Vendor: Amazon

Exam Code: DEA-C01

Exam Name: AWS Certified Data Engineer - Associate (DEA-C01) Exam

Version: DEMO

QUESTION 1
A company stores daily records of the financial performance of investment portfolios in .csv format in an Amazon S3 bucket. A data engineer uses AWS Glue crawlers to crawl the S3 data.
The data engineer must make the S3 data accessible daily in the AWS Glue Data Catalog.
Which solution will meet these requirements?

A. Create an IAM role that includes the AmazonS3FullAccess policy. Associate the role with the crawler. Specify the S3 bucket path of the source data as the crawler's data store. Create a daily schedule to run the crawler. Configure the output destination to a new path in the existing S3 bucket.
B. Create an IAM role that includes the AWSGlueServiceRole policy. Associate the role with the crawler. Specify the S3 bucket path of the source data as the crawler's data store. Create a daily schedule to run the crawler. Specify a database name for the output.
C. Create an IAM role that includes the AmazonS3FullAccess policy. Associate the role with the crawler. Specify the S3 bucket path of the source data as the crawler's data store. Allocate data processing units (DPUs) to run the crawler every day. Specify a database name for the output.
D. Create an IAM role that includes the AWSGlueServiceRole policy. Associate the role with the crawler. Specify the S3 bucket path of the source data as the crawler's data store. Allocate data processing units (DPUs) to run the crawler every day. Configure the output destination to a new path in the existing S3 bucket.

Answer: B
Explanation:
https://docs.aws.amazon.com/glue/latest/dg/tutorial-add-crawler.html

QUESTION 2
A company loads transaction data for each day into Amazon Redshift tables at the end of each day. The company wants to have the ability to track which tables have been loaded and which tables still need to be loaded.
A data engineer wants to store the load statuses of Redshift tables in an Amazon DynamoDB table. The data engineer creates an AWS Lambda function to publish the details of the load statuses to DynamoDB.
How should the data engineer invoke the Lambda function to write load statuses to the DynamoDB table?

A. Use a second Lambda function to invoke the first Lambda function based on Amazon CloudWatch events.
B. Use the Amazon Redshift Data API to publish an event to Amazon EventBridge. Configure an EventBridge rule to invoke the Lambda function.
C. Use the Amazon Redshift Data API to publish a message to an Amazon Simple Queue Service (Amazon SQS) queue. Configure the SQS queue to invoke the Lambda function.
D. Use a second Lambda function to invoke the first Lambda function based on AWS CloudTrail events.

Answer: B
Explanation:
https://docs.aws.amazon.com/redshift/latest/mgmt/data-api-monitoring-events.html

QUESTION 3
A data engineer needs to securely transfer 5 TB of data from an on-premises data center to an Amazon S3 bucket. Approximately 5% of the data changes every day. Updates to the data need to be regularly proliferated to the S3 bucket. The data includes files that are in multiple formats. The data engineer needs to automate the transfer process and must schedule the process to run

periodically.
Which AWS service should the data engineer use to transfer the data in the MOST operationally efficient way?

A. AWS DataSync
B. AWS Glue
C. AWS Direct Connect
D. Amazon S3 Transfer Acceleration

Answer: A
Explanation:
AWS DataSync is a managed data transfer service that simplifies and accelerates moving large amounts of data online between on-premises storage and Amazon S3, EFS, or FSx for Windows File Server. DataSync is optimized for efficient, incremental, and reliable transfers of large datasets, making it suitable for transferring 5 TB of data with daily updates.

QUESTION 4
A company uses an on-premises Microsoft SQL Server database to store financial transaction data. The company migrates the transaction data from the on-premises database to AWS at the end of each month. The company has noticed that the cost to migrate data from the on-premises database to an Amazon RDS for SQL Server database has increased recently.
The company requires a cost-effective solution to migrate the data to AWS. The solution must cause minimal downtown for the applications that access the database.
Which AWS service should the company use to meet these requirements?

A. AWS Lambda
B. AWS Database Migration Service (AWS DMS)
C. AWS Direct Connect
D. AWS DataSync

Answer: B
Explanation:
AWS Database Migration Service (DMS) is specifically designed for migrating data from various sources, including on-premises databases, to AWS with minimal downtime and disruption to applications. It supports homogeneous migrations (e.g., SQL Server to SQL Server) as well as heterogeneous migrations (e.g., SQL Server to Amazon RDS for SQL Server).

QUESTION 5
A data engineer is building a data pipeline on AWS by using AWS Glue extract, transform, and load (ETL) jobs. The data engineer needs to process data from Amazon RDS and MongoDB, perform transformations, and load the transformed data into Amazon Redshift for analytics. The data updates must occur every hour.
Which combination of tasks will meet these requirements with the LEAST operational overhead? (Choose two.)

A. Configure AWS Glue triggers to run the ETL jobs every hour.
B. Use AWS Glue DataBrew to clean and prepare the data for analytics.
C. Use AWS Lambda functions to schedule and run the ETL jobs every hour.
D. Use AWS Glue connections to establish connectivity between the data sources and Amazon Redshift.
E. Use the Redshift Data API to load transformed data into Amazon Redshift.

Answer: AD
Explanation:
AWS Glue triggers provide a simple and integrated way to schedule ETL jobs. By configuring these triggers to run hourly, the data engineer can ensure that the data processing and updates occur as required without the need for external scheduling tools or custom scripts. This approach is directly integrated with AWS Glue, reducing the complexity and operational overhead.
AWS Glue supports connections to various data sources, including Amazon RDS and MongoDB. By using AWS Glue connections, the data engineer can easily configure and manage the connectivity between these data sources and Amazon Redshift. This method leverages AWS Glue's built-in capabilities for data source integration, thus minimizing operational complexity and ensuring a seamless data flow from the sources to the destination (Amazon Redshift).

QUESTION 6
A company uses an Amazon Redshift cluster that runs on RA3 nodes. The company wants to scale read and write capacity to meet demand. A data engineer needs to identify a solution that will turn on concurrency scaling.
Which solution will meet this requirement?

A. Turn on concurrency scaling in workload management (WLM) for Redshift Serverless workgroups.
B. Turn on concurrency scaling at the workload management (WLM) queue level in the Redshift cluster.
C. Turn on concurrency scaling in the settings during the creation of any new Redshift cluster.
D. Turn on concurrency scaling for the daily usage quota for the Redshift cluster.

Answer: B
Explanation:
https://docs.aws.amazon.com/redshift/latest/dg/concurrency-scaling-queues.html

QUESTION 7
A data engineer must orchestrate a series of Amazon Athena queries that will run every day. Each query can run for more than 15 minutes.
Which combination of steps will meet these requirements MOST cost-effectively? (Choose two.)

A. Use an AWS Lambda function and the Athena Boto3 client start_query_execution API call to invoke the Athena queries programmatically.
B. Create an AWS Step Functions workflow and add two states. Add the first state before the Lambda function. Configure the second state as a Wait state to periodically check whether the Athena query has finished using the Athena Boto3 get_query_execution API call. Configure the workflow to invoke the next query when the current query has finished running.
C. Use an AWS Glue Python shell job and the Athena Boto3 client start_query_execution API call to invoke the Athena queries programmatically.
D. Use an AWS Glue Python shell script to run a sleep timer that checks every 5 minutes to determine whether the current Athena query has finished running successfully. Configure the Python shell script to invoke the next query when the current query has finished running.
E. Use Amazon Managed Workflows for Apache Airflow (Amazon MWAA) to orchestrate the Athena queries in AWS Batch.

Answer: AB

QUESTION 8

A company is migrating on-premises workloads to AWS. The company wants to reduce overall operational overhead. The company also wants to explore serverless options.
The company's current workloads use Apache Pig, Apache Oozie, Apache Spark, Apache Hbase, and Apache Flink. The on-premises workloads process petabytes of data in seconds. The company must maintain similar or better performance after the migration to AWS.
Which extract, transform, and load (ETL) service will meet these requirements?

A. AWS Glue
B. Amazon EMR
C. AWS Lambda
D. Amazon Redshift

Answer: B
Explanation:
Glue is like the more good-looking one, but weaker brother of EMR. So when it's about petabyte scales, let EMR do the work and have Glue stay away from the action.

QUESTION 9
A data engineer must use AWS services to ingest a dataset into an Amazon S3 data lake. The data engineer profiles the dataset and discovers that the dataset contains personally identifiable information (PII). The data engineer must implement a solution to profile the dataset and obfuscate the PII.
Which solution will meet this requirement with the LEAST operational effort?

A. Use an Amazon Kinesis Data Firehose delivery stream to process the dataset. Create an AWS Lambda transform function to identify the PII. Use an AWS SDK to obfuscate the PII. Set the S3 data lake as the target for the delivery stream.
B. Use the Detect PII transform in AWS Glue Studio to identify the PII. Obfuscate the PII. Use an AWS Step Functions state machine to orchestrate a data pipeline to ingest the data into the S3 data lake.
C. Use the Detect PII transform in AWS Glue Studio to identify the PII. Create a rule in AWS Glue Data Quality to obfuscate the PII. Use an AWS Step Functions state machine to orchestrate a data pipeline to ingest the data into the S3 data lake.
D. Ingest the dataset into Amazon DynamoDB. Create an AWS Lambda function to identify and obfuscate the PII in the DynamoDB table and to transform the data. Use the same Lambda function to ingest the data into the S3 data lake.

Answer: C

QUESTION 10
A company maintains multiple extract, transform, and load (ETL) workflows that ingest data from the company's operational databases into an Amazon S3 based data lake. The ETL workflows use AWS Glue and Amazon EMR to process data.
The company wants to improve the existing architecture to provide automated orchestration and to require minimal manual effort.
Which solution will meet these requirements with the LEAST operational overhead?

A. AWS Glue workflows
B. AWS Step Functions tasks
C. AWS Lambda functions
D. Amazon Managed Workflows for Apache Airflow (Amazon MWAA) workflows

Answer: B
Explanation:
https://docs.aws.amazon.com/step-functions/latest/dg/connect-emr.html
https://docs.aws.amazon.com/step-functions/latest/dg/connect-glue.html

QUESTION 11
A company currently stores all of its data in Amazon S3 by using the S3 Standard storage class.
A data engineer examined data access patterns to identify trends. During the first 6 months, most data files are accessed several times each day. Between 6 months and 2 years, most data files are accessed once or twice each month. After 2 years, data files are accessed only once or twice each year.
The data engineer needs to use an S3 Lifecycle policy to develop new data storage rules. The new storage solution must continue to provide high availability.
Which solution will meet these requirements in the MOST cost-effective way?

A. Transition objects to S3 One Zone-Infrequent Access (S3 One Zone-IA) after 6 months. Transfer objects to S3 Glacier Flexible Retrieval after 2 years.
B. Transition objects to S3 Standard-Infrequent Access (S3 Standard-IA) after 6 months. Transfer objects to S3 Glacier Flexible Retrieval after 2 years.
C. Transition objects to S3 Standard-Infrequent Access (S3 Standard-IA) after 6 months. Transfer objects to S3 Glacier Deep Archive after 2 years.
D. Transition objects to S3 One Zone-Infrequent Access (S3 One Zone-IA) after 6 months. Transfer objects to S3 Glacier Deep Archive after 2 years.

Answer: B

QUESTION 12
A company maintains an Amazon Redshift provisioned cluster that the company uses for extract, transform, and load (ETL) operations to support critical analysis tasks. A sales team within the company maintains a Redshift cluster that the sales team uses for business intelligence (BI) tasks.
The sales team recently requested access to the data that is in the ETL Redshift cluster so the team can perform weekly summary analysis tasks. The sales team needs to join data from the ETL cluster with data that is in the sales team's BI cluster.
The company needs a solution that will share the ETL cluster data with the sales team without interrupting the critical analysis tasks. The solution must minimize usage of the computing resources of the ETL cluster.
Which solution will meet these requirements?

A. Set up the sales team BI cluster as a consumer of the ETL cluster by using Redshift data sharing.
B. Create materialized views based on the sales team's requirements. Grant the sales team direct access to the ETL cluster.
C. Create database views based on the sales team's requirements. Grant the sales team direct access to the ETL cluster.
D. Unload a copy of the data from the ETL cluster to an Amazon S3 bucket every week. Create an Amazon Redshift Spectrum table based on the content of the ETL cluster.

Answer: A
Explanation:
https://docs.aws.amazon.com/redshift/latest/dg/data_sharing_intro.html
Supporting different kinds of business-critical workloads – Use a central extract, transform, and load (ETL) cluster that shares data with multiple business intelligence (BI) or analytic clusters. This approach provides read workload isolation and chargeback for individual workloads. You can

size and scale your individual workload compute according to the workload-specific requirements of price and performance.

QUESTION 13
A data engineer needs to join data from multiple sources to perform a one-time analysis job. The data is stored in Amazon DynamoDB, Amazon RDS, Amazon Redshift, and Amazon S3.
Which solution will meet this requirement MOST cost-effectively?

A. Use an Amazon EMR provisioned cluster to read from all sources. Use Apache Spark to join the data and perform the analysis.
B. Copy the data from DynamoDB, Amazon RDS, and Amazon Redshift into Amazon S3. Run Amazon Athena queries directly on the S3 files.
C. Use Amazon Athena Federated Query to join the data from all data sources.
D. Use Redshift Spectrum to query data from DynamoDB, Amazon RDS, and Amazon S3 directly from Redshift.

Answer: C
Explanation:
You can query these sources by using Federated Queries, which is a native feature of Athena.
The other options may increase costs and operational overhead, as they use more than one service to achieve the same result.
https://docs.aws.amazon.com/athena/latest/ug/connectors-available.html

QUESTION 14
A company wants to implement real-time analytics capabilities. The company wants to use Amazon Kinesis Data Streams and Amazon Redshift to ingest and process streaming data at the rate of several gigabytes per second. The company wants to derive near real-time insights by using existing business intelligence (BI) and analytics tools.
Which solution will meet these requirements with the LEAST operational overhead?

A. Use Kinesis Data Streams to stage data in Amazon S3. Use the COPY command to load data from Amazon S3 directly into Amazon Redshift to make the data immediately available for real-time analysis.
B. Access the data from Kinesis Data Streams by using SQL queries. Create materialized views directly on top of the stream. Refresh the materialized views regularly to query the most recent stream data.
C. Create an external schema in Amazon Redshift to map the data from Kinesis Data Streams to an Amazon Redshift object. Create a materialized view to read data from the stream. Set the materialized view to auto refresh.
D. Connect Kinesis Data Streams to Amazon Kinesis Data Firehose. Use Kinesis Data Firehose to stage the data in Amazon S3. Use the COPY command to load the data from Amazon S3 to a table in Amazon Redshift.

Answer: C
Explanation:
https://docs.aws.amazon.com/redshift/latest/dg/materialized-view-streaming-ingestion.html