Choose two.)

A. Create an IAM user for each BI team member who requires access. Create an IAM group for these users.

B. Create a database user for each BI team member who requires access. Create a database user group for these users.

C. Create an IAM policy that grants read and write permissions for schemas A and B to the BI IAM group and denies read and write permissions for schema C to the BI IAM group. Attach the policy to the BI IAM group.

D. Use the GRANT command to grant access to the BI database user group for schemas A and B. Use the REVOKE command to block access for the BI database user group for schema C.

E. Specify the WITH MANAGED ACCESS parameter during the creation of schema C.

Correct Answer: CD

QUESTION 13

A company owns manufacturing facilities with Internet of Things (IoT) devices installed to monitor safety data. The company has configured an Amazon Kinesis data stream as a source for an Amazon Kinesis Data Firehose delivery stream, which outputs data to Amazon S3. The company\\'s operations team wants to gain insights from the IoT data to monitor data quality at ingestion. The insights need to be derived in near-real time, and the output must be logged to Amazon DynamoDB for further analysis.

Which solution meets these requirements?

A. Create an Amazon Kinesis Data Analytics for SQL application to read and analyze the data in the data stream. Add an output configuration so that everything written to an in-application stream persists in a DynamoDB table.

B. Create an Amazon Kinesis Data Analytics for SQL application to read and analyze the data in the data stream. Add an output configuration so that everything written to an in-application stream is passed to an AWS Lambda function that saves the data in a DynamoDB table as persistent data.

C. Configure an AWS Lambda function to analyze the data in the Kinesis Data Firehose delivery stream. Save the output to a DynamoDB table.

D. Configure an AWS Lambda function to analyze the data in the Kinesis Data Firehose delivery stream and save the output to an S3 bucket. Schedule an AWS Glue job to periodically copy the data from the bucket to a DynamoDB table.

Correct Answer: B

QUESTION 14

A company wants to improve user satisfaction for its smart home system by adding more features to its recommendation engine. Each sensor asynchronously pushes its nested JSON data into Amazon Kinesis Data Streams using the Kinesis Producer Library (KPL) in Java. Statistics from a set of failed sensors showed that, when a sensor is malfunctioning, its recorded data is not always sent to the cloud.

The company needs a solution that offers near-real-time analytics on the data from the most updated sensors.

Which solution enables the company to meet these requirements?

A. Set the RecordMaxBufferedTime property of the KPL to "-1" to disable the buffering on the sensor side. Use Kinesis Data Analytics to enrich the data based on a company-developed anomaly detection SQL script. Push the enriched data to a fleet of Kinesis data streams and enable the data transformation feature to flatten the JSON file. Instantiate a dense storage Amazon Redshift cluster and use it as the destination for the Kinesis Data Firehose delivery stream.

B. Update the sensors code to use the PutRecord/PutRecords call from the Kinesis Data Streams API with the AWS SDK for Java. Use Kinesis Data Analytics to enrich the data based on a company-developed anomaly detection SQL script. Direct the output of KDA application to a Kinesis Data Firehose delivery stream, enable the data transformation feature to flatten the JSON file, and set the Kinesis Data Firehose destination to an Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster.

C. Set the RecordMaxBufferedTime property of the KPL to "0" to disable the buffering on the sensor side. Connect for each stream a dedicated Kinesis Data Firehose delivery stream and enable the data transformation feature to flatten the JSON file before sending it to an Amazon S3 bucket. Load the S3 data into an Amazon Redshift cluster.

D. Update the sensors code to use the PutRecord/PutRecords call from the Kinesis Data Streams API with the AWS SDK for Java. Use AWS Glue to fetch and process data from the stream using the Kinesis Client Library (KCL). Instantiate an Amazon Elasticsearch Service cluster and use AWS Lambda to directly push data into it.

Correct Answer: B

QUESTION 15

A company stores revenue data in Amazon Redshift. A data analyst needs to create a dashboard so that the company\\'s sales team can visualize historical revenue and accurately forecast revenue for the upcoming months. Which solution will MOST cost-effectively meet these requirements?

A. Create an Amazon QuickSight analysis by using the data in Amazon Redshift. Add a custom field in QuickSight that applies a linear regression function to the data. Publish the analysis as a dashboard.

B. Create a JavaScript dashboard by using D3.js charts and the data in Amazon Redshift. Export the data to Amazon SageMaker. Run a Python script to run a regression model to forecast revenue. Import the data back into Amazon Redshift. Add the new forecast information to the dashboard.

C. Create an Amazon QuickSight analysis by using the data in Amazon Redshift. Add a forecasting widget Publish the analysis as a dashboard.

D. Create an Amazon SageMaker model for forecasting. Integrate the model with an Amazon QuickSight dataset. Create a widget for the dataset. Publish the analysis as a dashboard.

Correct Answer: C

You can add a forecasting widget to your existing analysis, and publish it as a dashboard. Reference: https://docs.aws.amazon.com/quicksight/latest/user/forecasts-and-whatifs.html