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.pass4itsure.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 data analyst runs a large number of data manipulation language (DML) queries by using Amazon Athena with the JDBC driver. Recently, a query failed after it ran for 30 minutes. The query returned the following message:

java.sql.SQLException: Query timeout

The data analyst does not immediately need the query results. However, the data analyst needs a long-term solution for this problem.

Which solution will meet these requirements?

A. Split the query into smaller queries to search smaller subsets of data

B. In the settings for Athena, adjust the DML query timeout limit

C. In the Service Quotas console, request an increase for the DML query timeout

D. Save the tables as compressed .csv files

Correct Answer: C

Reference: https://docs.aws.amazon.com/athena/latest/ug/service-limits.html

QUESTION 2

A company\\'s system operators and security engineers need to analyze activities within specific date ranges of AWS CloudTrail logs. All log files are stored in an Amazon S3 bucket, and the size of the logs is more than 5 TB. The solution must be cost-effective and maximize query performance.

Which solution meets these requirements?

A. Copy the logs to a new S3 bucket with a prefix structure of . Use the date column as a partition key. Create a table on Amazon Athena based on the objects in the new bucket. Automatically add metadata partitions by using the MSCK REPAIR TABLE command in Athena. Use Athena to query the table and partitions.

B. Create a table on Amazon Athena. Manually add metadata partitions by using the ALTER TABLE ADD PARTITION statement, and use multiple columns for the partition key. Use Athena to query the table and partitions.

C. Launch an Amazon EMR cluster and use Amazon S3 as a data store for Apache HBase. Load the logs from the S3 bucket to an HBase table on Amazon EMR. Use Amazon Athena to query the table and partitions.

D. Create an AWS Glue job to copy the logs from the S3 source bucket to a new S3 bucket and create a table using Apache Parquet file format, Snappy as compression codec, and partition by date. Use Amazon Athena to query the table and partitions.

Correct Answer: D

QUESTION 3

A telecommunications company stores its call records as JSON files in an Amazon S3 bucket. The company uses

Amazon Athena to query the records and wants to improve query performance. The data is stored in data records that have up to 300 different columns. The most common query uses a subset of only 12 of the columns.

Which solution will improve the query performance?

A. Convert the data to Apache Parquet files by using native JSON Serializer/Deserializer (SerDe) libraries.

B. Convert the data to Apache Parquet files by using Amazon EMR. Compress the files by using Snappy.

C. Convert the data to Apache Parquet files by using Amazon EMR. Compress the files by using gzip.

D. Convert the data to Apache Avro files by using Athena. Compress the files by using bzip2.

Correct Answer: B

QUESTION 4

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 OpenSearch Service (Amazon Elasticsearch Service) cluster.

The company has very slow query performance on the Amazon OpenSearch Service (Amazon Elasticsearch Service) index and occasionally sees errors from Kinesis Data Firehose when attempting to write to the index. The Amazon OpenSearch Service (Amazon Elasticsearch Service) 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 OpenSearch Service (Amazon Elasticsearch Service)?

A. Increase the memory of the Amazon OpenSearch Service (Amazon Elasticsearch Service) master nodes.

B. Decrease the number of Amazon OpenSearch Service (Amazon Elasticsearch Service) data nodes.

C. Decrease the number of Amazon OpenSearch Service (Amazon Elasticsearch Service) shards for the index.

D. Increase the number of Amazon OpenSearch Service (Amazon Elasticsearch Service) shards for the index.

Correct Answer: C

Reference: https://www.bluematador.com/docs/troubleshooting/aws-elasticsearch-jvm-pressure#:~:text=Amazon%20recommends%20keeping%20JVM%20pressure,getting%20into%20a%20red%20state

QUESTION 5

An ecommerce company stores customer purchase data in Amazon RDS. The company wants a solution to store and analyze historical data. The most recent 6 months of data will be queried frequently for analytics workloads. This data is several terabytes large. Once a month, historical data for the last 5 years must be accessible and will be joined with the more recent data. The company wants to optimize performance and cost.

Which storage solution will meet these requirements?

A. Create a read replica of the RDS database to store the most recent 6 months of data. Copy the historical data into Amazon S3. Create an AWS Glue Data Catalog of the data in Amazon S3 and Amazon RDS. Run historical queries using Amazon Athena.

B. Use an ETL tool to incrementally load the most recent 6 months of data into an Amazon Redshift cluster. Run more frequent queries against this cluster. Create a read replica of the RDS database to run queries on the historical data.

C. Incrementally copy data from Amazon RDS to Amazon S3. Create an AWS Glue Data Catalog of the data in Amazon S3. Use Amazon Athena to query the data.

D. Incrementally copy data from Amazon RDS to Amazon S3. Load and store the most recent 6 months of data in Amazon Redshift. Configure an Amazon Redshift Spectrum table to connect to all historical data.

Correct Answer: D

The cost-effective way to query across S3 and Amazon RDS is using Amazon redshift spectrum.

Reference: https://docs.aws.amazon.com/redshift/latest/dg/c-using-spectrum.html https://www.upsolver.com/blog/aws-athena-pricing-redshift-comparison

QUESTION 6

A company has a data lake on AWS that ingests sources of data from multiple business units and uses Amazon Athena for queries. The storage layer is Amazon S3 using the AWS Glue Data Catalog. The company wants to make the data available to its data scientists and business analysts. However, the company first needs to manage data access for Athena based on user roles and responsibilities.

What should the company do to apply these access controls with the LEAST operational overhead?

A. Define security policy-based rules for the users and applications by role in AWS Lake Formation.

B. Define security policy-based rules for the users and applications by role in AWS Identity and Access Management (IAM).

C. Define security policy-based rules for the tables and columns by role in AWS Glue.

D. Define security policy-based rules for the tables and columns by role in AWS Identity and Access Management (IAM).

Correct Answer: D

QUESTION 7

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 8

A company wants find ways to expand its website business by analyzing customer orders and purchasing trends. To perform data analysis, a pipeline must support daily data ingestion from the production databases into a data lake that is built on Amazon S3. The website uses Amazon DynamoDB to store product details and Amazon Aurora PostgreSQL to store order details in production.

Which solution can be used to accomplish these goals with LEAST operational overhead?

A. Leverage AWS Database Migration Service (AWS DMS) to run two continuous data replication jobs from both Aurora PostgreSQL and DynamoDB into Amazon S3. Leverage AWS Glue for data cataloging.

B. Set up an AWS Lake Formation workflow with blueprints for Aurora PostgreSQL and an AWS Glue ETLjob for DynamoDB to ingest data into Amazon S3. Leverage AWS Glue for data cataloging.

C. Create a custom Python script to ingest data from both Aurora PostgreSQL and Amazon DynamoDB into Amazon S3 using the AWS SDK for Python (Boto3) library. Deploy the script on an Amazon EC2 instance and schedule the job to run daily using a cron job. Leverage AWS Glue for data cataloging.

D. Use Amazon EMR to ingest data from both Aurora PostgreSQL and DynamoDB into Amazon S3. Leverage Apache Hive on the same EMR cluster for data cataloging.

Correct Answer: A

QUESTION 9

A gaming company is building a serverless data lake. The company is ingesting streaming data into Amazon Kinesis Data Streams and is writing the data to Amazon S3 through Amazon Kinesis Data Firehose. The company is using 10 MB as the S3 buffer size and is using 90 seconds as the buffer interval. The company runs an AWS Glue ETL job to merge and transform the data to a different format before writing the data back to Amazon S3.

Recently, the company has experienced substantial growth in its data volume. The AWS Glue ETL jobs are frequently showing an OutOfMemoryError error.

Which solutions will resolve this issue without incurring additional costs? (Choose two.)

A. Place the small files into one S3 folder. Define one single table for the small S3 files in AWS Glue Data Catalog. Rerun the AWS Glue ETL jobs against this AWS Glue table

B. Create an AWS Lambda function to merge small S3 files and invoke them periodically. Run the AWS Glue ETL jobs after successful completion of the Lambda function.

C. Run the S3DistCp utility in Amazon EMR to merge a large number of small S3 files before running the AWS Glue ETL jobs.

D. Use the groupFiles setting in the AWS Glue ETL job to merge small S3 files and rerun AWS Glue ETL jobs. Most Voted

E. Update the Kinesis Data Firehose S3 buffer size to 128 MB. Update the buffer interval to 900 seconds.

Correct Answer: DE

Grouping files together reduces the memory footprint on the Spark driver as well as simplifying file split orchestration. Buffer size increase avoids creating small size files Reference: https://docs.aws.amazon.com/glue/latest/dg/grouping-input-files.html https://docs.aws.amazon.com/glue/latest/dg/grouping-input-files.html

QUESTION 10

A financial company recently added more features to its mobile app. This required the creation of a new topic named mobile_transfers in the existing Amazon Managed Streaming for Apache Kafka (Amazon MSK) cluster. A few days after adding this new topic, an Amazon CloudWatch alarm for the RootDiskUsed metric for the MSK cluster was raised.

How should a data specialist resolve this issue?

A. Expand the storage of the MSK broker. Configure storage auto-expansion.

B. Increase the storage for the Apache ZooKeeper nodes.

C. Update the MSK broker instance to the next larger type. Restart the MSK cluster.

D. Specify the Target-Volume-in-GiB parameter for the mobile_transfers topic.

Correct Answer: D

QUESTION 11

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 12

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 13

A human resources company maintains a 10-node Amazon Redshift cluster to run analytics queries on the company\\'s

data. The Amazon Redshift cluster contains a product table and a transactions table, and both tables have a product_sku column. The tables are over 100 GB in size. The majority of queries run on both tables.

Which distribution style should the company use for the two tables to achieve optimal query performance?

A. An EVEN distribution style for both tables

B. A KEY distribution style for both tables

C. An ALL distribution style for the product table and an EVEN distribution style for the transactions table

D. An EVEN distribution style for the product table and an KEY distribution style for the transactions table

Correct Answer: B

QUESTION 14

A large telecommunications company is planning to set up a data catalog and metadata management for multiple data sources running on AWS. The catalog will be used to maintain the metadata of all the objects stored in the data stores. The data stores are composed of structured sources like Amazon RDS and Amazon Redshift, and semistructured sources like JSON and XML files stored in Amazon S3. The catalog must be updated on a regular basis, be able to detect the changes to object metadata, and require the least possible administration.

Which solution meets these requirements?

A. Use Amazon Aurora as the data catalog. Create AWS Lambda functions that will connect and gather the metadata information from multiple sources and update the data catalog in Aurora. Schedule the Lambda functions periodically.

B. Use the AWS Glue Data Catalog as the central metadata repository. Use AWS Glue crawlers to connect to multiple data stores and update the Data Catalog with metadata changes. Schedule the crawlers periodically to update the metadata catalog.

C. Use Amazon DynamoDB as the data catalog. Create AWS Lambda functions that will connect and gather the metadata information from multiple sources and update the DynamoDB catalog. Schedule the Lambda functions periodically.

D. Use the AWS Glue Data Catalog as the central metadata repository. Extract the schema for RDS and Amazon Redshift sources and build the Data Catalog. Use AWS crawlers for data stored in Amazon S3 to infer the schema and automatically update the Data Catalog.

Correct Answer: D

Reference: https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hive-metastore-glue.html

QUESTION 15

A company stores its sales and marketing data that includes personally identifiable information (PII) in Amazon S3. The company allows its analysts to launch their own Amazon EMR cluster and run analytics reports with the data. To meet compliance requirements, the company must ensure the data is not publicly accessible throughout this process. A data engineer has secured Amazon S3 but must ensure the individual EMR clusters created by the analysts are not exposed to the public internet.

Which solution should the data engineer to meet this compliance requirement with LEAST amount of effort?

A. Create an EMR security configuration and ensure the security configuration is associated with the EMR clusters when they are created.

B. Check the security group of the EMR clusters regularly to ensure it does not allow inbound traffic from IPv4 0.0.0.0/0 or IPv6 ::/0.

C. Enable the block public access setting for Amazon EMR at the account level before any EMR cluster is created.

D. Use AWS WAF to block public internet access to the EMR clusters across the board.

Correct Answer: C

https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-block-public-access.html