Exam Questions DAS-C01

AWS Certified Data Analytics - Specialty

https://www.2passeasy.com/dumps/DAS-C01/

NEW QUESTION 1
A company wants to run analytics on its Elastic Load Balancing logs stored in Amazon S3. A data analyst needs to be able to query all data from a desired year, month, or day. The data analyst should also be able to query a subset of the columns. The company requires minimal operational overhead and the most cost-effective solution.
Which approach meets these requirements for optimizing and querying the log data?

A. Use an AWS Glue job nightly to transform new log files into .csv format and partition by year, month, and da
B. Use AWS Glue crawlers to detect new partition
C. Use Amazon Athena to query data.
D. Launch a long-running Amazon EMR cluster that continuously transforms new log files from Amazon S3 into its Hadoop Distributed File System (HDFS) storage and partitions by year, month, and da
E. Use Apache Presto to query the optimized format.
F. Launch a transient Amazon EMR cluster nightly to transform new log files into Apache ORC format and partition by year, month, and da
G. Use Amazon Redshift Spectrum to query the data.
H. Use an AWS Glue job nightly to transform new log files into Apache Parquet format and partition by year, month, and da
I. Use AWS Glue crawlers to detect new partition
J. Use Amazon Athena to querydata.

Answer: C

NEW QUESTION 2
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 3
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 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 a data lake on AWS that ingests sources of data from multiple business units and uses Amazon Athena for queries. The storage layer is Amazon S3 using the AWS Glue Data Catalog. The company wants to make the data available to its data scientists and business analysts. However, the company first needs to manage data access for Athena based on user roles and responsibilities.
What should the company do to apply these access controls with the LEAST operational overhead?

A. Define security policy-based rules for the users and applications by role in AWS Lake Formation.
B. Define security policy-based rules for the users and applications by role in AWS Identity and Access Management (IAM).
C. Define security policy-based rules for the tables and columns by role in AWS Glue.
D. Define security policy-based rules for the tables and columns by role in AWS Identity and Access Management (IAM).

Answer: D

NEW QUESTION 7
A retail company's data analytics team recently created multiple product sales analysis dashboards for the average selling price per product using Amazon QuickSight. The dashboards were created from .csv files uploaded to Amazon S3. The team is now planning to share the dashboards with the respective external product owners by creating individual users in Amazon QuickSight. For compliance and governance reasons, restricting access is a key requirement. The product owners should view only their respective product analysis in the dashboard reports.
Which approach should the data analytics team take to allow product owners to view only their products in the dashboard?

A. Separate the data by product and use S3 bucket policies for authorization.
B. Separate the data by product and use IAM policies for authorization.
C. Create a manifest file with row-level security.

D. Create dataset rules with row-level security.

Answer: D

Explanation:
https://docs.aws.amazon.com/quicksight/latest/user/restrict-access-to-a-data-set-using-row-level-security.html

NEW QUESTION 8
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 9
An airline has been collecting metrics on flight activities for analytics. A recently completed proof of concept demonstrates how the company provides insights to data analysts to improve on-time departures. The proof of concept used objects in Amazon S3, which contained the metrics in .csv format, and used Amazon Athena for querying the data. As the amount of data increases, the data analyst wants to optimize the storage solution to improve query performance.
Which options should the data analyst use to improve performance as the data lake grows? (Choose three.)

A. Add a randomized string to the beginning of the keys in S3 to get more throughput across partitions.
B. Use an S3 bucket in the same account as Athena.
C. Compress the objects to reduce the data transfer I/O.
D. Use an S3 bucket in the same Region as Athena.
E. Preprocess the .csv data to JSON to reduce I/O by fetching only the document keys needed by the query.
F. Preprocess the .csv data to Apache Parquet to reduce I/O by fetching only the data blocks needed for predicates.

Answer: CDF

Explanation:
https://aws.amazon.com/blogs/big-data/top-10-performance-tuning-tips-for-amazon-athena/

NEW QUESTION 10
An ecommerce company stores customer purchase data in Amazon RDS. The company wants a solution to store and analyze historical data. The most recent 6 months of data will be queried frequently for analytics workloads. This data is several terabytes large. Once a month, historical data for the last 5 years must be accessible and will be joined with the more recent data. The company wants to optimize performance and cost.
Which storage solution will meet these requirements?

A. Create a read replica of the RDS database to store the most recent 6 months of dat
B. Copy the historical data into Amazon S3. Create an AWS Glue Data Catalog of the data in Amazon S3 and Amazon RD
C. Run historical queries using Amazon Athena.
D. Use an ETL tool to incrementally load the most recent 6 months of data into an Amazon Redshift cluste
E. Run more frequent queries against this cluste
F. Create a read replica of the RDS database to run queries on the historical data.
G. Incrementally copy data from Amazon RDS to Amazon S3. Create an AWS Glue Data Catalog of the data in Amazon S3. Use Amazon Athena to query the data.
H. Incrementally copy data from Amazon RDS to Amazon S3. Load and store the most recent 6 months of data in Amazon Redshif
I. Configure an Amazon Redshift Spectrum table to connect to all historical data.

Answer: D

NEW QUESTION 10
A financial company uses Amazon S3 as its data lake and has set up a data warehouse using a multi-node Amazon Redshift cluster. The data files in the data lake are organized in folders based on the data source of each data file. All the data files are loaded to one table in the Amazon Redshift cluster using a separate COPY command for each data file location. With this approach, loading all the data files into Amazon Redshift takes a long time to complete. Users want a faster solution with little or no increase in cost while maintaining the segregation of the data files in the S3 data lake.
Which solution meets these requirements?

A. Use Amazon EMR to copy all the data files into one folder and issue a COPY command to load the data into Amazon Redshift.
B. Load all the data files in parallel to Amazon Aurora, and run an AWS Glue job to load the data into Amazon Redshift.
C. Use an AWS Glue job to copy all the data files into one folder and issue a COPY command to load the data into Amazon Redshift.
D. Create a manifest file that contains the data file locations and issue a COPY command to load the data into Amazon Redshift.

Answer: D

Explanation:
https://docs.aws.amazon.com/redshift/latest/dg/loading-data-files-using-manifest.html "You can use a manifest to ensure that the COPY command loads all of the required files, and only the required files, for a data load"

NEW QUESTION 13
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 18
An IoT company wants to release a new device that will collect data to track sleep overnight on an intelligent mattress. Sensors will send data that will be uploaded to an Amazon S3 bucket. About 2 MB of data is generated each night for each bed. Data must be processed and summarized for each user, and the results need to be available as soon as possible. Part of the process consists of time windowing and other functions. Based on tests with a Python script, every run will require about 1 GB of memory and will complete within a couple of minutes.
Which solution will run the script in the MOST cost-effective way?

A. AWS Lambda with a Python script
B. AWS Glue with a Scala job
C. Amazon EMR with an Apache Spark script
D. AWS Glue with a PySpark job

Answer: A

NEW QUESTION 22
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 Routin
B. Enable VPC Flow Logs to monitor traffic.
C. Allow access to the Amazon Redshift database using AWS IAM onl
D. Log access using AWS CloudTrail.
E. Enable audit logging for Amazon Redshift using the AWS Management Console or the AWS CLI.
F. Enable and download audit reports from AWS Artifact.

Answer: C

NEW QUESTION 26
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 30
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 34
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 36
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 39
A media company is using Amazon QuickSight dashboards to visualize its national sales data. The dashboard is using a dataset with these fields: ID, date, time_zone, city, state, country, longitude, latitude, sales_volume, and number_of_items.
To modify ongoing campaigns, the company wants an interactive and intuitive visualization of which states across the country recorded a significantly lower sales volume compared to the national average.
Which addition to the company's QuickSight dashboard will meet this requirement?

A. A geospatial color-coded chart of sales volume data across the country.
B. A pivot table of sales volume data summed up at the state level.
C. A drill-down layer for state-level sales volume data.
D. A drill through to other dashboards containing state-level sales volume data.

Answer: B

NEW QUESTION 42
A company leverages Amazon Athena for ad-hoc queries against data stored in Amazon S3. The company wants to implement additional controls to separate query execution and query history among users, teams, or applications running in the same AWS account to comply with internal security policies.
Which solution meets these requirements?

A. Create an S3 bucket for each given use case, create an S3 bucket policy that grants permissions to appropriate individual IAM user
B. and apply the S3 bucket policy to the S3 bucket.
C. Create an Athena workgroup for each given use case, apply tags to the workgroup, and create an IAM policy using the tags to apply appropriate permissions to the workgroup.
D. Create an IAM role for each given use case, assign appropriate permissions to the role for the given use case, and add the role to associate the role with Athena.
E. Create an AWS Glue Data Catalog resource policy for each given use case that grants permissions to appropriate individual IAM users, and apply the resource policy to the specific tables used by Athena.

Answer: B

Explanation:

https://docs.aws.amazon.com/athena/latest/ug/user-created-workgroups.html
Amazon Athena Workgroups - A new resource type that can be used to separate query execution and query history between Users, Teams, or Applications running under the same AWS account https://aws.amazon.com/about-aws/whats-new/2019/02/athena_workgroups/

NEW QUESTION 44
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 49
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 53
A global pharmaceutical company receives test results for new drugs from various testing facilities worldwide. The results are sent in millions of 1 KB-sized JSON objects to an Amazon S3 bucket owned by the company. The data engineering team needs to process those files, convert them into Apache Parquet format, and load them into Amazon Redshift for data analysts to perform dashboard reporting. The engineering team uses AWS Glue to process the objects, AWS Step Functions for process orchestration, and Amazon CloudWatch for job scheduling.
More testing facilities were recently added, and the time to process files is increasing. What will MOST efficiently decrease the data processing time?

A. Use AWS Lambda to group the small files into larger file
B. Write the files back to Amazon S3. Process the files using AWS Glue and load them into Amazon Redshift tables.
C. Use the AWS Glue dynamic frame file grouping option while ingesting the raw input file
D. Process the files and load them into Amazon Redshift tables.
E. Use the Amazon Redshift COPY command to move the files from Amazon S3 into Amazon Redshift tables directl
F. Process the files in Amazon Redshift.
G. Use Amazon EMR instead of AWS Glue to group the small input file
H. Process the files in Amazon EMR and load them into Amazon Redshift tables.

Answer: A

NEW QUESTION 55
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 minute
C. The job inserts reference data into Amazon Redshift.
D. Send reference data to Amazon Kinesis Data Stream

E. Configure the Kinesis data stream to directly load the reference data into Amazon Redshift in real time.
F. Send the reference data to an Amazon Kinesis Data Firehose delivery strea
G. Configure Kinesis with a buffer interval of 60 seconds and to directly load the data into Amazon Redshift.

Answer: A

NEW QUESTION 57
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 58
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 60
A manufacturing company uses Amazon Connect to manage its contact center and Salesforce to manage its customer relationship management (CRM) data. The data engineering team must build a pipeline to ingest data from the contact center and CRM system into a data lake that is built on Amazon S3.
What is the MOST efficient way to collect data in the data lake with the LEAST operational overhead?

A. Use Amazon Kinesis Data Streams to ingest Amazon Connect data and Amazon AppFlow to ingest Salesforce data.
B. Use Amazon Kinesis Data Firehose to ingest Amazon Connect data and Amazon Kinesis Data Streams to ingest Salesforce data.
C. Use Amazon Kinesis Data Firehose to ingest Amazon Connect data and Amazon AppFlow to ingest Salesforce data.
D. Use Amazon AppFlow to ingest Amazon Connect data and Amazon Kinesis Data Firehose to ingest Salesforce data.

Answer: B

NEW QUESTION 62
A company analyzes its data in an Amazon Redshift data warehouse, which currently has a cluster of three dense storage nodes. Due to a recent business acquisition, the company needs to load an additional 4 TB of user data into Amazon Redshift. The engineering team will combine all the user data and apply complex calculations that require I/O intensive resources. The company needs to adjust the cluster's capacity to support the change in analytical and storage requirements.
Which solution meets these requirements?

A. Resize the cluster using elastic resize with dense compute nodes.
B. Resize the cluster using classic resize with dense compute nodes.
C. Resize the cluster using elastic resize with dense storage nodes.
D. Resize the cluster using classic resize with dense storage nodes.

Answer: C

NEW QUESTION 63
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 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 69
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 the 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 retrie
B. Decrease the timeout valu
C. Increase the job concurrency.
D. Keep the number of retries at 0. Decrease the timeout valu
E. Increase the job concurrency.
F. Keep the number of retries at 0. Decrease the timeout valu
G. Keep the job concurrency at 1.
H. Keep the number of retries at 0. Increase the timeout valu
I. Keep the job concurrency at 1.

Answer: B

NEW QUESTION 71
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 72
A financial company hosts a data lake in Amazon S3 and a data warehouse on an Amazon Redshift cluster. The company uses Amazon QuickSight to build dashboards and wants to secure access from its on-premises Active Directory to Amazon QuickSight.
How should the data be secured?

A. Use an Active Directory connector and single sign-on (SSO) in a corporate network environment.
B. Use a VPC endpoint to connect to Amazon S3 from Amazon QuickSight and an IAM role to authenticate Amazon Redshift.
C. Establish a secure connection by creating an S3 endpoint to connect Amazon QuickSight and a VPC endpoint to connect to Amazon Redshift.
D. Place Amazon QuickSight and Amazon Redshift in the security group and use an Amazon S3 endpoint to connect Amazon QuickSight to Amazon S3.

Answer: A

Explanation:
https://docs.aws.amazon.com/quicksight/latest/user/directory-integration.html

NEW QUESTION 74
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 79
A bank operates in a regulated environment. The compliance requirements for the country in which the bank operates say that customer data for each state should only be accessible by the bank's employees located in the same state. Bank employees in one state should NOT be able to access data for customers who have provided a home address in a different state.
The bank's marketing team has hired a data analyst to gather insights from customer data for a new campaign being launched in certain states. Currently, data linking each customer account to its home state is stored in a tabular .csv file within a single Amazon S3 folder in a private S3 bucket. The total size of the S3 folder is 2 GB uncompressed. Due to the country's compliance requirements, the marketing team is not able to access this folder.
The data analyst is responsible for ensuring that the marketing team gets one-time access to customer data for their campaign analytics project, while being subject to all the compliance requirements and controls.
Which solution should the data analyst implement to meet the desired requirements with the LEAST amount of setup effort?

A. Re-arrange data in Amazon S3 to store customer data about each state in a different S3 folder within the same bucke
B. Set up S3 bucket policies to provide marketing employees with appropriate data access under compliance control
C. Delete the bucket policies after the project.
D. Load tabular data from Amazon S3 to an Amazon EMR cluster using s3DistC
E. Implement a customHadoop-based row-level security solution on the Hadoop Distributed File System (HDFS) to provide marketing employees with appropriate data access under compliance control
F. Terminate the EMR cluster after the project.
G. Load tabular data from Amazon S3 to Amazon Redshift with the COPY comman
H. Use the built-in row- level security feature in Amazon Redshift to provide marketing employees with appropriate data access under compliance control
I. Delete the Amazon Redshift tables after the project.
J. Load tabular data from Amazon S3 to Amazon QuickSight Enterprise edition by directly importing it as a data sourc
K. Use the built-in row-level security feature in Amazon QuickSight to provide marketing employees with appropriate data access under compliance control
L. Delete Amazon QuickSight data sources after the project is complete.

Answer: C

NEW QUESTION 84
A company uses the Amazon Kinesis SDK to write data to Kinesis Data Streams. Compliance requirements state that the data must be encrypted at rest using a key that can be rotated. The company wants to meet this encryption requirement with minimal coding effort.
How can these requirements be met?

A. Create a customer master key (CMK) in AWS KM
B. Assign the CMK an alia
C. Use the AWS Encryption SDK, providing it with the key alias to encrypt and decrypt the data.
D. Create a customer master key (CMK) in AWS KM
E. Assign the CMK an alia
F. Enable server-side encryption on the Kinesis data stream using the CMK alias as the KMS master key.
G. Create a customer master key (CMK) in AWS KM
H. Create an AWS Lambda function to encrypt and decrypt the dat
I. Set the KMS key ID in the function's environment variables.
J. Enable server-side encryption on the Kinesis data stream using the default KMS key for Kinesis Data Streams.

Answer: B

NEW QUESTION 87
A streaming application is reading data from Amazon Kinesis Data Streams and immediately writing the data to an Amazon S3 bucket every 10 seconds. The application is reading data from hundreds of shards. The batch interval cannot be changed due to a separate requirement. The data is being accessed by Amazon Athena. Users are seeing degradation in query performance as time progresses.
Which action can help improve query performance?

A. Merge the files in Amazon S3 to form larger files.
B. Increase the number of shards in Kinesis Data Streams.
C. Add more memory and CPU capacity to the streaming application.
D. Write the files to multiple S3 buckets.

Answer: A

Explanation:
https://aws.amazon.com/blogs/big-data/top-10-performance-tuning-tips-for-amazon-athena/

NEW QUESTION 92
A banking company is currently using an Amazon Redshift cluster with dense storage (DS) nodes to store sensitive data. An audit found that the cluster is unencrypted. Compliance requirements state that a database with sensitive data must be encrypted through a hardware security module (HSM) with automated key rotation.
Which combination of steps is required to achieve compliance? (Choose two.)

A. Set up a trusted connection with HSM using a client and server certificate with automatic key rotation.
B. Modify the cluster with an HSM encryption option and automatic key rotation.
C. Create a new HSM-encrypted Amazon Redshift cluster and migrate the data to the new cluster.
D. Enable HSM with key rotation through the AWS CLI.
E. Enable Elliptic Curve Diffie-Hellman Ephemeral (ECDHE) encryption in the HSM.

Answer: BD

NEW QUESTION 93
A media content company has a streaming playback application. The company wants to collect and analyze the data to provide near-real-time feedback on playback issues. The company needs to consume this data and return results within 30 seconds according to the service-level agreement (SLA). The company needs the consumer to identify playback issues, such as quality during a specified timeframe. The data will be emitted as JSON and may change schemas over time.
Which solution will allow the company to collect data for processing while meeting these requirements?

A. Send the data to Amazon Kinesis Data Firehose with delivery to Amazon S3. Configure an S3 event trigger an AWS Lambda function to process the dat
B. The Lambda function will consume the data and process it to identify potential playback issue
C. Persist the raw data to Amazon S3.
D. Send the data to Amazon Managed Streaming for Kafka and configure an Amazon Kinesis Analytics for Java application as the consume
E. The application will consume the data and process it to identify potential playback issue
F. Persist the raw data to Amazon DynamoDB.
G. Send the data to Amazon Kinesis Data Firehose with delivery to Amazon S3. Configure Amazon S3 to trigger an event for AWS Lambda to proces
H. The Lambda function will consume the data and process it to identify potential playback issue
I. Persist the raw data to Amazon DynamoDB.
J. Send the data to Amazon Kinesis Data Streams and configure an Amazon Kinesis Analytics for Java application as the consume
K. The application will consume the data and process it to identify potential playback issue
L. Persist the raw data to Amazon S3.

Answer: D

Explanation:
https://aws.amazon.com/blogs/aws/new-amazon-kinesis-data-analytics-for-java/

NEW QUESTION 96
A company is streaming its high-volume billing data (100 MBps) to Amazon Kinesis Data Streams. A data analyst partitioned the data on account_id to ensure that all records belonging to an account go to the same Kinesis shard and order is maintained. While building a custom consumer using the Kinesis Java SDK, the data analyst notices that, sometimes, the messages arrive out of order for account_id. Upon further investigation, the data analyst discovers the messages that are out of order seem to be arriving from different shards for the same account_id and are seen when a stream resize runs.
What is an explanation for this behavior and what is the solution?

A. There are multiple shards in a stream and order needs to be maintained in the shar
B. The data analyst needs to make sure there is only a single shard in the stream and no stream resize runs.
C. The hash key generation process for the records is not working correctl
D. The data analyst should generate an explicit hash key on the producer side so the records are directed to the appropriate shard accurately.
E. The records are not being received by Kinesis Data Streams in orde
F. The producer should use the PutRecords API call instead of the PutRecord API call with the SequenceNumberForOrdering parameter.
G. The consumer is not processing the parent shard completely before processing the child shards after a stream resiz
H. The data analyst should process the parent shard completely first before processing the child shards.

Answer: D

Explanation:
https://docs.aws.amazon.com/streams/latest/dev/kinesis-using-sdk-java-after-resharding.html the parent shards that remain after the reshard could still contain data that you haven't read yet that was added to the stream before the reshard. If you read data from the child shards before having read all data from the parent shards, you could read data for a particular hash key out of the order given by the data records' sequence numbers. Therefore, assuming that the order of the data is important, you should, after a reshard, always continue to read data from the parent shards until it is exhausted. Only then should you begin reading data from the child shards.

NEW QUESTION 100
A software company hosts an application on AWS, and new features are released weekly. As part of the application testing process, a solution must be developed that analyzes logs from each Amazon EC2 instance to ensure that the application is working as expected after each deployment. The collection and analysis solution should be highly available with the ability to display new information with minimal delays.
Which method should the company use to collect and analyze the logs?

A. Enable detailed monitoring on Amazon EC2, use Amazon CloudWatch agent to store logs in Amazon S3, and use Amazon Athena for fast, interactive log analytics.
B. Use the Amazon Kinesis Producer Library (KPL) agent on Amazon EC2 to collect and send data to Kinesis Data Streams to further push the data to Amazon Elasticsearch Service and visualize using Amazon QuickSight.
C. Use the Amazon Kinesis Producer Library (KPL) agent on Amazon EC2 to collect and send data to Kinesis Data Firehose to further push the data to Amazon Elasticsearch Service and Kibana.
D. Use Amazon CloudWatch subscriptions to get access to a real-time feed of logs and have the logs delivered to Amazon Kinesis Data Streams to further push the data to Amazon Elasticsearch Service and Kibana.

Answer: D

NEW QUESTION 102
An online retailer is rebuilding its inventory management system and inventory reordering system to automatically reorder products by using Amazon Kinesis Data Streams. The inventory management system uses the Kinesis Producer Library (KPL) to publish data to a stream. The inventory reordering system uses the Kinesis Client Library (KCL) to consume data from the stream. The stream has been configured to scale as needed. Just before production deployment, the retailer discovers that the inventory reordering system is receiving duplicated data.
Which factors could be causing the duplicated data? (Choose two.)

A. The producer has a network-related timeout.
B. The stream's value for the IteratorAgeMilliseconds metric is too high.
C. There was a change in the number of shards, record processors, or both.
D. The AggregationEnabled configuration property was set to true.
E. The max_records configuration property was set to a number that is too high.

Answer: BD

NEW QUESTION 103
A marketing company is storing its campaign response data in Amazon S3. A consistent set of sources has generated the data for each campaign. The data is saved into Amazon S3 as .csv files. A business analyst will use Amazon Athena to analyze each campaign's data. The company needs the cost of ongoing data analysis with Athena to be minimized.
Which combination of actions should a data analytics specialist take to meet these requirements? (Choose two.)

A. Convert the .csv files to Apache Parquet.
B. Convert the .csv files to Apache Avro.
C. Partition the data by campaign.
D. Partition the data by source.
E. Compress the .csv files.

Answer: AC

Explanation:
https://aws.amazon.com/blogs/big-data/top-10-performance-tuning-tips-for-amazon-athena/

NEW QUESTION 108
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 111
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 112
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 HDFSconnector 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

NEW QUESTION 116
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 119
A company has an encrypted Amazon Redshift cluster. The company recently enabled Amazon Redshift audit logs and needs to ensure that the audit logs are also encrypted at rest. The logs are retained for 1 year. The auditor queries the logs once a month.
What is the MOST cost-effective way to meet these requirements?

A. Encrypt the Amazon S3 bucket where the logs are stored by using AWS Key Management Service (AWS KMS). Copy the data into the Amazon Redshift cluster from Amazon S3 on a daily basi
B. Query the data as required.
C. Disable encryption on the Amazon Redshift cluster, configure audit logging, and encrypt the Amazon Redshift cluste
D. Use Amazon Redshift Spectrum to query the data as required.
E. Enable default encryption on the Amazon S3 bucket where the logs are stored by using AES-256 encryptio
F. Copy the data into the Amazon Redshift cluster from Amazon S3 on a daily basi
G. Query the data as required.
H. Enable default encryption on the Amazon S3 bucket where the logs are stored by using AES-256 encryptio
I. Use Amazon Redshift Spectrum to query the data as required.

Answer: A

NEW QUESTION 124
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 127
A marketing company wants to improve its reporting and business intelligence capabilities. During the planning phase, the company interviewed the relevant stakeholders and discovered that:
 The operations team reports are run hourly for the current month's data.
 The sales team wants to use multiple Amazon QuickSight dashboards to show a rolling view of the last 30 days based on several categories.
 The sales team also wants to view the data as soon as it reaches the reporting backend.
 The finance team's reports are run daily for last month's data and once a month for the last 24 months of data.
Currently, there is 400 TB of data in the system with an expected additional 100 TB added every month. The company is looking for a solution that is as cost-effective as possible.
Which solution meets the company's requirements?

A. Store the last 24 months of data in Amazon Redshif
B. Configure Amazon QuickSight with Amazon Redshift as the data source.
C. Store the last 2 months of data in Amazon Redshift and the rest of the months in Amazon S3. Set up an external schema and table for Amazon Redshift Spectru
D. Configure Amazon QuickSight with Amazon Redshift as the data source.
E. Store the last 24 months of data in Amazon S3 and query it using Amazon Redshift Spectrum.Configure Amazon QuickSight with Amazon Redshift Spectrum as

the data source.
F. Store the last 2 months of data in Amazon Redshift and the rest of the months in Amazon S3. Use a long- running Amazon EMR with Apache Spark cluster to query the data as neede
G. Configure Amazon QuickSight with Amazon EMR as the data source.

Answer: B

NEW QUESTION 129
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 131
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 133
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 134
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 136
......