Q&A

Up-to-date Questions and Answers from authentic resources to improve knowledge and pass the exam at very first attempt. ----- Guaranteed.

KILL EXAMS

DAS-C01 Dumps
DAS-C01 Braindumps
DAS-C01 Real Questions
DAS-C01 Practice Test
DAS-C01 Actual Questions

PASS

killexams.com

Amazon

DAS-C01

AWS Certified Data Analytics - Specialty (DAS-C01)

ORDER FULL VERSION

A company wants to provide its data analysts with uninterrupted access to the data in its Amazon Redshift cluster. All data is streamed to an Amazon S3 bucket with Amazon Kinesis Data Firehose. An AWS Glue job that is scheduled to run every 5 minutes issues a COPY command to move the data into Amazon Redshift.
The amount of data delivered is uneven throughout then day, and cluster utilization is high during certain periods. The COPY command usually completes within a couple of seconds. However, when load spike occurs, locks can exist and data can be missed. Currently, the AWS Glue job is configured to run without retries, with timeout at 5 minutes and concurrency at 1.
How should a data analytics specialist configure the AWS Glue job to optimize fault tolerance and improve data availability in the Amazon Redshift cluster?

- A. Increase the number of retries. Decrease the timeout value. Increase the job concurrency.
- B. Keep the number of retries at 0. Decrease the timeout value. Increase the job concurrency.
- C. Keep the number of retries at 0. Decrease the timeout value. Keep the job concurrency at 1.
- D. Keep the number of retries at 0. Increase the timeout value. Keep the job concurrency at 1.

Answer: B

Question: 94

A retail company leverages Amazon Athena for ad-hoc queries against an AWS Glue Data Catalog. The data analytics team manages the data catalog and data access for the company. The data analytics team wants to separate queries and manage the cost of running those queries by different workloads and teams.
Ideally, the data analysts want to group the queries run by different users within a team, store the query results in individual Amazon S3 buckets specific to each team, and enforce cost constraints on the queries run against the Data Catalog.
Which solution meets these requirements?

- A. Create IAM groups and resource tags for each team within the company. Set up IAM policies that control user access and actions on the Data Catalog resources.
- B. Create Athena resource groups for each team within the company and assign users to these groups. Add S3 bucket names and other query configurations to the properties list for the resource groups.
- C. Create Athena workgroups for each team within the company. Set up IAM workgroup policies that control user access and actions on the workgroup resources.
- D. Create Athena query groups for each team within the company and assign users to the groups.

Answer: A

Question: 95

A manufacturing company uses Amazon S3 to store its data. The company wants to use AWS Lake Formation to provide granular-level security on those data assets. The data is in Apache Parquet format. The company has set a deadline for a consultant to build a data lake.

How should the consultant create the MOST cost-effective solution that meets these requirements?

- A. Run Lake Formation blueprints to move the data to Lake Formation. Once Lake Formation has the data, apply permissions on Lake Formation.
- B. To create the data catalog, run an AWS Glue crawler on the existing Parquet data. Register the Amazon S3 path and then apply permissions through Lake Formation to provide granular-level security.
- C. Install Apache Ranger on an Amazon EC2 instance and integrate with Amazon EMR. Using Ranger policies, create role-based access control for the existing data assets in Amazon S3.
- D. Create multiple IAM roles for different users and groups. Assign IAM roles to different data assets in Amazon S3 to create table-based and column-based access controls.

Answer: C

Question: 96

A company has an application that uses the Amazon Kinesis Client Library (KCL) to read records from a Kinesis data stream.
After a successful marketing campaign, the application experienced a significant increase in usage. As a result, a data analyst had to split some shards in the data stream. When the shards were split, the application started throwing an ExpiredIteratorExceptions error sporadically.
What should the data analyst do to resolve this?

- A. Increase the number of threads that process the stream records.
- B. Increase the provisioned read capacity units assigned to the streams Amazon DynamoDB table.
- C. Increase the provisioned write capacity units assigned to the streams Amazon DynamoDB table.
- D. Decrease the provisioned write capacity units assigned to the streams Amazon DynamoDB table.

Answer: C

Question: 97

A company is building a service to monitor fleets of vehicles. The company collects IoT data from a device in each vehicle and loads the data into Amazon
Redshift in near-real time. Fleet owners upload .csv files containing vehicle reference data into Amazon S3 at different times throughout the day. A nightly process loads the vehicle reference data from Amazon S3 into Amazon Redshift. The company joins the IoT data from the device and the vehicle reference data to power reporting and dashboards. Fleet owners are frustrated by waiting a day for the dashboards to update.
Which solution would provide the SHORTEST delay between uploading reference data to Amazon S3 and the change showing up in the owners dashboards?

- A. Use S3 event notifications to trigger an AWS Lambda function to copy the vehicle reference data into Amazon Redshift immediately when the reference data is uploaded to Amazon S3.
- B. Create and schedule an AWS Glue Spark job to run every 5 minutes. The job inserts reference data into Amazon Redshift.
- C. Send reference data to Amazon Kinesis Data Streams. Configure the Kinesis data stream to directly load the reference data into Amazon Redshift in real time.
- D. Send the reference data to an Amazon Kinesis Data Firehose delivery stream. Configure Kinesis with a buffer interval of 60 seconds and to directly load the data into Amazon Redshift.

Answer: A

Question: 98

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

Answer: BCF

Question: 99

A retail company wants to use Amazon QuickSight to generate dashboards for web and in-store sales. A group of 50 business intelligence professionals will develop and use the dashboards. Once ready, the dashboards will be shared with a group of 1,000 users.
The sales data comes from different stores and is uploaded to Amazon S3 every 24 hours. The data is partitioned by year and month, and is stored in Apache
Parquet format. The company is using the AWS Glue Data Catalog as its main data catalog and Amazon Athena for querying. The total size of the uncompressed data that the dashboards query from at any point is 200 GB.
Which configuration will provide the MOST cost-effective solution that meets these requirements?

- A. Load the data into an Amazon Redshift cluster by using the COPY command. Configure 50 author users and 1,000 reader users. Use QuickSight Enterprise edition. Configure an Amazon Redshift data source with a direct query option.
- B. Use QuickSight Standard edition. Configure 50 author users and 1,000 reader users. Configure an Athena data source with a direct query option.
- C. Use QuickSight Enterprise edition. Configure 50 author users and 1,000 reader users. Configure an Athena data source and import the data into SPICE. Automatically refresh every 24 hours.
- D. Use QuickSight Enterprise edition. Configure 1 administrator and 1,000 reader users. Configure an S3 data source and import the data into SPICE. Automatically refresh every 24 hours.

Answer: C

Question: 100

A central government organization is collecting events from various internal applications using Amazon Managed Streaming for Apache Kafka (Amazon MSK).
The organization has configured a separate Kafka topic for each application to separate the data. For security reasons, the Kafka cluster has been configured to only allow TLS encrypted data and it encrypts the data at rest.
A recent application update showed that one of the applications was configured incorrectly, resulting in writing data to a Kafka topic that belongs to another application. This resulted in multiple errors in the analytics pipeline as data from different applications appeared on the same topic. After this incident, the organization wants to prevent applications from writing to a topic different than the one they should write to.
Which solution meets these requirements with the least amount of effort?

- A. Create a different Amazon EC2 security group for each application. Configure each security group to have access to a specific topic in the Amazon MSK cluster. Attach the security group to each application based on the topic that the applications should read and write to.
- B. Install Kafka Connect on each application instance and configure each Kafka Connect instance to write to a specific topic only.
- C. Use Kafka ACLs and configure read and write permissions for each topic. Use the distinguished name of the clients TLS certificates as the principal of the ACL.
- D. Create a different Amazon EC2 security group for each application. Create an Amazon MSK cluster and Kafka topic for each application. Configure each security group to have access to the specific cluster.

Answer: B

Question: 101

A company wants to collect and process events data from different departments in near-real time. Before storing the data in Amazon S3, the company needs to clean the data by standardizing the format of the address and

timestamp columns. The data varies in size based on the overall load at each particular point in time. A single data record can be 100 KB-10 MB.
How should a data analytics specialist design the solution for data ingestion?

- A. Use Amazon Kinesis Data Streams. Configure a stream for the raw data. Use a Kinesis Agent to write data to the stream. Create an Amazon Kinesis Data Analytics application that reads data from the raw stream, cleanses it, and stores the output to Amazon S3.
- B. Use Amazon Kinesis Data Firehose. Configure a Firehose delivery stream with a preprocessing AWS Lambda function for data cleansing. Use a Kinesis Agent to write data to the delivery stream. Configure Kinesis Data Firehose to deliver the data to Amazon S3.
- C. Use Amazon Managed Streaming for Apache Kafka. Configure a topic for the raw data. Use a Kafka producer to write data to the topic. Create an application on Amazon EC2 that reads data from the topic by using the Apache Kafka consumer API, cleanses the data, and writes to Amazon S3.
- D. Use Amazon Simple Queue Service (Amazon SQS). Configure an AWS Lambda function to read events from the SQS queue and upload the events to Amazon S3.

Answer: B

Question: 102

An operations team notices that a few AWS Glue jobs for a given ETL application are failing. The AWS Glue jobs read a large number of small JOSN files from an
Amazon S3 bucket and write the data to a different S3 bucket in Apache Parquet format with no major transformations. Upon initial investigation, a data engineer notices the following error message in the History tab on the AWS Glue console: Command Failed with Exit Code 1.
Upon further investigation, the data engineer notices that the driver memory profile of the failed jobs crosses the safe threshold of 50% usage quickly and reaches
90"95% soon after. The average memory usage across all executors continues to be less than 4%.
The data engineer also notices the following error while examining the related Amazon CloudWatch Logs.

```
# java.lang.OutOfMemoryError: Java heap space
# -XX: OnOutOfMemoryError= "kill -9 %p"
# Executing /bin/sh -c "kill -9 12039"...
```

What should the data engineer do to solve the failure in the MOST cost-effective way?

- A. Change the worker type from Standard to G.2X.
- B. Modify the AWS Glue ETL code to use the groupFiles: inPartition feature.
- C. Increase the fetch size setting by using AWS Glue dynamics frame.
- D. Modify maximum capacity to increase the total maximum data processing units (DPUs) used.

Answer: D

Question: 103

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 companys requirements?

- A. Kinesis Agent
- B. Kinesis Producer Library (KPL)
- C. Kinesis Data Firehose
- D. Kinesis SDK

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

SAMPLE QUESTIONS