Amazon-Web-Services

Exam Questions DAS-C01

AWS Certified Data Analytics - Specialty

NEW QUESTION 1
A data analyst is using AWS Glue to organize, cleanse, validate, and format a 200 GB dataset. The data analyst triggered the job to run with the Standard worker type. After 3 hours, the AWS Glue job status is still RUNNING. Logs from the job run show no error codes. The data analyst wants to improve the job execution time without overprovisioning.
Which actions should the data analyst take?

A. Enable job bookmarks in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the executor-cores job parameter.
B. Enable job metrics in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the maximum capacity job parameter.
C. Enable job metrics in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the spark.yarn.executor.memoryOverhead job parameter.
D. Enable job bookmarks in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the num-executors job parameter.

Answer: B

NEW QUESTION 2
An online retailer needs to deploy a product sales reporting solution. The source data is exported from an external online transaction processing (OLTP) system for reporting. Roll-up data is calculated each day for the previous day’s activities. The reporting system has the following requirements:
Have the daily roll-up data readily available for 1 year.
After 1 year, archive the daily roll-up data for occasional but immediate access.
The source data exports stored in the reporting system must be retained for 5 years. Query access will be needed only for re-evaluation, which may occur within the first 90 days.
Which combination of actions will meet these requirements while keeping storage costs to a minimum? (Choose two.)

A. Store the source data initially in the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage clas
B. Apply a lifecycle configuration that changes the storage class to Amazon S3 Glacier Deep Archive 90 days after creation, and then deletes the data 5 years after creation.
C. Store the source data initially in the Amazon S3 Glacier storage clas
D. Apply a lifecycle configuration that changes the storage class from Amazon S3 Glacier to Amazon S3 Glacier Deep Archive 90 days after creation, and then deletes the data 5 years after creation.
E. Store the daily roll-up data initially in the Amazon S3 Standard storage clas
F. Apply a lifecycle configuration that changes the storage class to Amazon S3 Glacier Deep Archive 1 year after data creation.
G. Store the daily roll-up data initially in the Amazon S3 Standard storage clas
H. Apply a lifecycle configuration that changes the storage class to Amazon S3 Standard-Infrequent Access (S3 Standard-IA) 1 year afterdata creation.
I. Store the daily roll-up data initially in the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage clas
J. Apply a lifecycle configuration that changes the storage class to Amazon S3 Glacier 1 year after data creation.

Answer: AD

NEW QUESTION 3
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 catalo
B. Create AWS Lambda functions that will connect and gather themetadata information from multiple sources and update the data catalog in Auror
C. Schedule the Lambda functions periodically.
D. Use the AWS Glue Data Catalog as the central metadata repositor
E. Use AWS Glue crawlers to connect to multiple data stores and update the Data Catalog with metadata change
F. Schedule the crawlers periodically to update the metadata catalog.
G. Use Amazon DynamoDB as the data catalo
H. Create AWS Lambda functions that will connect and gather the metadata information from multiple sources and update the DynamoDB catalo
I. Schedule the Lambda functions periodically.
J. Use the AWS Glue Data Catalog as the central metadata repositor
K. Extract the schema for RDS and Amazon Redshift sources and build the Data Catalo
L. Use AWS crawlers for data stored in Amazon S3 to infer the schema and automatically update the Data Catalog.

Answer: D

NEW QUESTION 4
A power utility company is deploying thousands of smart meters to obtain real-time updates about power consumption. The company is using Amazon Kinesis Data Streams to collect the data streams from smart meters. The consumer application uses the Kinesis Client Library (KCL) to retrieve the stream data. The company has only one consumer application.
The company observes an average of 1 second of latency from the moment that a record is written to the stream until the record is read by a consumer application. The company must reduce this latency to 500 milliseconds.
Which solution meets these requirements?

A. Use enhanced fan-out in Kinesis Data Streams.
B. Increase the number of shards for the Kinesis data stream.
C. Reduce the propagation delay by overriding the KCL default settings.
D. Develop consumers by using Amazon Kinesis Data Firehose.

Answer: C

Explanation:
The KCL defaults are set to follow the best practice of polling every 1 second. This default results in average propagation delays that are typically below 1 second.

NEW QUESTION 5
A company's marketing team has asked for help in identifying a high performing long-term storage service for their data based on the following requirements:

- The data size is approximately 32 TB uncompressed.
- There is a low volume of single-row inserts each day.
- There is a high volume of aggregation queries each day.
- Multiple complex joins are performed.
- The queries typically involve a small subset of the columns in a table. Which storage service will provide the MOST performant solution?

A. Amazon Aurora MySQL
B. Amazon Redshift
C. Amazon Neptune
D. Amazon Elasticsearch

Answer: B

NEW QUESTION 6
A company has developed several AWS Glue jobs to validate and transform its data from Amazon S3 and load it into Amazon RDS for MySQL in batches once every day. The ETL jobs read the S3 data using a DynamicFrame. Currently, the ETL developers are experiencing challenges in processing only the incremental data on every run, as the AWS Glue job processes all the S3 input data on each run.
Which approach would allow the developers to solve the issue with minimal coding effort?

A. Have the ETL jobs read the data from Amazon S3 using a DataFrame.
B. Enable job bookmarks on the AWS Glue jobs.
C. Create custom logic on the ETL jobs to track the processed S3 objects.
D. Have the ETL jobs delete the processed objects or data from Amazon S3 after each run.

Answer: B

NEW QUESTION 7
A data analyst is using Amazon QuickSight for data visualization across multiple datasets generated by applications. Each application stores files within a separate Amazon S3 bucket. AWS Glue Data Catalog is used as a central catalog across all application data in Amazon S3. A new application stores its data within a separate S3 bucket. After updating the catalog to include the new application data source, the data analyst created a new Amazon QuickSight data source from an Amazon Athena table, but the import into SPICE failed.
How should the data analyst resolve the issue?

A. Edit the permissions for the AWS Glue Data Catalog from within the Amazon QuickSight console.
B. Edit the permissions for the new S3 bucket from within the Amazon QuickSight console.
C. Edit the permissions for the AWS Glue Data Catalog from within the AWS Glue console.
D. Edit the permissions for the new S3 bucket from within the S3 console.

Answer: B

NEW QUESTION 8
A global company has different sub-organizations, and each sub-organization sells its products and services in various countries. The company's senior leadership wants to quickly identify which sub-organization is the strongest performer in each country. All sales data is stored in Amazon S3 in Parquet format.
Which approach can provide the visuals that senior leadership requested with the least amount of effort?

A. Use Amazon QuickSight with Amazon Athena as the data sourc
B. Use heat maps as the visual type.
C. Use Amazon QuickSight with Amazon S3 as the data sourc
D. Use heat maps as the visual type.
E. Use Amazon QuickSight with Amazon Athena as the data sourc
F. Use pivot tables as the visual type.
G. Use Amazon QuickSight with Amazon S3 as the data sourc
H. Use pivot tables as the visual type.

Answer: A

NEW QUESTION 9
An insurance company has raw data in JSON format that is sent without a predefined schedule through an Amazon Kinesis Data Firehose delivery stream to an Amazon S3 bucket. An AWS Glue crawler is scheduled to run every 8 hours to update the schema in the data catalog of the tables stored in the S3 bucket. Data analysts analyze the data using Apache Spark SQL on Amazon EMR set up with AWS Glue Data Catalog as the metastore. Data analysts say that, occasionally, the data they receive is stale. A data engineer needs to provide access to the most up-to-date data.
Which solution meets these requirements?

A. Create an external schema based on the AWS Glue Data Catalog on the existing Amazon Redshift cluster to query new data in Amazon S3 with Amazon Redshift Spectrum.
B. Use Amazon CloudWatch Events with the rate (1 hour) expression to execute the AWS Glue crawler every hour.
C. Using the AWS CLI, modify the execution schedule of the AWS Glue crawler from 8 hours to 1 minute.
D. Run the AWS Glue crawler from an AWS Lambda function triggered by an S3:ObjectCreated:* event notification on the S3 bucket.

Answer: D

Explanation:
https://docs.aws.amazon.com/AmazonS3/latest/dev/NotificationHowTo.html "you can use a wildcard (for example, s3:ObjectCreated:*) to request notification when an object is created regardless of the API used" "AWS Lambda can run custom code in response to Amazon S3 bucket events. You upload your custom code to AWS Lambda and create what is called a Lambda function. When Amazon S3 detects an event of a specific type (for example, an object created event), it can publish the event to AWS Lambda and invoke your function in Lambda. In response, AWS Lambda runs your function."

NEW QUESTION 10
A technology company is creating a dashboard that will visualize and analyze time-sensitive data. The data will come in through Amazon Kinesis Data Firehose with the butter interval set to 60 seconds. The dashboard must support near-real-time data.
Which visualization solution will meet these requirements?

A. Select Amazon Elasticsearch Service (Amazon ES) as the endpoint for Kinesis Data Firehos
B. Set up a Kibana dashboard using the data in Amazon ES with the desired analyses and visualizations.
C. Select Amazon S3 as the endpoint for Kinesis Data Firehos
D. Read data into an Amazon SageMaker Jupyter notebook and carry out the desired analyses and visualizations.
E. Select Amazon Redshift as the endpoint for Kinesis Data Firehos
F. Connect Amazon QuickSight with SPICE to Amazon Redshift to create the desired analyses and visualizations.
G. Select Amazon S3 as the endpoint for Kinesis Data Firehos
H. Use AWS Glue to catalog the data and Amazon Athena to query i
I. Connect Amazon QuickSight with SPICE to Athena to create the desired analyses and visualizations.

Answer: A

NEW QUESTION 10
A financial company uses Apache Hive on Amazon EMR for ad-hoc queries. Users are complaining of sluggish performance.
A data analyst notes the following:

- Approximately 90% of queries are submitted 1 hour after the market opens.
- Hadoop Distributed File System (HDFS) utilization never exceeds 10%.

Which solution would help address the performance issues?

A. Create instance fleet configurations for core and task node
B. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch CapacityRemainingGB metri
C. Create an automatic scaling policy to scale in the instance fleet based on the CloudWatch CapacityRemainingGB metric.
D. Create instance fleet configurations for core and task node
E. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch YARNMemoryAvailablePercentage metri
F. Create an automatic scaling policy to scale in the instance fleet based on the CloudWatch YARNMemoryAvailablePercentage metric.
G. Create instance group configurations for core and task node
H. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch CapacityRemainingGB metri
I. Create anautomatic scaling policy to scale in the instance groups based on the CloudWatch CapacityRemainingGB metric.
J. Create instance group configurations for core and task node
K. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch YARNMemoryAvailablePercentage metri
L. Create an automatic scaling policy to scale in the instance groups based on the CloudWatch YARNMemoryAvailablePercentage metric.

Answer: D

Explanation:
https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html

NEW QUESTION 14
A company's data analyst needs to ensure that queries executed in Amazon Athena cannot scan more than a prescribed amount of data for cost control purposes. Queries that exceed the prescribed threshold must be canceled immediately.
What should the data analyst do to achieve this?

A. Configure Athena to invoke an AWS Lambda function that terminates queries when the prescribed threshold is crossed.
B. For each workgroup, set the control limit for each query to the prescribed threshold.
C. Enforce the prescribed threshold on all Amazon S3 bucket policies
D. For each workgroup, set the workgroup-wide data usage control limit to the prescribed threshold.

Answer: B

Explanation:
https://docs.aws.amazon.com/athena/latest/ug/manage-queries-control-costs-with-workgroups.html

NEW QUESTION 15
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 Driver
B. Ingest incremental records only using job bookmarks.
C. Use AWS Glue to connect to the data source using JDBC Driver
D. Store the last updated key in an Amazon DynamoDB table and ingest the data using the updated key as a filter.
E. Use AWS Glue to connect to the data source using JDBC Drivers and ingest the entire datase
F. Use appropriate Apache Spark libraries to compare the dataset, and find the delta.
G. Use AWS Glue to connect to the data source using JDBC Drivers and ingest the full dat
H. Use AWSDataSync to ensure the delta only is written into Amazon S3.

Answer: A

Explanation:
https://docs.aws.amazon.com/glue/latest/dg/monitor-continuations.html

NEW QUESTION 20
A media company wants to perform machine learning and analytics on the data residing in its Amazon S3 data lake. There are two data transformation requirements that will enable the consumers within the company to create reports:

- Daily transformations of 300 GB of data with different file formats landing in Amazon S3 at a scheduled time.
- One-time transformations of terabytes of archived data residing in the S3 data lake.

Which combination of solutions cost-effectively meets the company's requirements for transforming the data? (Choose three.)

A. For daily incoming data, use AWS Glue crawlers to scan and identify the schema.
B. For daily incoming data, use Amazon Athena to scan and identify the schema.
C. For daily incoming data, use Amazon Redshift to perform transformations.
D. For daily incoming data, use AWS Glue workflows with AWS Glue jobs to perform transformations.
E. For archived data, use Amazon EMR to perform data transformations.
F. For archived data, use Amazon SageMaker to perform data transformations.

Answer: ADE

NEW QUESTION 25
A real estate company has a mission-critical application using Apache HBase in Amazon EMR. Amazon EMR is configured with a single master node. The company has over 5 TB of data stored on an Hadoop Distributed File System (HDFS). The company wants a cost-effective solution to make its HBase data highly available.
Which architectural pattern meets company's requirements?

A. Use Spot Instances for core and task nodes and a Reserved Instance for the EMR master node.Configurethe EMR cluster with multiple master node
B. Schedule automated snapshots using AmazonEventBridge.
C. Store the data on an EMR File System (EMRFS) instead of HDF
D. Enable EMRFS consistent view.Create an EMR HBase cluster with multiple master node
E. Point the HBase root directory to an Amazon S3 bucket.
F. Store the data on an EMR File System (EMRFS) instead of HDFS and enable EMRFS consistent view.Run two separate EMR clusters in two different Availability Zone
G. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.
H. Store the data on an EMR File System (EMRFS) instead of HDFS and enable EMRFS consistent view.Create a primary EMR HBase cluster with multiple master node
I. Create a secondary EMR HBase read- replica cluster in a separate Availability Zon
J. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.

Answer: D

NEW QUESTION 28
A telecommunications company is looking for an anomaly-detection solution to identify fraudulent calls. The company currently uses Amazon Kinesis to stream voice call records in a JSON format from its on-premises database to Amazon S3. The existing dataset contains voice call records with 200 columns. To detect fraudulent calls, the solution would need to look at 5 of these columns only.
The company is interested in a cost-effective solution using AWS that requires minimal effort and experience in anomaly-detection algorithms.
Which solution meets these requirements?

A. Use an AWS Glue job to transform the data from JSON to Apache Parque
B. Use AWS Glue crawlers to discover the schema and build the AWS Glue Data Catalo
C. Use Amazon Athena to create a table with a subset of column
D. Use Amazon QuickSight to visualize the data and then use Amazon QuickSight machine learning-powered anomaly detection.
E. Use Kinesis Data Firehose to detect anomalies on a data stream from Kinesis by running SQL queries, which compute an anomaly score for all calls and store the output in Amazon RD
F. Use Amazon Athena to build a dataset and Amazon QuickSight to visualize the results.
G. Use an AWS Glue job to transform the data from JSON to Apache Parque
H. Use AWS Glue crawlers to discover the schema and build the AWS Glue Data Catalo
I. Use Amazon SageMaker to build an anomaly detection model that can detect fraudulent calls by ingesting data from Amazon S3.
J. Use Kinesis Data Analytics to detect anomalies on a data stream from Kinesis by running SQL queries, which compute an anomaly score for all call
K. Connect Amazon QuickSight to Kinesis Data Analytics to visualize the anomaly scores.

Answer: A

NEW QUESTION 32
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 applicatio
B. Configure each security group to have access to a specific topic in the Amazon MSK cluste
C. Attach the security group to each application based on the topic that the applications should read and write to.
D. Install Kafka Connect on each application instance and configure each Kafka Connect instance to write to a specific topic only.
E. Use Kafka ACLs and configure read and write permissions for each topi
F. Use the distinguished name of the clients' TLS certificates as the principal of the ACL.
G. Create a different Amazon EC2 security group for each applicatio

H. Create an Amazon MSK cluster and Kafka topic for each applicatio
I. Configure each security group to have access to the specific cluster.

Answer: B

NEW QUESTION 34
A regional energy company collects voltage data from sensors attached to buildings. To address any known dangerous conditions, the company wants to be alerted when a sequence of two voltage drops is detected within 10 minutes of a voltage spike at the same building. It is important to ensure that all messages are delivered as quickly as possible. The system must be fully managed and highly available. The company also needs a solution that will automatically scale up as it covers additional cites with this monitoring feature. The alerting system is subscribed to an Amazon SNS topic for remediation.
Which solution meets these requirements?

A. Create an Amazon Managed Streaming for Kafka cluster to ingest the data, and use an Apache Spark Streaming with Apache Kafka consumer API in an automatically scaled Amazon EMR cluster to process the incoming dat
B. Use the Spark Streaming application to detect the known event sequence and send the SNS message.
C. Create a REST-based web service using Amazon API Gateway in front of an AWS Lambda function.Create an Amazon RDS for PostgreSQL database with sufficient Provisioned IOPS (PIOPS). In the Lambda function, store incoming events in the RDS database and query the latest data to detect the known event sequence and send the SNS message.
D. Create an Amazon Kinesis Data Firehose delivery stream to capture the incoming sensor dat
E. Use an AWS Lambda transformation function to detect the known event sequence and send the SNS message.
F. Create an Amazon Kinesis data stream to capture the incoming sensor data and create another stream for alert message
G. Set up AWS Application Auto Scaling on bot
H. Create a Kinesis Data Analytics for Java application to detect the known event sequence, and add a message to the message strea
I. Configure an AWS Lambda function to poll the message stream and publish to the SNS topic.

Answer: D

NEW QUESTION 39
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 Stream
B. Configure a stream for the raw dat
C. Use a Kinesis Agent to write data to the strea
D. Create an Amazon Kinesis Data Analytics application that reads data from the raw stream, cleanses it, and stores the output to Amazon S3.
E. Use Amazon Kinesis Data Firehos
F. Configure a Firehose delivery stream with a preprocessing AWS Lambda function for data cleansin
G. Use a Kinesis Agent to write data to the delivery strea
H. Configure Kinesis Data Firehose to deliver the data to Amazon S3.
I. Use Amazon Managed Streaming for Apache Kafk
J. Configure a topic for the raw dat
K. Use a Kafka producer to write data to the topi
L. 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.
M. 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

NEW QUESTION 44
A hospital uses wearable medical sensor devices to collect data from patients. The hospital is architecting a near-real-time solution that can ingest the data securely at scale. The solution should also be able to remove the patient's protected health information (PHI) from the streaming data and store the data in durable storage.
Which solution meets these requirements with the least operational overhead?

A. Ingest the data using Amazon Kinesis Data Streams, which invokes an AWS Lambda function using Kinesis Client Library (KCL) to remove all PH
B. Write the data in Amazon S3.
C. Ingest the data using Amazon Kinesis Data Firehose to write the data to Amazon S3. Have Amazon S3 trigger an AWS Lambda function that parses the sensor data to remove all PHI in Amazon S3.
D. Ingest the data using Amazon Kinesis Data Streams to write the data to Amazon S3. Have the data stream launch an AWS Lambda function that parses the sensor data and removes all PHI in Amazon S3.
E. Ingest the data using Amazon Kinesis Data Firehose to write the data to Amazon S3. Implement a transformation AWS Lambda function that parses the sensor data to remove all PHI.

Answer: D

Explanation:
https://aws.amazon.com/blogs/big-data/persist-streaming-data-to-amazon-s3-using-amazon-kinesis-firehose-and

NEW QUESTION 48
A data analytics specialist is building an automated ETL ingestion pipeline using AWS Glue to ingest compressed files that have been uploaded to an Amazon S3 bucket. The ingestion pipeline should support incremental data processing.
Which AWS Glue feature should the data analytics specialist use to meet this requirement?

A. Workflows
B. Triggers
C. Job bookmarks
D. Classifiers

Answer: C

NEW QUESTION 51
A company uses Amazon Redshift for its data warehousing needs. ETL jobs run every night to load data, apply business rules, and create aggregate tables for reporting. The company's data analysis, data science, and business intelligence teams use the data warehouse during regular business hours. The workload management is set to auto, and separate queues exist for each team with the priority set to NORMAL.
Recently, a sudden spike of read queries from the data analysis team has occurred at least twice daily, and queries wait in line for cluster resources. The company needs a solution that enables the data analysis team to avoid query queuing without impacting latency and the query times of other teams.
Which solution meets these requirements?

A. Increase the query priority to HIGHEST for the data analysis queue.
B. Configure the data analysis queue to enable concurrency scaling.
C. Create a query monitoring rule to add more cluster capacity for the data analysis queue when queries are waiting for resources.
D. Use workload management query queue hopping to route the query to the next matching queue.

Answer: D

NEW QUESTION 53
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 strea
B. Decrypt the sensitive data using the same KMS ke
C. Save the output to a restricted S3 bucket for the finance tea
D. Create a finance table in Amazon Redshift that is accessible to the finance team onl
E. Use the COPY command to load the data from Amazon S3 to the finance table.
F. Create an AWS Lambda function to process the DynamoDB strea
G. Save the output to a restricted S3 bucket for the finance tea
H. Create a finance table in Amazon Redshift that is accessible to the finance team onl
I. 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.
J. 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 Redshif
K. In Hive, select the data from DynamoDB and then insert the output to the finance table in Amazon Redshift.
L. Create an Amazon EMR cluste
M. Create Apache Hive tables that reference the data stored inDynamoD
N. Insert the output to the restricted Amazon S3 bucket for the finance tea
O. 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

NEW QUESTION 57
A manufacturing company wants to create an operational analytics dashboard to visualize metrics from equipment in near-real time. The company uses Amazon Kinesis Data Streams to stream the data to other applications. The dashboard must automatically refresh every 5 seconds. A data analytics specialist must design a solution that requires the least possible implementation effort.
Which solution meets these requirements?

A. Use Amazon Kinesis Data Firehose to store the data in Amazon S3. Use Amazon QuickSight to build the dashboard.
B. Use Apache Spark Streaming on Amazon EMR to read the data in near-real tim
C. Develop a custom application for the dashboard by using D3.js.
D. Use Amazon Kinesis Data Firehose to push the data into an Amazon Elasticsearch Service (Amazon ES) cluste
E. Visualize the data by using a Kibana dashboard.
F. Use AWS Glue streaming ETL to store the data in Amazon S3. Use Amazon QuickSight to build the dashboard.

Answer: B

NEW QUESTION 61
An education provider’s learning management system (LMS) is hosted in a 100 TB data lake that is built on Amazon S3. The provider’s LMS supports hundreds of schools. The provider wants to build an advanced analytics reporting platform using Amazon Redshift to handle complex queries with optimal performance.
System users will query the most recent 4 months of data 95% of the time while 5% of the queries will leverage data from the previous 12 months.
Which solution meets these requirements in the MOST cost-effective way?

A. Store the most recent 4 months of data in the Amazon Redshift cluste
B. Use Amazon Redshift Spectrum to query data in the data lak
C. Use S3 lifecycle management rules to store data from the previous 12 months in Amazon S3 Glacier storage.
D. Leverage DS2 nodes for the Amazon Redshift cluste
E. Migrate all data from Amazon S3 to Amazon Redshif
F. Decommission the data lake.
G. Store the most recent 4 months of data in the Amazon Redshift cluste
H. Use Amazon Redshift Spectrum to query data in the data lak
I. Ensure the S3 Standard storage class is in use with objects in the data lake.
J. Store the most recent 4 months of data in the Amazon Redshift cluste
K. Use Amazon Redshift federated queries to join cluster data with the data lake to reduce cost
L. Ensure the S3 Standard storage class is in use with objects in the data lake.

Answer: C

NEW QUESTION 65
A large ride-sharing company has thousands of drivers globally serving millions of unique customers every day. The company has decided to migrate an existing data mart to Amazon Redshift. The existing schema includes the following tables.
A trips fact table for information on completed rides. A drivers dimension table for driver profiles. A customers fact table holding customer profile information.
The company analyzes trip details by date and destination to examine profitability by region. The drivers data rarely changes. The customers data frequently changes.
What table design provides optimal query performance?

A. Use DISTSTYLE KEY (destination) for the trips table and sort by dat
B. Use DISTSTYLE ALL for the drivers and customers tables.
C. Use DISTSTYLE EVEN for the trips table and sort by dat
D. Use DISTSTYLE ALL for the drivers table.Use DISTSTYLE EVEN for the customers table.
E. Use DISTSTYLE KEY (destination) for the trips table and sort by dat
F. Use DISTSTYLE ALL for the drivers tabl
G. Use DISTSTYLE EVEN for the customers table.
H. Use DISTSTYLE EVEN for the drivers table and sort by dat
I. Use DISTSTYLE ALL for both fact tables.

Answer: C

Explanation:
https://www.matillion.com/resources/blog/aws-redshift-performance-choosing-the-right-distribution-styles/#:~:t
https://docs.aws.amazon.com/redshift/latest/dg/c_best-practices-best-dist-key.html

NEW QUESTION 67
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

Answer: B

NEW QUESTION 71
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 comman
B. Configure 50 author users and 1,000 reader user
C. Use QuickSight Enterprise editio
D. Configure an Amazon Redshift data source with a direct query option.
E. Use QuickSight Standard editio
F. Configure 50 author users and 1,000 reader user
G. Configure an Athena data source with a direct query option.
H. Use QuickSight Enterprise editio
I. Configure 50 author users and 1,000 reader user
J. Configure an Athena data source and import the data into SPIC
K. Automatically refresh every 24 hours.
L. Use QuickSight Enterprise editio
M. Configure 1 administrator and 1,000 reader user
N. Configure an S3 data source and import the data into SPIC
O. Automatically refresh every 24 hours.

Answer: C

NEW QUESTION 73
A company hosts an on-premises PostgreSQL database that contains historical data. An internal legacy application uses the database for read-only activities. The company's business team wants to move the data to a data lake in Amazon S3 as soon as possible and enrich the data for analytics.
The company has set up an AWS Direct Connect connection between its VPC and its on-premises network. A data analytics specialist must design a solution that achieves the business team's goals with the least operational overhead.
Which solution meets these requirements?

A. Upload the data from the on-premises PostgreSQL database to Amazon S3 by using a customized batch upload proces
B. Use the AWS Glue crawler to catalog the data in Amazon S3. Use an AWS Glue job to enrich and store the result in a separate S3 bucket in Apache Parquet forma
C. Use Amazon Athena to query the data.
D. Create an Amazon RDS for PostgreSQL database and use AWS Database Migration Service (AWS DMS) to migrate the data into Amazon RD
E. Use AWS Data Pipeline to copy and enrich the data from the Amazon RDS for PostgreSQL table and move the data to Amazon S3. Use Amazon Athena to querythe data.
F. Configure an AWS Glue crawler to use a JDBC connection to catalog the data in the on-premises databas
G. Use an AWS Glue job to enrich the data and save the result to Amazon S3 in Apache Parquet forma
H. Create an Amazon Redshift cluster and use Amazon Redshift Spectrum to query the data.

I. Configure an AWS Glue crawler to use a JDBC connection to catalog the data in the on-premises databas
J. Use an AWS Glue job to enrich the data and save the result to Amazon S3 in Apache Parquet forma
K. Use Amazon Athena to query the data.

Answer: B

NEW QUESTION 74
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: ACE

Explanation:
https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-ha.html "Note : The cluster can reside only in one Availability Zone or subnet."

NEW QUESTION 77
An online gaming company is using an Amazon Kinesis Data Analytics SQL application with a Kinesis data stream as its source. The source sends three non-null fields to the application: player_id, score, and us_5_digit_zip_code.
A data analyst has a .csv mapping file that maps a small number of us_5_digit_zip_code values to a territory code. The data analyst needs to include the territory code, if one exists, as an additional output of the Kinesis Data Analytics application.
How should the data analyst meet this requirement while minimizing costs?

A. Store the contents of the mapping file in an Amazon DynamoDB tabl
B. Preprocess the records as they arrive in the Kinesis Data Analytics application with an AWS Lambda function that fetches the mapping and supplements each record to include the territory code, if one exist
C. Change the SQL query in the application to include the new field in the SELECT statement.
D. Store the mapping file in an Amazon S3 bucket and configure the reference data column headers for the.csv file in the Kinesis Data Analytics applicatio
E. Change the SQL query in the application to include a join to the file’s S3 Amazon Resource Name (ARN), and add the territory code field to the SELECT columns.
F. Store the mapping file in an Amazon S3 bucket and configure it as a reference data source for the Kinesis Data Analytics applicatio
G. Change the SQL query in the application to include a join to the reference table and add the territory code field to the SELECT columns.
H. Store the contents of the mapping file in an Amazon DynamoDB tabl
I. Change the Kinesis DataAnalytics application to send its output to an AWS Lambda function that fetches the mapping and supplements each record to include the territory code, if one exist
J. Forward the record from the Lambda function to the original application destination.

Answer: C

NEW QUESTION 79
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 Formatio
B. Once Lake Formation has the data, apply permissions on Lake Formation.
C. To create the data catalog, run an AWS Glue crawler on the existing Parquet dat
D. Register the Amazon S3 path and then apply permissions through Lake Formation to provide granular-level security.
E. Install Apache Ranger on an Amazon EC2 instance and integrate with Amazon EM
F. Using Ranger policies, create role-based access control for the existing data assets in Amazon S3.
G. Create multiple IAM roles for different users and group
H. Assign IAM roles to different data assets in Amazon S3 to create table-based and column-based access controls.

Answer: A

Explanation:
https://aws.amazon.com/blogs/big-data/building-securing-and-managing-data-lakes-with-aws-lake-formation/

NEW QUESTION 83
A company has an application that ingests streaming data. The company needs to analyze this stream over a 5-minute timeframe to evaluate the stream for anomalies with Random Cut Forest (RCF) and summarize the current count of status codes. The source and summarized data should be persisted for future use.
Which approach would enable the desired outcome while keeping data persistence costs low?

A. Ingest the data stream with Amazon Kinesis Data Stream
B. Have an AWS Lambda consumer evaluate the stream, collect the number status codes, and evaluate the data against a previously trained RCF mode
C. Persist the source and results as a time series to Amazon DynamoDB.
D. Ingest the data stream with Amazon Kinesis Data Stream
E. Have a Kinesis Data Analytics application evaluate the stream over a 5-minute window using the RCF function and summarize the count of status code
F. Persist the source and results to Amazon S3 through output delivery to Kinesis Data Firehouse.
G. Ingest the data stream with Amazon Kinesis Data Firehose with a delivery frequency of 1 minute or 1 MB in Amazon S3. Ensure Amazon S3 triggers an event to invoke an AWS Lambda consumer that evaluates the batch data, collects the number status codes, and evaluates the data against a previouslytrained RCF

mode
H. Persist the source and results as a time series to Amazon DynamoDB.
I. Ingest the data stream with Amazon Kinesis Data Firehose with a delivery frequency of 5 minutes or 1 MB into Amazon S3. Have a Kinesis Data Analytics application evaluate the stream over a 1-minute window using the RCF function and summarize the count of status code
J. Persist the results to Amazon S3 through a Kinesis Data Analytics output to an AWS Lambda integration.

Answer: B

NEW QUESTION 84
A marketing company is using Amazon EMR clusters for its workloads. The company manually installs third party libraries on the clusters by logging in to the master nodes. A data analyst needs to create an automated solution to replace the manual process.
Which options can fulfill these requirements? (Choose two.)

A. Place the required installation scripts in Amazon S3 and execute them using custom bootstrap actions.
B. Place the required installation scripts in Amazon S3 and execute them through Apache Spark in Amazon EMR.
C. Install the required third-party libraries in the existing EMR master nod
D. Create an AMI out of that master node and use that custom AMI to re-create the EMR cluster.
E. Use an Amazon DynamoDB table to store the list of required application
F. Trigger an AWS Lambda function with DynamoDB Streams to install the software.
G. Launch an Amazon EC2 instance with Amazon Linux and install the required third-party libraries on the instanc
H. Create an AMI and use that AMI to create the EMR cluster.

Answer: AE

Explanation:
https://aws.amazon.com/about-aws/whats-new/2017/07/amazon-emr-now-supports-launching-clusters-with-cust
https://docs.aws.amazon.com/de_de/emr/latest/ManagementGuide/emr-plan-bootstrap.html

NEW QUESTION 89
An online retail company is migrating its reporting system to AWS. The company's legacy system runs data processing on online transactions using a complex series of nested Apache Hive queries. Transactional data is exported from the online system to the reporting system several times a day. Schemas in the files are stable between updates.
A data analyst wants to quickly migrate the data processing to AWS, so any code changes should be minimized. To keep storage costs low, the data analyst decides to store the data in Amazon S3. It is vital that the data from the reports and associated analytics is completely up to date based on the data in Amazon S3.
Which solution meets these requirements?

A. Create an AWS Glue Data Catalog to manage the Hive metadat
B. Create an AWS Glue crawler over Amazon S3 that runs when data is refreshed to ensure that data changes are update
C. Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
D. Create an AWS Glue Data Catalog to manage the Hive metadat
E. Create an Amazon EMR cluster with consistent view enable
F. Run emrfs sync before each analytics step to ensure data changes are update
G. Create an EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
H. Create an Amazon Athena table with CREATE TABLE AS SELECT (CTAS) to ensure data is refreshed from underlying queries against the raw datase
I. Create an AWS Glue Data Catalog to manage the Hive metadata over the CTAS tabl
J. Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
K. Use an S3 Select query to ensure that the data is properly update
L. Create an AWS Glue Data Catalog to manage the Hive metadata over the S3 Select tabl
M. Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.

Answer: A

NEW QUESTION 92
A company needs to collect streaming data from several sources and store the data in the AWS Cloud. The dataset is heavily structured, but analysts need to perform several complex SQL queries and need consistent performance. Some of the data is queried more frequently than the rest. The company wants a solution that meets its performance requirements in a cost-effective manner.
Which solution meets these requirements?

A. Use Amazon Managed Streaming for Apache Kafka to ingest the data to save it to Amazon S3. Use Amazon Athena to perform SQL queries over the ingested data.
B. Use Amazon Managed Streaming for Apache Kafka to ingest the data to save it to Amazon Redshift.Enable Amazon Redshift workload management (WLM) to prioritize workloads.
C. Use Amazon Kinesis Data Firehose to ingest the data to save it to Amazon Redshif
D. Enable Amazon Redshift workload management (WLM) to prioritize workloads.
E. Use Amazon Kinesis Data Firehose to ingest the data to save it to Amazon S3. Load frequently queried data to Amazon Redshift using the COPY comman
F. Use Amazon Redshift Spectrum for less frequently queried data.

Answer: B

NEW QUESTION 97
A transportation company uses IoT sensors attached to trucks to collect vehicle data for its global delivery fleet. The company currently sends the sensor data in small .csv files to Amazon S3. The files are then loaded into a 10-node Amazon Redshift cluster with two slices per node and queried using both Amazon Athena and Amazon Redshift. The company wants to optimize the files to reduce the cost of querying and also improve the speed of data loading into the Amazon Redshift cluster.
Which solution meets these requirements?

A. Use AWS Glue to convert all the files from .csv to a single large Apache Parquet fil
B. COPY the file into Amazon Redshift and query the file with Athena from Amazon S3.
C. Use Amazon EMR to convert each .csv file to Apache Avr

D. COPY the files into Amazon Redshift and query the file with Athena from Amazon S3.
E. Use AWS Glue to convert the files from .csv to a single large Apache ORC fil
F. COPY the file into Amazon Redshift and query the file with Athena from Amazon S3.
G. Use AWS Glue to convert the files from .csv to Apache Parquet to create 20 Parquet file
H. COPY the files into Amazon Redshift and query the files with Athena from Amazon S3.

Answer: D

NEW QUESTION 100
A company is sending historical datasets to Amazon S3 for storage. A data engineer at the company wants to make these datasets available for analysis using Amazon Athena. The engineer also wants to encrypt the Athena query results in an S3 results location by using AWS solutions for encryption. The requirements for encrypting the query results are as follows:
Use custom keys for encryption of the primary dataset query results.
Use generic encryption for all other query results.
Provide an audit trail for the primary dataset queries that shows when the keys were used and by whom. Which solution meets these requirements?

A. Use server-side encryption with S3 managed encryption keys (SSE-S3) for the primary datase
B. Use SSE-S3 for the other datasets.
C. Use server-side encryption with customer-provided encryption keys (SSE-C) for the primary dataset.Use server-side encryption with S3 managed encryption keys (SSE-S3) for the other datasets.
D. Use server-side encryption with AWS KMS managed customer master keys (SSE-KMS CMKs) for the primary datase
E. Use server-side encryption with S3 managed encryption keys (SSE-S3) for the other datasets.
F. Use client-side encryption with AWS Key Management Service (AWS KMS) customer managed keys for the primary datase
G. Use S3 client-side encryption with client-side keys for the other datasets.

Answer: A

NEW QUESTION 105
A manufacturing company has been collecting IoT sensor data from devices on its factory floor for a year and is storing the data in Amazon Redshift for daily analysis. A data analyst has determined that, at an expected ingestion rate of about 2 TB per day, the cluster will be undersized in less than 4 months. A long-term solution is needed. The data analyst has indicated that most queries only reference the most recent 13 months of data, yet there are also quarterly reports that need to query all the data generated from the past 7 years. The chief technology officer (CTO) is concerned about the costs, administrative effort, and performance of a long-term solution.
Which solution should the data analyst use to meet these requirements?

A. Create a daily job in AWS Glue to UNLOAD records older than 13 months to Amazon S3 and delete those records from Amazon Redshif
B. Create an external table in Amazon Redshift to point to the S3 locatio
C. Use Amazon Redshift Spectrum to join to data that is older than 13 months.
D. Take a snapshot of the Amazon Redshift cluste
E. Restore the cluster to a new cluster using dense storage nodes with additional storage capacity.
F. Execute a CREATE TABLE AS SELECT (CTAS) statement to move records that are older than 13 months to quarterly partitioned data in Amazon Redshift Spectrum backed by Amazon S3.
G. Unload all the tables in Amazon Redshift to an Amazon S3 bucket using S3 Intelligent-Tierin
H. Use AWS Glue to crawl the S3 bucket location to create external tables in an AWS Glue Data Catalo
I. Create an Amazon EMR cluster using Auto Scaling for any daily analytics needs, and use Amazon Athena for the quarterly reports, with both using the same AWS Glue Data Catalog.

Answer: A

NEW QUESTION 106
A healthcare company uses AWS data and analytics tools to collect, ingest, and store electronic health record (EHR) data about its patients. The raw EHR data is stored in Amazon S3 in JSON format partitioned by hour, day, and year and is updated every hour. The company wants to maintain the data catalog and metadata in an AWS Glue Data Catalog to be able to access the data using Amazon Athena or Amazon Redshift Spectrum for analytics.
When defining tables in the Data Catalog, the company has the following requirements:
Choose the catalog table name and do not rely on the catalog table naming algorithm. Keep the table updated with new partitions loaded in the respective S3 bucket prefixes.
Which solution meets these requirements with minimal effort?

A. Run an AWS Glue crawler that connects to one or more data stores, determines the data structures, and writes tables in the Data Catalog.
B. Use the AWS Glue console to manually create a table in the Data Catalog and schedule an AWS Lambda function to update the table partitions hourly.
C. Use the AWS Glue API CreateTable operation to create a table in the Data Catalo
D. Create an AWS Glue crawler and specify the table as the source.
E. Create an Apache Hive catalog in Amazon EMR with the table schema definition in Amazon S3, and update the table partition with a scheduled jo
F. Migrate the Hive catalog to the Data Catalog.

Answer: C

Explanation:
Updating Manually Created Data Catalog Tables Using Crawlers: To do this, when you define a crawler, instead of specifying one or more data stores as the source of a crawl, you specify one or more existing Data Catalog tables. The crawler then crawls the data stores specified by the catalog tables. In this case, no new tables are created; instead, your manually created tables are updated.

NEW QUESTION 108
A university intends to use Amazon Kinesis Data Firehose to collect JSON-formatted batches of water quality readings in Amazon S3. The readings are from 50 sensors scattered across a local lake. Students will query the stored data using Amazon Athena to observe changes in a captured metric over time, such as water temperature or acidity. Interest has grown in the study, prompting the university to reconsider how data will be stored.
Which data format and partitioning choices will MOST significantly reduce costs? (Choose two.)

A. Store the data in Apache Avro format using Snappy compression.

B. Partition the data by year, month, and day.
C. Store the data in Apache ORC format using no compression.
D. Store the data in Apache Parquet format using Snappy compression.
E. Partition the data by sensor, year, month, and day.

Answer: CD

NEW QUESTION 112
A company is planning to do a proof of concept for a machine learning (ML) project using Amazon SageMaker with a subset of existing on-premises data hosted in the company's 3 TB data warehouse. For part of the project, AWS Direct Connect is established and tested. To prepare the data for ML, data analysts are performing data curation. The data analysts want to perform multiple step, including mapping, dropping null fields, resolving choice, and splitting fields. The company needs the fastest solution to curate the data for this project.
Which solution meets these requirements?

A. Ingest data into Amazon S3 using AWS DataSync and use Apache Spark scrips to curate the data in an Amazon EMR cluste
B. Store the curated data in Amazon S3 for ML processing.
C. Create custom ETL jobs on-premises to curate the dat
D. Use AWS DMS to ingest data into Amazon S3 for ML processing.
E. Ingest data into Amazon S3 using AWS DM
F. Use AWS Glue to perform data curation and store the data in Amazon S3 for ML processing.
G. Take a full backup of the data store and ship the backup files using AWS Snowbal
H. Upload Snowball data into Amazon S3 and schedule data curation jobs using AWS Batch to prepare the data for ML.

Answer: C

NEW QUESTION 113
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

NEW QUESTION 117
A company has a marketing department and a finance department. The departments are storing data in Amazon S3 in their own AWS accounts in AWS Organizations. Both departments use AWS Lake Formation to catalog and secure their data. The departments have some databases and tables that share common names.
The marketing department needs to securely access some tables from the finance department. Which two steps are required for this process? (Choose two.)

A. The finance department grants Lake Formation permissions for the tables to the external account for the marketing department.
B. The finance department creates cross-account IAM permissions to the table for the marketing department role.
C. The marketing department creates an IAM role that has permissions to the Lake Formation tables.

Answer: AB

Explanation:
Granting Lake Formation Permissions Creating an IAM role (AWS CLI)

NEW QUESTION 122
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 connecto
B. Run the query from MySQL Workbench instead of Athena directly.
C. Use Athena to extract the data and store it in Apache Parquet format on a daily basi
D. Query the extracted data.
E. Run a daily AWS Glue ETL job to convert the data files to Apache Parquet and to partition the converted file
F. Create a periodic AWS Glue crawler to automatically crawl the partitioned data on a daily basis.
G. Run a daily AWS Glue ETL job to compress the data files by using the .gzip forma
H. Query the compressed data.
I. Run a daily AWS Glue ETL job to compress the data files by using the .lzo forma
J. Query the compressed data.

Answer: BC

NEW QUESTION 127
A financial services company needs to aggregate daily stock trade data from the exchanges into a data store.
The company requires that data be streamed directly into the data store, but also occasionally allows data to be modified using SQL. The solution should integrate complex, analytic queries running with minimal latency. The solution must provide a business intelligence dashboard that enables viewing of the top contributors to

anomalies in stock prices.
Which solution meets the company’s requirements?

A. Use Amazon Kinesis Data Firehose to stream data to Amazon S3. Use Amazon Athena as a data source for Amazon QuickSight to create a business intelligence dashboard.
B. Use Amazon Kinesis Data Streams to stream data to Amazon Redshif
C. Use Amazon Redshift as a data source for Amazon QuickSight to create a business intelligence dashboard.
D. Use Amazon Kinesis Data Firehose to stream data to Amazon Redshif
E. Use Amazon Redshift as a data source for Amazon QuickSight to create a business intelligence dashboard.
F. Use Amazon Kinesis Data Streams to stream data to Amazon S3. Use Amazon Athena as a data source for Amazon QuickSight to create a business intelligence dashboard.

Answer: C

NEW QUESTION 132
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 Redshif
C. Query the data with Amazon Redshift.
D. Query all the datasets in place with Apache Spark SQL running on an AWS Glue developer endpoint.
E. Query all the datasets in place with Apache Presto running on Amazon EMR.

Answer: C

NEW QUESTION 134
......