Vendor: Amazon

Exam Code: DAS-C01

Exam Name: AWS Certified Data Analytics - Specialty (DAS-C01) Exam

Version: DEMO

QUESTION 1
A company has developed an Apache Hive script to batch process data stared in Amazon S3. The script needs to run once every day and store the output in Amazon S3. The company tested the script, and it completes within 30 minutes on a small local three-node cluster.

Which solution is the MOST cost-effective for scheduling and executing the script?

A. Create an AWS Lambda function to spin up an Amazon EMR cluster with a Hive execution step. Set KeepJobFlowAliveWhenNoSteps to false and disable the termination protection flag. Use Amazon CloudWatch Events to schedule the Lambda function to run daily.
B. Use the AWS Management Console to spin up an Amazon EMR cluster with Python Hue. Hive, and Apache Oozie.
Set the termination protection flag to true and use Spot Instances for the core nodes of the cluster.
Configure an Oozie workflow in the cluster to invoke the Hive script daily.
C. Create an AWS Glue job with the Hive script to perform the batch operation.
Configure the job to run once a day using a time-based schedule.
D. Use AWS Lambda layers and load the Hive runtime to AWS Lambda and copy the Hive script. Schedule the Lambda function to run daily by creating a workflow using AWS Step Functions.

Answer: A
Explanation:
Not B because we are not supposed to run core nodes in spot instances, just task nodes and it is more expensive because to schedule with oozie, our cluster have to be up all the time. It is not C because glue cannot run hive script, and it is not c because lambda cannot run hive scripts also.
https://docs.aws.amazon.com/AmazonCloudWatch/latest/events/RunLambdaSchedule.html

QUESTION 2
A company wants to improve the data load time of a sales data dashboard. Data has been collected as .csv files and stored within an Amazon S3 bucket that is partitioned by date. The data is then loaded to an Amazon Redshift data warehouse for frequent analysis. The data volume is up to 500 GB per day.

Which solution will improve the data loading performance?

A. Compress .csv files and use an INSERT statement to ingest data into Amazon Redshift.
B. Split large .csv files, then use a COPY command to load data into Amazon Redshift.
C. Use Amazon Kinesis Data Firehose to ingest data into Amazon Redshift.
D. Load the .csv files in an unsorted key order and vacuum the table in Amazon Redshift.

Answer: B
Explanation:
https://docs.aws.amazon.com/redshift/latest/dg/c_loading-data-best-practices.html
Also, Vaccum command will help with clearing up space on the cluster but not improve data loading time.

QUESTION 3
A mortgage company has a microservice for accepting payments. This microservice uses the Amazon DynamoDB encryption client with AWS KMS managed keys to encrypt the sensitive data before writing the data to DynamoDB. The finance team should be able to load this data into Amazon Redshift and aggregate the values within the sensitive fields. The Amazon Redshift cluster is shared with other data analysts from different business units.

Which steps should a data analyst take to accomplish this task efficiently and securely?

A. Create an AWS Lambda function to process the DynamoDB stream. Decrypt the sensitive data using the same KMS key.
Save the output to a restricted S3 bucket for the finance team.
Create a finance table in Amazon Redshift that is accessible to the finance team only.
Use the COPY command to load the data from Amazon S3 to the finance table.
B. Create an AWS Lambda function to process the DynamoDB stream.
Save the output to a restricted S3 bucket for the finance team.
Create a finance table in Amazon Redshift that is accessible to the finance team only.
Use the COPY command with the IAM role that has access to the KMS key to load the data from S3 to the finance table.
C. Create an Amazon EMR cluster with an EMR_EC2_DefaultRole role that has access to the KMS key.
Create Apache Hive tables that reference the data stored in DynamoDB and the finance table in Amazon Redshift.
In Hive, select the data from DynamoDB and then insert the output to the finance table in Amazon Redshift.
D. Create an Amazon EMR cluster. Create Apache Hive tables that reference the data stored in DynamoDB.
Insert the output to the restricted Amazon S3 bucket for the finance team.
Use the COPY command with the IAM role that has access to the KMS key to load the data from Amazon S3 to the finance table in Amazon Redshift.

Answer: B

QUESTION 4
A company is building a data lake and needs to ingest data from a relational database that has time-series data. The company wants to use managed services to accomplish this. The process needs to be scheduled daily and bring incremental data only from the source into Amazon S3.

What is the MOST cost-effective approach to meet these requirements?

A. Use AWS Glue to connect to the data source using JDBC Drivers.
Ingest incremental records only using job bookmarks.
B. Use AWS Glue to connect to the data source using JDBC Drivers.
Store the last updated key in an Amazon DynamoDB table and ingest the data using the updated key as a filter.
C. Use AWS Glue to connect to the data source using JDBC Drivers and ingest the entire dataset.
Use appropriate Apache Spark libraries to compare the dataset, and find the delta.
D. Use AWS Glue to connect to the data source using JDBC Drivers and ingest the full data.
Use AWS DataSync to ensure the delta only is written into Amazon S3.

Answer: A

QUESTION 5
An Amazon Redshift database contains sensitive user data. Logging is necessary to meet compliance requirements. The logs must contain database authentication attempts, connections, and disconnections. The logs must also contain each query run against the database and record which database user ran each query.

Which steps will create the required logs?

A. Enable Amazon Redshift Enhanced VPC Routing. Enable VPC Flow Logs to monitor traffic.
B. Allow access to the Amazon Redshift database using AWS IAM only. Log access using AWS CloudTrail.
C. Enable audit logging for Amazon Redshift using the AWS Management Console or the AWS CLI.
D. Enable and download audit reports from AWS Artifact.

Answer: C
Explanation:
https://docs.aws.amazon.com/redshift/latest/mgmt/db-auditing.html

QUESTION 6
A company that monitors weather conditions from remote construction sites is setting up a solution to collect temperature data from the following two weather stations.

- Station A, which has 10 sensors
- Station B, which has five sensors

These weather stations were placed by onsite subject-matter experts.

Each sensor has a unique ID. The data collected from each sensor will be collected using Amazon Kinesis Data Streams.

Based on the total incoming and outgoing data throughput, a single Amazon Kinesis data stream with two shards is created. Two partition keys are created based on the station names. During testing, there is a bottleneck on data coming from Station A, but not from Station B. Upon review, it is confirmed that the total stream throughput is still less than the allocated Kinesis Data Streams throughput.

How can this bottleneck be resolved without increasing the overall cost and complexity of the solution, while retaining the data collection quality requirements?

A. Increase the number of shards in Kinesis Data Streams to increase the level of parallelism.
B. Create a separate Kinesis data stream for Station A with two shards, and stream Station A sensor data to the new stream.
C. Modify the partition key to use the sensor ID instead of the station name.
D. Reduce the number of sensors in Station A from 10 to 5 sensors.

Answer: C

QUESTION 7
Once a month, a company receives a 100 MB .csv file compressed with gzip. The file contains 50,000 property listing records and is stored in Amazon S3 Glacier. The company needs its data analyst to query a subset of the data for a specific vendor.

What is the most cost-effective solution?

A. Load the data into Amazon S3 and query it with Amazon S3 Select.
B. Query the data from Amazon S3 Glacier directly with Amazon Glacier Select.
C. Load the data to Amazon S3 and query it with Amazon Athena.
D. Load the data to Amazon S3 and query it with Amazon Redshift Spectrum.

Answer: A
Explanation:
https://aws.amazon.com/blogs/aws/s3-glacier-select/

QUESTION 8
A retail company is building its data warehouse solution using Amazon Redshift. As a part of that effort, the company is loading hundreds of files into the fact table created in its Amazon Redshift cluster. The company wants the solution to achieve the highest throughput and optimally use cluster resources when loading data into the company's fact table.

How should the company meet these requirements?

A. Use multiple COPY commands to load the data into the Amazon Redshift cluster.
B. Use S3DistCp to load multiple files into the Hadoop Distributed File System (HDFS) and use an HDFS connector to ingest the data into the Amazon Redshift cluster.
C. Use LOAD commands equal to the number of Amazon Redshift cluster nodes and load the data in parallel into each node.
D. Use a single COPY command to load the data into the Amazon Redshift cluster.

Answer: D
Explanation:
https://docs.aws.amazon.com/redshift/latest/dg/c_best-practices-single-copy-command.html

QUESTION 9
A data analyst is designing a solution to interactively query datasets with SQL using a JDBC connection. Users will join data stored in Amazon S3 in Apache ORC format with data stored in Amazon Elasticsearch Service (Amazon ES) and Amazon Aurora MySQL.

Which solution will provide the MOST up-to-date results?

A. Use AWS Glue jobs to ETL data from Amazon ES and Aurora MySQL to Amazon S3. Query the data with Amazon Athena.
B. Use Amazon DMS to stream data from Amazon ES and Aurora MySQL to Amazon Redshift. Query the data with Amazon Redshift.
C. Query all the datasets in place with Apache Spark SQL running on an AWS Glue developer endpoint.
D. Query all the datasets in place with Apache Presto running on Amazon EMR.

Answer: D
Explanation:
Presto is an open source distributed SQL query engine for running interactive analytic queries against data sources of all sizes ranging from gigabytes to petabytes.

Thank You for Trying Our Product

Passleader Certification Exam Features:

★ More than **99,900** Satisfied Customers Worldwide.

★ Average **99.9%** Success Rate.

★ **Free Update** to match latest and real exam scenarios.

★ **Instant Download** Access! No Setup required.

★ Questions & Answers are downloadable in **PDF** format and **VCE** test engine format.

★ Multi-Platform capabilities - **Windows, Laptop, Mac, Android, iPhone, iPod, iPad**.

★ **100%** Guaranteed Success or **100%** Money Back Guarantee.

★ **Fast**, helpful support **24x7**.

View list of all certification exams: http://www.passleader.com/all-products.html

10% Discount Coupon Code: **ASTR14**