Amazon-Web-Services

Exam Questions DAS-C01

AWS Certified Data Analytics - Specialty

NEW QUESTION 1
A hospital uses an electronic health records (EHR) system to collect two types of data
• Patient information, which includes a patient's name and address
• Diagnostic tests conducted and the results of these tests
Patient information is expected to change periodically Existing diagnostic test data never changes and only new records are added
The hospital runs an Amazon Redshift cluster with four dc2.large nodes and wants to automate the ingestion of the patient information and diagnostic test data into respective Amazon Redshift tables for analysis The EHR system exports data as CSV files to an Amazon S3 bucket on a daily basis Two sets of CSV files are generated One set of files is for patient information with updates, deletes, and inserts The other set of files is for new diagnostic test data only
What is the MOST cost-effective solution to meet these requirements?

A. Use Amazon EMR with Apache Hud
B. Run daily ETL jobs using Apache Spark and the Amazon Redshift JDBC driver
C. Use an AWS Glue crawler to catalog the data in Amazon S3 Use Amazon Redshift Spectrum to perform scheduled queries of the data in Amazon S3 and ingest the data into the patient information table and the diagnostic tests table.
D. Use an AWS Lambda function to run a COPY command that appends new diagnostic test data to the diagnostic tests table Run another COPY command to load the patient information data into the staging tables Use a stored procedure to handle create update, and delete operations for the patient information table
E. Use AWS Database Migration Service (AWS DMS) to collect and process change data capture (CDC) records Use the COPY command to load patient information data into the staging table
F. Use a stored procedure to handle create, update and delete operations for the patient information table

Answer: B

NEW QUESTION 2
A market data company aggregates external data sources to create a detailed view of product consumption in different countries. The company wants to sell this data to external parties through a subscription. To achieve this goal, the company needs to make its data securely available to external parties who are also AWS users.
What should the company do to meet these requirements with the LEAST operational overhead?

A. Store the data in Amazon S3. Share the data by using presigned URLs for security.
B. Store the data in Amazon S3. Share the data by using S3 bucket ACLs.
C. Upload the data to AWS Data Exchange for storag
D. Share the data by using presigned URLs for security.
E. Upload the data to AWS Data Exchange for storag
F. Share the data by using the AWS Data Exchange sharing wizard.

Answer: A

NEW QUESTION 3
A company hosts an Apache Flink application on premises. The application processes data from several Apache Kafka clusters. The data originates from a variety of sources, such as web applications mobile apps and operational databases The company has migrated some of these sources to AWS and now wants to migrate the Flink application. The company must ensure that data that resides in databases within the VPC does not traverse the internet The application must be able to process all the data that comes from the company's AWS solution, on-premises resources and the public internet
Which solution will meet these requirements with the LEAST operational overhead?

A. Implement Flink on Amazon EC2 within the company's VPC Create Amazon Managed Streaming for Apache Kafka (Amazon MSK) clusters in the VPC to collect data that comes from applications and databases within the VPC Use Amazon Kinesis Data Streams to collect data that comes from the public internet Configure Flink to have sources from Kinesis Data Streams Amazon MSK and any on-premises Kafka clusters by using AWS Client VPN or AWS Direct Connect
B. Implement Flink on Amazon EC2 within the company's VPC Use Amazon Kinesis Data Streams to collect data that comes from applications and databases within the VPC and the public internet Configure Flink to have sources from Kinesis Data Streams and any on-premises Kafka clusters by using AWS Client VPN or AWS Direct Connect
C. Create an Amazon Kinesis Data Analytics application by uploading the compiled Flink jar file Use Amazon Kinesis Data Streams to collect data that comes from applications and databases within the VPC and the public internet Configure the Kinesis Data Analytics application to have sources from Kinesis Data Streams and any on-premises Kafka clusters by using AWS Client VPN or AWS Direct Connect
D. Create an Amazon Kinesis Data Analytics application by uploading the compiled Flink jar file Create Amazon Managed Streaming for Apache Kafka (Amazon MSK) clusters in the company's VPC to collect data that comes from applications and databases within the VPC Use Amazon Kinesis Data Streams to collect data that comes from the public internet Configure the Kinesis Data Analytics application to have sources from Kinesis Data Stream
E. Amazon MSK and any on-premises Kafka clusters by using AWS Client VPN or AWS Direct Connect

Answer: D

NEW QUESTION 4
A company has several Amazon EC2 instances sitting behind an Application Load Balancer (ALB) The company wants its IT Infrastructure team to analyze the IP addresses coming into the company's ALB The ALB is configured to store access logs in Amazon S3 The access logs create about 1 TB of data each day, and access to the data will be infrequent The company needs a solution that is scalable, cost-effective and has minimal maintenance requirements
Which solution meets these requirements?

A. Copy the data into Amazon Redshift and query the data
B. Use Amazon EMR and Apache Hive to query the S3 data
C. Use Amazon Athena to query the S3 data
D. Use Amazon Redshift Spectrum to query the S3 data

Answer: D

NEW QUESTION 5
Three teams of data analysts use Apache Hive on an Amazon EMR cluster with the EMR File System (EMRFS) to query data stored within each teams Amazon S3 bucket. The EMR cluster has Kerberos enabled and is configured to authenticate users from the corporate Active Directory. The data is highly sensitive, so access must be limited to the members of each team.

Which steps will satisfy the security requirements?

A. For the EMR cluster Amazon EC2 instances, create a service role that grants no access to Amazon S3.Create three additional IAM roles, each granting access to each team's specific bucke
B. Add the additional IAM roles to the cluster's EMR role for the EC2 trust polic
C. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.
D. For the EMR cluster Amazon EC2 instances, create a service role that grants no access to Amazon S3.Create three additional IAM roles, each granting access to each team's specific bucke
E. Add the service role for the EMR cluster EC2 instances to the trust policies for the additional IAM role
F. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.
G. For the EMR cluster Amazon EC2 instances, create a service role that grants full access to Amazon S3.Create three additional IAM roles, each granting access to each team's specific bucke
H. Add the service role for the EMR cluster EC2 instances to the trust polices for the additional IAM role
I. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.
J. For the EMR cluster Amazon EC2 instances, create a service role that grants full access to Amazon S3.Create three additional IAM roles, each granting access to each team's specific bucke
K. Add the service role for the EMR cluster EC2 instances to the trust polices for the base IAM role
L. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.

Answer: C

NEW QUESTION 6
A software company wants to use instrumentation data to detect and resolve errors to improve application recovery time. The company requires API usage anomalies, like error rate and response time spikes, to be detected in near-real time (NRT) The company also requires that data analysts have access to dashboards for log analysis in NRT
Which solution meets these requirements'?

A. Use Amazon Kinesis Data Firehose as the data transport layer for logging data Use Amazon Kinesis Data Analytics to uncover the NRT API usage anomalies Use Kinesis Data Firehose to deliver log data to Amazon OpenSearch Service (Amazon Elasticsearch Service) for search, log analytics, and application monitoring Use OpenSearch Dashboards (Kibana) in Amazon OpenSearch Service (Amazon Elasticsearch Service) for the dashboards.
B. Use Amazon Kinesis Data Analytics as the data transport layer for logging dat
C. Use Amazon Kinesis Data Streams to uncover NRT monitoring metric
D. Use Amazon Kinesis Data Firehose to deliver log data to Amazon OpenSearch Service (Amazon Elasticsearch Service) for search, log analytics, and application monitoring Use Amazon QuickSight for the dashboards
E. Use Amazon Kinesis Data Analytics as the data transport layer for logging data and to uncover NRT monitoring metrics Use Amazon Kinesis Data Firehose to deliver log data to Amazon OpenSearch Service (Amazon Elasticsearch Service) for search, log analytics, and application monitoring Use OpenSearch Dashboards (Kibana) in Amazon OpenSearch Service (Amazon Elasticsearch Service) for the dashboards
F. Use Amazon Kinesis Data Firehose as the data transport layer for logging data Use Amazon Kinesis Data Analytics to uncover NRT monitoring metrics Use Amazon Kinesis Data Streams to deliver logdata to Amazon OpenSearch Service (Amazon Elasticsearch Service) for search, log analytics, and application monitoring Use Amazon QuickSight for the dashboards.

Answer: C

NEW QUESTION 7
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 8
A company uses Amazon kinesis Data Streams to ingest and process customer behavior information from application users each day. A data analytics specialist notices that its data stream is throttling. The specialist has turned on enhanced monitoring for the Kinesis data stream and has verified that the data stream did not exceed the data limits. The specialist discovers that there are hot shards
Which solution will resolve this issue?

A. Use a random partition key to ingest the records.
B. Increase the number of shards Split the size of the log records.
C. Limit the number of records that are sent each second by the producer to match the capacity of the stream.
D. Decrease the size of the records that are sent from the producer to match the capacity of the stream.

Answer: A

NEW QUESTION 9
A company is building an analytical solution that includes Amazon S3 as data lake storage and Amazon Redshift for data warehousing. The company wants to use Amazon Redshift Spectrum to query the data that is stored in Amazon S3.
Which steps should the company take to improve performance when the company uses Amazon Redshift Spectrum to query the S3 data files? (Select THREE)
Use gzip compression with individual file sizes of 1-5 GB

A. Use a columnar storage file format
B. Partition the data based on the most common query predicates
C. Split the data into KB-sized files.
D. Keep all files about the same size.

E. Use file formats that are not splittable

Answer: BCD

NEW QUESTION 10
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 10
A company wants to optimize the cost of its data and analytics platform. The company is ingesting a number of .c sv and JSON files in Amazon S3 from various data sources. Incoming data is expected to be 50 GB each day. The company is using Amazon Athena to query the raw data in Amazon S3 directly. Most queries aggregate data from the past 12 months, and data that is older than 5 years is infrequently queried. The typical query scans about 500 MB of data and is expected to return results in less than 1 minute. The raw data must be retained indefinitely for compliance requirements.
Which solution meets the company's requirements?

A. Use an AWS Glue ETL job to compress, partition, and convert the data into a columnar data forma
B. Use Athena to query the processed datase
C. Configure a lifecycle policy to move the processed data into the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class 5 years after object creation.Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after object creation.
D. Use an AWS Glue ETL job to partition and convert the data into a row-based data forma
E. Use Athena to query the processed datase
F. Configure a lifecycle policy to move the data into the Amazon S3 Standard- Infrequent Access (S3 Standard-IA) storage class 5 years after object creatio
G. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after object creation.
H. Use an AWS Glue ETL job to compress, partition, and convert the data into a columnar data forma
I. Use Athena to query the processed datase
J. Configure a lifecycle policy to move the processed data into the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class 5 years after the object was last accesse
K. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier forlong-term archival 7 days after the last date the object was accessed.
L. Use an AWS Glue ETL job to partition and convert the data into a row-based data forma
M. Use Athena to query the processed datase
N. Configure a lifecycle policy to move the data into the Amazon S3 Standard- Infrequent Access (S3 Standard-IA) storage class 5 years after the object was last accesse
O. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after the last date the object was accessed.

Answer: A

NEW QUESTION 12
A team of data scientists plans to analyze market trend data for their company's new investment strategy. The trend data comes from five different data sources in large volumes. The team wants to utilize Amazon Kinesis to support their use case. The team uses SQL-like queries to analyze trends and wants to send notifications based on certain significant patterns in the trends. Additionally, the data scientists want to save the data to Amazon S3 for archival and historical re-processing, and use AWS managed services wherever possible. The team wants to implement the lowest-cost solution.
Which solution meets these requirements?

A. Publish data to one Kinesis data strea
B. Deploy a custom application using the Kinesis Client Library (KCL) for analyzing trends, and send notifications using Amazon SN
C. Configure Kinesis Data Firehose on the Kinesis data stream to persist data to an S3 bucket.
D. Publish data to one Kinesis data strea
E. Deploy Kinesis Data Analytic to the stream for analyzing trends, and configure an AWS Lambda function as an output to send notifications using Amazon SN
F. Configure Kinesis Data Firehose on the Kinesis data stream to persist data to an S3 bucket.
G. Publish data to two Kinesis data stream
H. Deploy Kinesis Data Analytics to the first stream for analyzing trends, and configure an AWS Lambda function as an output to send notifications using Amazon SN
I. Configure Kinesis Data Firehose on the second Kinesis data stream to persist data to an S3 bucket.
J. Publish data to two Kinesis data stream
K. Deploy a custom application using the Kinesis Client Library (KCL) to the first stream for analyzing trends, and send notifications using Amazon SN
L. Configure Kinesis Data Firehose on the second Kinesis data stream to persist data to an S3 bucket.

Answer: B

NEW QUESTION 13
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 14
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 18
A company stores its sales and marketing data that includes personally identifiable information (PII) in Amazon S3. The company allows its analysts to launch their own Amazon EMR cluster and run analytics reports with the data. To meet compliance requirements, the company must ensure the data is not publicly accessible throughout this process. A data engineer has secured Amazon S3 but must ensure the individual EMR clusters created by the analysts are not exposed to the public internet.
Which solution should the data engineer to meet this compliance requirement with LEAST amount of effort?

A. Create an EMR security configuration and ensure the security configuration is associated with the EMR clusters when they are created.
B. Check the security group of the EMR clusters regularly to ensure it does not allow inbound traffic from IPv4 0.0.0.0/0 or IPv6 ::/0.
C. Enable the block public access setting for Amazon EMR at the account level before any EMR cluster is created.
D. Use AWS WAF to block public internet access to the EMR clusters across the board.

Answer: C

Explanation:
https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-block-public-access.html

NEW QUESTION 22
A bank wants to migrate a Teradata data warehouse to the AWS Cloud The bank needs a solution for reading large amounts of data and requires the highest possible performance. The solution also must maintain the separation of storage and compute
Which solution meets these requirements?

A. Use Amazon Athena to query the data in Amazon S3
B. Use Amazon Redshift with dense compute nodes to query the data in Amazon Redshift managed storage
C. Use Amazon Redshift with RA3 nodes to query the data in Amazon Redshift managed storage
D. Use PrestoDB on Amazon EMR to query the data in Amazon S3

Answer: C

NEW QUESTION 26
A company recently created a test AWS account to use for a development environment The company also created a production AWS account in another AWS Region As part of its security testing the company wants to send log data from Amazon CloudWatch Logs in its production account to an Amazon Kinesis data stream in its test account
Which solution will allow the company to accomplish this goal?

A. Create a subscription filter in the production accounts CloudWatch Logs to target the Kinesis data stream in the test account as its destination In the test account create an 1AM role that grants access to the Kinesis data stream and the CloudWatch Logs resources in the production account
B. In the test account create an 1AM role that grants access to the Kinesis data stream and the CloudWatch Logs resources in the production account Create a destination data stream in Kinesis Data Streams in the test account with an 1AM role and a trust policy that allow CloudWatch Logs in the production account to write to the test account
C. In the test account, create an 1AM role that grants access to the Kinesis data stream and the CloudWatch Logs resources in the production account Create a destination data stream in Kinesis Data Streams in the test account with an 1AM role and a trust policy that allow CloudWatch Logs in the production account to

write to the test account
D. Create a destination data stream in Kinesis Data Streams in the test account with an 1AM role and a trust policy that allow CloudWatch Logs in the production account to write to the test account Create a subscription filter in the production accounts CloudWatch Logs to target the Kinesis data stream in the test account as its destination

Answer: D

NEW QUESTION 28
A retail company has 15 stores across 6 cities in the United States. Once a month, the sales team requests a visualization in Amazon QuickSight that provides the ability to easily identify revenue trends across cities and stores. The visualization also helps identify outliers that need to be examined with further analysis.
Which visual type in QuickSight meets the sales team's requirements?

A. Geospatial chart
B. Line chart
C. Heat map
D. Tree map

Answer: A

NEW QUESTION 31
An operations team notices that a few AWS Glue jobs for a given ETL application are failing. The AWS Glue jobs read a large number of small JSON files from an Amazon S3 bucket and write the data to a different S3 bucket in Apache Parquet format with no major transformations. Upon initial investigation, a data engineer notices the following error message in the History tab on the AWS Glue console: “Command Failed with Exit Code 1.”
Upon further investigation, the data engineer notices that the driver memory profile of the failed jobs crosses the safe threshold of 50% usage quickly and reaches 90–95% soon after. The average memory usage across all executors continues to be less than 4%.
The data engineer also notices the following error while examining the related Amazon CloudWatch Logs. What should the data engineer do to solve the failure in the MOST cost-effective way?

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

Answer: B

Explanation:
https://docs.aws.amazon.com/glue/latest/dg/monitor-profile-debug-oom-abnormalities.html#monitor-debug-oom

NEW QUESTION 32
A medical company has a system with sensor devices that read metrics and send them in real time to an Amazon Kinesis data stream. The Kinesis data stream has multiple shards. The company needs to calculate the average value of a numeric metric every second and set an alarm for whenever the value is above one threshold or below another threshold. The alarm must be sent to Amazon Simple Notification Service (Amazon SNS) in less than 30 seconds.
Which architecture meets these requirements?

A. Use an Amazon Kinesis Data Firehose delivery stream to read the data from the Kinesis data stream with an AWS Lambda transformation function that calculates the average per second and sends the alarm to Amazon SNS.
B. Use an AWS Lambda function to read from the Kinesis data stream to calculate the average per second and sent the alarm to Amazon SNS.
C. Use an Amazon Kinesis Data Firehose deliver stream to read the data from the Kinesis data stream and store it on Amazon S3. Have Amazon S3 trigger an AWS Lambda function that calculates the average per second and sends the alarm to Amazon SNS.
D. Use an Amazon Kinesis Data Analytics application to read from the Kinesis data stream and calculate the average per secon
E. Send the results to an AWS Lambda function that sends the alarm to Amazon SNS.

Answer: D

NEW QUESTION 36
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 37

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 42
A gaming company is collecting cllckstream data into multiple Amazon Kinesis data streams. The company uses Amazon Kinesis Data Firehose delivery streams to store the data in JSON format in Amazon S3 Data scientists use Amazon Athena to query the most recent data and derive business insights. The company wants to reduce its Athena costs without having to recreate the data pipeline. The company prefers a solution that will require less management effort
Which set of actions can the data scientists take immediately to reduce costs?

A. Change the Kinesis Data Firehose output format to Apache Parquet Provide a custom S3 object YYYYMMDD prefix expression and specify a large buffer size For the existing data, run an AWS Glue ETL job to combine and convert small JSON files to large Parquet files and add the YYYYMMDD prefix Use ALTER TABLE ADD PARTITION to reflect the partition on the existing Athena table.
B. Create an Apache Spark Job that combines and converts JSON files to Apache Parquet files Launch an Amazon EMR ephemeral cluster daily to run the Spark job to create new Parquet files in a different S3 location Use ALTER TABLE SET LOCATION to reflect the new S3 location on the existing Athena table.
C. Create a Kinesis data stream as a delivery target for Kinesis Data Firehose Run Apache Flink on Amazon Kinesis Data Analytics on the stream to read the streaming data, aggregate ikand save it to Amazon S3 in Apache Parquet format with a custom S3 object YYYYMMDD prefix Use ALTER TABLE ADD PARTITION to reflect the partition on the existing Athena table
D. Integrate an AWS Lambda function with Kinesis Data Firehose to convert source records to Apache Parquet and write them to Amazon S3 In parallel, run an AWS Glue ETL job to combine and convert existing JSON files to large Parquet files Create a custom S3 object YYYYMMDD prefix Use ALTER TABLE ADD PARTITION to reflect the partition on the existing Athena table.

Answer: D

NEW QUESTION 46
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 Catalog.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 48
A company using Amazon QuickSight Enterprise edition has thousands of dashboards analyses and datasets. The company struggles to manage and assign permissions for granting users access to various items within QuickSight. The company wants to make it easier to implement sharing and permissions management.
Which solution should the company implement to simplify permissions management?

A. Use QuickSight folders to organize dashboards, analyses, and datasets Assign individual users permissions to these folders
B. Use QuickSight folders to organize dashboards analyses, and datasets Assign group permissions by using these folders.
C. Use AWS 1AM resource-based policies to assign group permissions to QuickSight items
D. Use QuickSight user management APIs to provision group permissions based on dashboard naming conventions

Answer: C

NEW QUESTION 51
A company with a video streaming website wants to analyze user behavior to make recommendations to users in real time Clickstream data is being sent to Amazon Kinesis Data Streams and reference data is stored in Amazon S3 The company wants a solution that can use standard SQL quenes The solution must also provide a way to look up pre-calculated reference data while making recommendations
Which solution meets these requirements?

A. Use an AWS Glue Python shell job to process incoming data from Kinesis Data Streams Use the Boto3 library to write data to Amazon Redshift
B. Use AWS Glue streaming and Scale to process incoming data from Kinesis Data Streams Use the AWS Glue connector to write data to Amazon Redshift

C. Use Amazon Kinesis Data Analytics to create an in-application table based upon the reference data Process incoming data from Kinesis Data Streams Use a data stream to write results to Amazon Redshift
D. Use Amazon Kinesis Data Analytics to create an in-application table based upon the reference dataProcess incoming data from Kinesis Data Streams Use an Amazon Kinesis Data Firehose delivery stream to write results to Amazon Redshift

Answer: D

NEW QUESTION 55
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 60
An online retail company with millions of users around the globe wants to improve its ecommerce analytics capabilities. Currently, clickstream data is uploaded directly to Amazon S3 as compressed files. Several times each day, an application running on Amazon EC2 processes the data and makes search options and reports available for visualization by editors and marketers. The company wants to make website clicks and aggregated data available to editors and marketers in minutes to enable them to connect with users more effectively.
Which options will help meet these requirements in the MOST efficient way? (Choose two.)

A. Use Amazon Kinesis Data Firehose to upload compressed and batched clickstream records to Amazon Elasticsearch Service.
B. Upload clickstream records to Amazon S3 as compressed file
C. Then use AWS Lambda to send data to Amazon Elasticsearch Service from Amazon S3.
D. Use Amazon Elasticsearch Service deployed on Amazon EC2 to aggregate, filter, and process the data.Refresh content performance dashboards in near-real time.
E. Use Kibana to aggregate, filter, and visualize the data stored in Amazon Elasticsearch Servic
F. Refresh content performance dashboards in near-real time.
G. Upload clickstream records from Amazon S3 to Amazon Kinesis Data Streams and use a Kinesis Data Streams consumer to send records to Amazon Elasticsearch Service.

Answer: AD

NEW QUESTION 64
A company wants to enrich application logs in near-real-time and use the enriched dataset for further analysis. The application is running on Amazon EC2 instances across multiple Availability Zones and storing its logs using Amazon CloudWatch Logs. The enrichment source is stored in an Amazon DynamoDB table.
Which solution meets the requirements for the event collection and enrichment?

A. Use a CloudWatch Logs subscription to send the data to Amazon Kinesis Data Firehos
B. Use AWS Lambda to transform the data in the Kinesis Data Firehose delivery stream and enrich it with the data inthe DynamoDB tabl
C. Configure Amazon S3 as the Kinesis Data Firehose delivery destination.
D. Export the raw logs to Amazon S3 on an hourly basis using the AWS CL
E. Use AWS Glue crawlers to catalog the log
F. Set up an AWS Glue connection for the DynamoDB table and set up an AWS Glue ETL job to enrich the dat
G. Store the enriched data in Amazon S3.
H. Configure the application to write the logs locally and use Amazon Kinesis Agent to send the data to Amazon Kinesis Data Stream
I. Configure a Kinesis Data Analytics SQL application with the Kinesis data stream as the sourc
J. Join the SQL application input stream with DynamoDB records, and then store the enriched output stream in Amazon S3 using Amazon Kinesis Data Firehose.
K. Export the raw logs to Amazon S3 on an hourly basis using the AWS CL
L. Use Apache Spark SQL on Amazon EMR to read the logs from Amazon S3 and enrich the records with the data from DynamoD
M. Store the enriched data in Amazon S3.

Answer: A

Explanation:
https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/SubscriptionFilters.html#FirehoseExample

NEW QUESTION 66
A company that monitors weather conditions from remote construction sites is setting up a solution to collect temperature data from the following two weather stations.

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

These weather stations were placed by onsite subject-matter experts.
Each sensor has a unique ID. The data collected from each sensor will be collected using Amazon Kinesis Data Streams.
Based on the total incoming and outgoing data throughput, a single Amazon Kinesis data stream with two shards is created. Two partition keys are created based

on the station names. During testing, there is a bottleneck on data coming from Station A, but not from Station B. Upon review, it is confirmed that the total stream throughput is still less than the allocated Kinesis Data Streams throughput.
How can this bottleneck be resolved without increasing the overall cost and complexity of the solution, while retaining the data collection quality requirements?

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

Answer: C

Explanation:
https://docs.aws.amazon.com/streams/latest/dev/kinesis-using-sdk-java-resharding.html
"Splitting increases the number of shards in your stream and therefore increases the data capacity of the stream. Because you are charged on a per-shard basis, splitting increases the cost of your stream"

NEW QUESTION 68
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 71
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 75
A retail company is building its data warehouse solution using Amazon Redshift. As a part of that effort, the company is loading hundreds of files into the fact table created in its Amazon Redshift cluster. The company wants the solution to achieve the highest throughput and optimally use cluster resources when loading data into the company's fact table.
How should the company meet these requirements?

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

Answer: D

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

NEW QUESTION 76
A company uses Amazon Elasticsearch Service (Amazon ES) to store and analyze its website clickstream data. The company ingests 1 TB of data daily using Amazon Kinesis Data Firehose and stores one day's worth of data in an Amazon ES cluster.
The company has very slow query performance on the Amazon ES index and occasionally sees errors from Kinesis Data Firehose when attempting to write to the index. The Amazon ES cluster has 10 nodes running a single index and 3 dedicated master nodes. Each data node has 1.5 TB of Amazon EBS storage attached and the cluster is configured with 1,000 shards. Occasionally, JVMMemoryPressure errors are found in the cluster logs.
Which solution will improve the performance of Amazon ES?

A. Increase the memory of the Amazon ES master nodes.
B. Decrease the number of Amazon ES data nodes.

C. Decrease the number of Amazon ES shards for the index.
D. Increase the number of Amazon ES shards for the index.

Answer: C

Explanation:
https://aws.amazon.com/premiumsupport/knowledge-center/high-jvm-memory-pressure-elasticsearch/

NEW QUESTION 79
A mobile gaming company wants to capture data from its gaming app and make the data available for analysis immediately. The data record size will be approximately 20 KB. The company is concerned about achieving optimal throughput from each device. Additionally, the company wants to develop a data stream processing application with dedicated throughput for each consumer.
Which solution would achieve this goal?

A. Have the app call the PutRecords API to send data to Amazon Kinesis Data Stream
B. Use the enhanced fan-out feature while consuming the data.
C. Have the app call the PutRecordBatch API to send data to Amazon Kinesis Data Firehos
D. Submit a support case to enable dedicated throughput on the account.
E. Have the app use Amazon Kinesis Producer Library (KPL) to send data to Kinesis Data Firehos
F. Use the enhanced fan-out feature while consuming the data.
G. Have the app call the PutRecords API to send data to Amazon Kinesis Data Stream
H. Host the stream- processing application on Amazon EC2 with Auto Scaling.

Answer: A

Explanation:
https://docs.aws.amazon.com/streams/latest/dev/enhanced-consumers.html

NEW QUESTION 84
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 86
A company launched a service that produces millions of messages every day and uses Amazon Kinesis Data Streams as the streaming service.
The company uses the Kinesis SDK to write data to Kinesis Data Streams. A few months after launch, a data analyst found that write performance is significantly reduced. The data analyst investigated the metrics and determined that Kinesis is throttling the write requests. The data analyst wants to address this issue without significant changes to the architecture.
Which actions should the data analyst take to resolve this issue? (Choose two.)

A. Increase the Kinesis Data Streams retention period to reduce throttling.
B. Replace the Kinesis API-based data ingestion mechanism with Kinesis Agent.
C. Increase the number of shards in the stream using the UpdateShardCount API.
D. Choose partition keys in a way that results in a uniform record distribution across shards.
E. Customize the application code to include retry logic to improve performance.

Answer: CD

Explanation:
https://aws.amazon.com/blogs/big-data/under-the-hood-scaling-your-kinesis-data-streams/

NEW QUESTION 90
A company has a data warehouse in Amazon Redshift that is approximately 500 TB in size. New data is imported every few hours and read-only queries are run throughout the day and evening. There is a particularly heavy load with no writes for several hours each morning on business days. During those hours, some queries are queued and take a long time to execute. The company needs to optimize query execution and avoid any downtime.
What is the MOST cost-effective solution?

A. Enable concurrency scaling in the workload management (WLM) queue.
B. Add more nodes using the AWS Management Console during peak hour

C. Set the distribution style to ALL.
D. Use elastic resize to quickly add nodes during peak time
E. Remove the nodes when they are not needed.
F. Use a snapshot, restore, and resize operatio
G. Switch to the new target cluster.

Answer: A

Explanation:
https://docs.aws.amazon.com/redshift/latest/dg/cm-c-implementing-workload-management.html

NEW QUESTION 91
A large company has a central data lake to run analytics across different departments. Each department uses a separate AWS account and stores its data in an Amazon S3 bucket in that account. Each AWS account uses the AWS Glue Data Catalog as its data catalog. There are different data lake access requirements based on roles. Associate analysts should only have read access to their departmental data. Senior data analysts can have access in multiple departments including theirs, but for a subset of columns only.
Which solution achieves these required access patterns to minimize costs and administrative tasks?

A. Consolidate all AWS accounts into one accoun
B. Create different S3 buckets for each department and move all the data from every account to the central data lake accoun
C. Migrate the individual data catalogs into a central data catalog and apply fine-grained permissions to give to each user the required access to tables and databases in AWS Glue and Amazon S3.
D. Keep the account structure and the individual AWS Glue catalogs on each accoun
E. Add a central data lake account and use AWS Glue to catalog data from various account
F. Configure cross-account access for AWS Glue crawlers to scan the data in each departmental S3 bucket to identify the schema and populate the catalo
G. Add the senior data analysts into the central account and apply highly detailed access controls in the Data Catalog and Amazon S3.
H. Set up an individual AWS account for the central data lak
I. Use AWS Lake Formation to catalog the cross- account location
J. On each individual S3 bucket, modify the bucket policy to grant S3 permissions to the Lake Formation service-linked rol
K. Use Lake Formation permissions to addfine-grained access controls to allow senior analysts to view specific tables and columns.
L. Set up an individual AWS account for the central data lake and configure a central S3 bucke
M. Use an AWS Lake Formation blueprint to move the data from the various buckets into the central S3 bucke
N. On each individual bucket, modify the bucket policy to grant S3 permissions to the Lake Formation service-linked rol
O. Use Lake Formation permissions to add fine-grained access controls for both associate and senior analysts to view specific tables and columns.

Answer: C

Explanation:
Lake Formation provides secure and granular access to data through a new grant/revoke permissions model that augments AWS Identity and Access Management (IAM) policies. Analysts and data scientists can use the full portfolio of AWS analytics and machine learning services, such as Amazon Athena, to access the data. The configured Lake Formation security policies help ensure that users can access only the data that they are authorized to access. Source : https://docs.aws.amazon.com/lake-formation/latest/dg/how-it-works.html

NEW QUESTION 94
A company uses an Amazon EMR cluster with 50 nodes to process operational data and make the data available for data analysts These jobs run nightly use Apache Hive with the Apache Jez framework as a processing model and write results to Hadoop Distributed File System (HDFS) In the last few weeks, jobs are failing and are producing the following error message
"File could only be replicated to 0 nodes instead of 1"
A data analytics specialist checks the DataNode logs the NameNode logs and network connectivity for potential issues that could have prevented HDFS from replicating data The data analytics specialist rules out these factors as causes for the issue
Which solution will prevent the jobs from failing'?

A. Monitor the HDFSUtilization metri
B. If the value crosses a user-defined threshold add task nodes to the EMR cluster
C. Monitor the HDFSUtilization metri.c If the value crosses a user-defined threshold add core nodes to the EMR cluster
D. Monitor the MemoryAllocatedMB metri
E. If the value crosses a user-defined threshold, add task nodes to the EMR cluster
F. Monitor the MemoryAllocatedMB metri
G. If the value crosses a user-defined threshold, add core nodes to the EMR cluster.

Answer: C

NEW QUESTION 95
A data engineering team within a shared workspace company wants to build a centralized logging system for all weblogs generated by the space reservation system. The company has a fleet of Amazon EC2 instances that process requests for shared space reservations on its website. The data engineering team wants to ingest all weblogs into a service that will provide a near-real-time search engine. The team does not want to manage the maintenance and operation of the logging system.
Which solution allows the data engineering team to efficiently set up the web logging system within AWS?

A. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis data stream to CloudWatc
B. Choose Amazon Elasticsearch Service as the end destination of the weblogs.
C. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis Data Firehose delivery stream to CloudWatc
D. Choose Amazon Elasticsearch Service as the end destination of the weblogs.
E. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis data stream to CloudWatc
F. Configure Splunk as the end destination of the weblogs.
G. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis Firehose delivery stream to CloudWatc
H. Configure Amazon DynamoDB as the end destination of the weblogs.

Answer: B

Explanation:
https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/CWL_ES_Stream.html

NEW QUESTION 99
A company has an application that uses the Amazon Kinesis Client Library (KCL) to read records from a Kinesis data stream.
After a successful marketing campaign, the application experienced a significant increase in usage. As a result, a data analyst had to split some shards in the data stream. When the shards were split, the application started
throwing an ExpiredIteratorExceptions error sporadically. What should the data analyst do to resolve this?

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

Answer: C

NEW QUESTION 100
A large company receives files from external parties in Amazon EC2 throughout the day. At the end of the day, the files are combined into a single file, compressed into a gzip file, and uploaded to Amazon S3. The total size of all the files is close to 100 GB daily. Once the files are uploaded to Amazon S3, an AWS Batch program executes a COPY command to load the files into an Amazon Redshift cluster.
Which program modification will accelerate the COPY process?

A. Upload the individual files to Amazon S3 and run the COPY command as soon as the files become available.
B. Split the number of files so they are equal to a multiple of the number of slices in the Amazon Redshift cluste
C. Gzip and upload the files to Amazon S3. Run the COPY command on the files.
D. Split the number of files so they are equal to a multiple of the number of compute nodes in the Amazon Redshift cluste
E. Gzip and upload the files to Amazon S3. Run the COPY command on the files.
F. Apply sharding by breaking up the files so the distkey columns with the same values go to the same file.Gzip and upload the sharded files to Amazon S3. Run the COPY command on the files.

Answer: B

NEW QUESTION 101
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 after data 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 106
A manufacturing company has many IoT devices in different facilities across the world The company is using Amazon Kinesis Data Streams to collect the data from the devices
The company's operations team has started to observe many WnteThroughputExceeded exceptions The operations team determines that the reason is the number of records that are being written to certain shards The data contains device ID capture date measurement type, measurement value and facility ID The facility ID is used as the partition key
Which action will resolve this issue?

A. Change the partition key from facility ID to a randomly generated key
B. Increase the number of shards
C. Archive the data on the producers' side
D. Change the partition key from facility ID to capture date

Answer: B

NEW QUESTION 107
A retail company stores order invoices in an Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster Indices on the cluster are created monthly Once a new month begins, no new writes are made to any of the indices from the previous months The company has been expanding the storage on the Amazon OpenSearch Service {Amazon Elasticsearch Service) cluster to avoid running out of space, but the company wants to reduce costs Most searches on the cluster are on the most recent 3 months of data while the audit team requires infrequent access to older data to generate periodic reports The most recent 3 months of data must be quickly available for queries, but the audit team can tolerate slower queries if the solution saves on cluster costs
Which of the following is the MOST operationally efficient solution to meet these requirements?

A. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to store the indices in Amazon S3 Glacier When the audit team requires the archived data restore thearchived indices back to the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster
B. Archive indices that are older than 3 months by taking manual snapshots and storing the snapshots in Amazon S3 When the audit team requires the archived data, restore the archived indices back to the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster
C. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to migrate the indices to Amazon OpenSearch Service (Amazon Elasticsearch Service) UltraWarm storage
D. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to migrate the indices to Amazon OpenSearch Service (Amazon Elasticsearch Service) UltraWarm storage When the audit team requires the older data: migrate the indices in UltraWarm storage back to hot storage

Answer: D

NEW QUESTION 111
A company uses Amazon Redshift as its data warehouse. A new table has columns that contain sensitive data. The data in the table will eventually be referenced by several existing queries that run many times a day.
A data analyst needs to load 100 billion rows of data into the new table. Before doing so, the data analyst must ensure that only members of the auditing group can read the columns containing sensitive data.
How can the data analyst meet these requirements with the lowest maintenance overhead?

A. Load all the data into the new table and grant the auditing group permission to read from the tabl
B. Load all the data except for the columns containing sensitive data into a second tabl
C. Grant the appropriate users read-only permissions to the second table.
D. Load all the data into the new table and grant the auditing group permission to read from the tabl
E. Use the GRANT SQL command to allow read-only access to a subset of columns to the appropriate users.
F. Load all the data into the new table and grant all users read-only permissions to non-sensitive columns.Attach an IAM policy to the auditing group with explicit ALLOW access to the sensitive data columns.
G. Load all the data into the new table and grant the auditing group permission to read from the table.Create a view of the new table that contains all the columns, except for those considered sensitive, and grant the appropriate users read-only permissions to the table.

Answer: B

Explanation:
https://aws.amazon.com/blogs/big-data/achieve-finer-grained-data-security-with-column-level-access-control-in

NEW QUESTION 115
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 119
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 124
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 126
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 131
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 132
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 134
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 136
A company operates toll services for highways across the country and collects data that is used to understand usage patterns. Analysts have requested the ability to run traffic reports in near-real time. The company is interested in building an ingestion pipeline that loads all the data into an Amazon Redshift cluster and alerts operations personnel when toll traffic for a particular toll station does not meet a specified threshold. Station data and the corresponding threshold values are stored in Amazon S3.
Which approach is the MOST efficient way to meet these requirements?

A. Use Amazon Kinesis Data Firehose to collect data and deliver it to Amazon Redshift and Amazon Kinesis Data Analytics simultaneousl
B. Create a reference data source in Kinesis Data Analytics to temporarily store the threshold values from Amazon S3 and compare the count of vehicles for a particular toll station against its corresponding threshold valu
C. Use AWS Lambda to publish an Amazon Simple Notification Service (Amazon SNS) notification if the threshold is not met.
D. Use Amazon Kinesis Data Streams to collect all the data from toll station
E. Create a stream in Kinesis Data Streams to temporarily store the threshold values from Amazon S3. Send both streams to Amazon Kinesis Data Analytics to compare the count of vehicles for a particular toll station against its corresponding threshold valu
F. Use AWS Lambda to publish an Amazon Simple Notification Service (Amazon SNS) notification if the threshold is not me
G. Connect Amazon Kinesis Data Firehose to Kinesis Data Streams to deliver the data to Amazon Redshift.
H. Use Amazon Kinesis Data Firehose to collect data and deliver it to Amazon Redshif
I. Then, automatically trigger an AWS Lambda function that queries the data in Amazon Redshift, compares the count of vehicles for a particular toll station against its corresponding threshold values read from Amazon S3, and publishes an Amazon Simple Notification Service (Amazon SNS) notification if the threshold is not met.
J. Use Amazon Kinesis Data Firehose to collect data and deliver it to Amazon Redshift and Amazon Kinesis Data Analytics simultaneousl
K. Use Kinesis Data Analytics to compare the count of vehicles against the threshold value for the station stored in a table as an in-application stream based on information stored in Amazon S3. Configure an AWS Lambda function as an output for the application that will publish an Amazon Simple Queue Service (Amazon SQS) notification to alert operations personnel if the threshold is not met.

Answer: D

NEW QUESTION 140
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 in DynamoD
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 141
A banking company wants to collect large volumes of transactional data using Amazon Kinesis Data Streams for real-time analytics. The company uses PutRecord to send data to Amazon Kinesis, and has observed network outages during certain times of the day. The company wants to obtain exactly once semantics for the entire processing pipeline.
What should the company do to obtain these characteristics?

A. Design the application so it can remove duplicates during processing be embedding a unique ID in each record.
B. Rely on the processing semantics of Amazon Kinesis Data Analytics to avoid duplicate processing of events.
C. Design the data producer so events are not ingested into Kinesis Data Streams multiple times.
D. Rely on the exactly one processing semantics of Apache Flink and Apache Spark Streaming included in Amazon EMR.

Answer: A

NEW QUESTION 144
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 145
A large retailer has successfully migrated to an Amazon S3 data lake architecture. The company's marketing team is using Amazon Redshift and Amazon QuickSight to analyze data, and derive and visualize insights. To ensure the marketing team has the most up-to-date actionable information, a data analyst implements nightly refreshes of Amazon Redshift using terabytes of updates from the previous day.
After the first nightly refresh, users report that half of the most popular dashboards that had been running correctly before the refresh are now running much slower. Amazon CloudWatch does not show any alerts.
What is the MOST likely cause for the performance degradation?

A. The dashboards are suffering from inefficient SQL queries.
B. The cluster is undersized for the queries being run by the dashboards.
C. The nightly data refreshes are causing a lingering transaction that cannot be automatically closed by Amazon Redshift due to ongoing user workloads.
D. The nightly data refreshes left the dashboard tables in need of a vacuum operation that could not be automatically performed by Amazon Redshift due to ongoing user workloads.

Answer: D

Explanation:
https://github.com/awsdocs/amazon-redshift-developer-guide/issues/21

NEW QUESTION 148
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.Configure the EMR cluster with multiple master node
B. Schedule automated snapshots using Amazon EventBridge.
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 153
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 156
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 158
A company is planning to create a data lake in Amazon S3. The company wants to create tiered storage based on access patterns and cost objectives. The solution must include support for JDBC connections from legacy clients, metadata management that allows federation for access control, and batch-based ETL using PySpark and Scala. Operational management should be limited.
Which combination of components can meet these requirements? (Choose three.)

A. AWS Glue Data Catalog for metadata management
B. Amazon EMR with Apache Spark for ETL
C. AWS Glue for Scala-based ETL
D. Amazon EMR with Apache Hive for JDBC clients
E. Amazon Athena for querying data in Amazon S3 using JDBC drivers
F. Amazon EMR with Apache Hive, using an Amazon RDS with MySQL-compatible backed metastore

Answer: BEF

NEW QUESTION 159
An analytics software as a service (SaaS) provider wants to offer its customers business intelligence <BI) reporting capabilities that are self-service The provider is using Amazon QuickSight to build these reports The data for the reports resides in a multi-tenant database, but each customer should only be able to access their own data
The provider wants to give customers two user role options
• Read-only users for individuals who only need to view dashboards
• Power users for individuals who are allowed to create and share new dashboards with other users Which QuickSight feature allows the provider to meet these requirements'?

A. Embedded dashboards
B. Table calculations
C. Isolated namespaces
D. SPICE

Answer: A

NEW QUESTION 162
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 167
A large university has adopted a strategic goal of increasing diversity among enrolled students. The data analytics team is creating a dashboard with data visualizations to enable stakeholders to view historical trends. All access must be authenticated using Microsoft Active Directory. All data in transit and at rest must be encrypted.
Which solution meets these requirements?

A. Amazon QuickSight Standard edition configured to perform identity federation using SAML 2.0. and the default encryption settings.
B. Amazon QuickSight Enterprise edition configured to perform identity federation using SAML 2.0 and the default encryption settings.
C. Amazon QuckSight Standard edition using AD Connector to authenticate using Active Directory.Configure Amazon QuickSight to use customer-provided keys imported into AWS KMS.
D. Amazon QuickSight Enterprise edition using AD Connector to authenticate using Active Directory.Configure Amazon QuickSight to use customer-provided keys imported into AWS KMS.

Answer: D

NEW QUESTION 172
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 177
An online retail company uses Amazon Redshift to store historical sales transactions. The company is required to encrypt data at rest in the clusters to comply with the Payment Card Industry Data Security Standard (PCI DSS). A corporate governance policy mandates management of encryption keys using an on-premises hardware security module (HSM).
Which solution meets these requirements?

A. Create and manage encryption keys using AWS CloudHSM Classi
B. Launch an Amazon Redshift cluster in a VPC with the option to use CloudHSM Classic for key management.
C. Create a VPC and establish a VPN connection between the VPC and the on-premises networ
D. Create an HSM connection and client certificate for the on-premises HS
E. Launch a cluster in the VPC with theoption to use the on-premises HSM to store keys.
F. Create an HSM connection and client certificate for the on-premises HS
G. Enable HSM encryption on the existing unencrypted cluster by modifying the cluste
H. Connect to the VPC where the Amazon Redshift cluster resides from the on-premises network using a VPN.
I. Create a replica of the on-premises HSM in AWS CloudHS
J. Launch a cluster in a VPC with the option to use CloudHSM to store keys.

Answer: B

NEW QUESTION 181
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 185
A transport company wants to track vehicular movements by capturing geolocation records. The records are 10 B in size and up to 10,000 records are captured each second. Data transmission delays of a few minutes are acceptable, considering unreliable network conditions. The transport company decided to use Amazon Kinesis Data Streams to ingest the data. The company is looking for a reliable mechanism to send data to Kinesis Data Streams while maximizing the throughput efficiency of the Kinesis shards.
Which solution will meet the company's requirements?

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

Answer: B

NEW QUESTION 189
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 192
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 the metadata 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 193
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 196
A company needs to store objects containing log data in JSON format. The objects are generated by eight applications running in AWS. Six of the applications generate a total of 500 KiB of data per second, and two of the applications can generate up to 2 MiB of data per second.
A data engineer wants to implement a scalable solution to capture and store usage data in an Amazon S3
bucket. The usage data objects need to be reformatted, converted to .csv format, and then compressed before they are stored in Amazon S3. The company requires the solution to include the least custom code possible and has authorized the data engineer to request a service quota increase if needed.
Which solution meets these requirements?

A. Configure an Amazon Kinesis Data Firehose delivery stream for each applicatio
B. Write AWS Lambda functions to read log data objects from the stream for each applicatio
C. Have the function perform reformatting and .csv conversio
D. Enable compression on all the delivery streams.
E. Configure an Amazon Kinesis data stream with one shard per applicatio
F. Write an AWS Lambda function to read usage data objects from the shard
G. Have the function perform .csv conversion, reformatting, and compression of the dat
H. Have the function store the output in Amazon S3.
I. Configure an Amazon Kinesis data stream for each applicatio
J. Write an AWS Lambda function to read usage data objects from the stream for each applicatio
K. Have the function perform .csv conversion, reformatting, and compression of the dat
L. Have the function store the output in Amazon S3.
M. Store usage data objects in an Amazon DynamoDB tabl
N. Configure a DynamoDB stream to copy the objects to an S3 bucke
O. Configure an AWS Lambda function to be triggered when objects are written to the S3 bucke
P. Have the function convert the objects into .csv format.

Answer: A

NEW QUESTION 200
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 205
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 207
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 212
A retail company leverages Amazon Athena for ad-hoc queries against an AWS Glue Data Catalog. The data analytics team manages the data catalog and data access for the company. The data analytics team wants to separate queries and manage the cost of running those queries by different workloads and teams. Ideally, the data analysts want to group the queries run by different users within a team, store the query results in individual Amazon S3 buckets specific to each team, and enforce cost constraints on the queries run against the Data Catalog.
Which solution meets these requirements?

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

Answer: C

Explanation:
https://aws.amazon.com/about-aws/whats-new/2019/02/athena_workgroups/

NEW QUESTION 216

.....

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questons and Answers in PDF Format

DAS-C01 Practice Exam Features:

* DAS-C01 Questions and Answers Updated Frequently

* DAS-C01 Practice Questions Verified by Expert Senior Certified Staff

* DAS-C01 Most Realistic Questions that Guarantee you a Pass on Your FirstTry

* DAS-C01 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year