DAS-C01 Dumps

AWS Certified Data Analytics - Specialty

https://www.certleader.com/DAS-C01-dumps.html

NEW QUESTION 1

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 2

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 3

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 4

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

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

Answer: B

NEW QUESTION 5

An ecommerce company is migrating its business intelligence environment from on premises to the AWS Cloud. The company will use Amazon Redshift in a public subnet and Amazon QuickSight. The tables already are loaded into Amazon Redshift and can be accessed by a SQL tool.
The company starts QuickSight for the first time. During the creation of the data source, a data analytics specialist enters all the information and tries to validate the connection. An error with the following message occurs: "Creating a connection to your data source timed out."

How should the data analytics specialist resolve this error?

A. Grant the SELECT permission on Amazon Redshift tables.
B. Add the QuickSight IP address range into the Amazon Redshift security group.
C. Create an IAM role for QuickSight to access Amazon Redshift.
D. Use a QuickSight admin user for creating the dataset.

Answer: A

Explanation:
Connection to the database times out
Your client connection to the database appears to hang or time out when running long queries, such as a COPY command. In this case, you might observe that the Amazon Redshift console displays that the query has completed, but the client tool itself still appears to be running the query. The results of the query might be missing or incomplete depending on when the connection stopped.

NEW QUESTION 6
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 7
A company owns facilities with IoT devices installed across the world. The company is using Amazon Kinesis Data Streams to stream data from the devices to Amazon S3. The company's operations team wants to get insights from the IoT data to monitor data quality at ingestion. The insights need to be derived in near-real time, and the output must be logged to Amazon DynamoDB for further analysis.
Which solution meets these requirements?

A. Connect Amazon Kinesis Data Analytics to analyze the stream dat
B. Save the output to DynamoDB by using the default output from Kinesis Data Analytics.
C. Connect Amazon Kinesis Data Analytics to analyze the stream dat
D. Save the output to DynamoDB by using an AWS Lambda function.
E. Connect Amazon Kinesis Data Firehose to analyze the stream data by using an AWS Lambda function.Save the output to DynamoDB by using the default output from Kinesis Data Firehose.
F. Connect Amazon Kinesis Data Firehose to analyze the stream data by using an AWS Lambda function.Save the data to Amazon S3. Then run an AWS Glue job on schedule to ingest the data into DynamoDB.

Answer: C

NEW QUESTION 8
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 9
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 10
A company has a marketing department and a finance department. The departments are storing data in Amazon S3 in their own AWS accounts in AWS Organizations. Both departments use AWS Lake Formation to catalog and secure their data. The departments have some databases and tables that share common names.
The marketing department needs to securely access some tables from the finance department. Which two steps are required for this process? (Choose two.)

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

Answer: AB

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

NEW QUESTION 10
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 15
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 20
A hospital is building a research data lake to ingest data from electronic health records (EHR) systems from multiple hospitals and clinics. The EHR systems are independent of each other and do not have a common patient identifier. The data engineering team is not experienced in machine learning (ML) and has been asked to generate a unique patient identifier for the ingested records.
Which solution will accomplish this task?

A. An AWS Glue ETL job with the FindMatches transform
B. Amazon Kendra
C. Amazon SageMaker Ground Truth
D. An AWS Glue ETL job with the ResolveChoice transform

Answer: A

Explanation:
Matching Records with AWS Lake Formation FindMatches

NEW QUESTION 25
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 28
A media analytics company consumes a stream of social media posts. The posts are sent to an Amazon Kinesis data stream partitioned on user_id. An AWS Lambda function retrieves the records and validates the content before loading the posts into an Amazon Elasticsearch cluster. The validation process needs to receive the posts for a given user in the order they were received. A data analyst has noticed that, during peak hours, the social media platform posts take more than an hour to appear in the Elasticsearch cluster.
What should the data analyst do reduce this latency?

A. Migrate the validation process to Amazon Kinesis Data Firehose.
B. Migrate the Lambda consumers from standard data stream iterators to an HTTP/2 stream consumer.
C. Increase the number of shards in the stream.
D. Configure multiple Lambda functions to process the stream.

Answer: D

NEW QUESTION 33
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 34
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 35
A company receives data from its vendor in JSON format with a timestamp in the file name. The vendor uploads the data to an Amazon S3 bucket, and the data is registered into the company's data lake for analysis and reporting. The company has configured an S3 Lifecycle policy to archive all files to S3 Glacier after 5 days.
The company wants to ensure that its AWS Glue crawler catalogs data only from S3 Standard storage and ignores the archived files. A data analytics specialist must implement a solution to achieve this goal without changing the current S3 bucket configuration.
Which solution meets these requirements?

A. Use the exclude patterns feature of AWS Glue to identify the S3 Glacier files for the crawler to exclude.
B. Schedule an automation job that uses AWS Lambda to move files from the original S3 bucket to a new S3 bucket for S3 Glacier storage.
C. Use the excludeStorageClasses property in the AWS Glue Data Catalog table to exclude files on S3 Glacier storage
D. Use the include patterns feature of AWS Glue to identify the S3 Standard files for the crawler to include.

Answer: A

NEW QUESTION 39
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 42
A company wants to improve user satisfaction for its smart home system by adding more features to its recommendation engine. Each sensor asynchronously pushes its nested JSON data into Amazon Kinesis Data Streams using the Kinesis Producer Library (KPL) in Java. Statistics from a set of failed sensors showed that, when a sensor is malfunctioning, its recorded data is not always sent to the cloud.
The company needs a solution that offers near-real-time analytics on the data from the most updated sensors. Which solution enables the company to meet these requirements?

A. Set the RecordMaxBufferedTime property of the KPL to "1" to disable the buffering on the sensor side.Use Kinesis Data Analytics to enrich the data based on a company-developed anomaly detection SQL scrip
B. Push the enriched data to a fleet of Kinesis data streams and enable the data transformation feature to flatten the JSON fil
C. Instantiate a dense storage Amazon Redshift cluster and use it as the destination for the Kinesis Data Firehose delivery stream.
D. Update the sensors code to use the PutRecord/PutRecords call from the Kinesis Data Streams API with the AWS SDK for Jav
E. Use Kinesis Data Analytics to enrich the data based on a company-developed anomaly detection SQL scrip
F. Direct the output of KDA application to a Kinesis Data Firehose delivery stream, enable the data transformation feature to flatten the JSON file, and set the Kinesis Data Firehose destination to an Amazon Elasticsearch Service cluster.
G. Set the RecordMaxBufferedTime property of the KPL to "0" to disable the buffering on the sensor side.Connect for each stream a dedicated Kinesis Data Firehose delivery stream and enable the data transformation feature to flatten the JSON file before sending it to an Amazon S3 bucke
H. Load the S3 data into an Amazon Redshift cluster.
I. Update the sensors code to use the PutRecord/PutRecords call from the Kinesis Data Streams API with the AWS SDK for Jav
J. Use AWS Glue to fetch and process data from the stream using the Kinesis Client Library (KCL). Instantiate an Amazon Elasticsearch Service cluster and use AWS Lambda to directly push data into it.

Answer: B

Explanation:
https://docs.aws.amazon.com/streams/latest/dev/developing-producers-with-kpl.html
The KPL can incur an additional processing delay of up to RecordMaxBufferedTime within the library (user-configurable). Larger values of RecordMaxBufferedTime results in higher packing efficiencies and better performance. Applications that cannot tolerate this additional delay may need to use the AWS SDK directly.

NEW QUESTION 46
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 51
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 52
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 57
A hospital uses wearable medical sensor devices to collect data from patients. The hospital is architecting a near-real-time solution that can ingest the data securely at scale. The solution should also be able to remove the patient's protected health information (PHI) from the streaming data and store the data in durable storage.
Which solution meets these requirements with the least operational overhead?

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

Answer: D

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

NEW QUESTION 59
A company is migrating its existing on-premises ETL jobs to Amazon EMR. The code consists of a series of jobs written in Java. The company needs to reduce overhead for the system administrators without changing the underlying code. Due to the sensitivity of the data, compliance requires that the company use root device volume encryption on all nodes in the cluster. Corporate standards require that environments be provisioned though AWS CloudFormation when possible.
Which solution satisfies these requirements?

A. Install open-source Hadoop on Amazon EC2 instances with encrypted root device volume
B. Configure the cluster in the CloudFormation template.
C. Use a CloudFormation template to launch an EMR cluste
D. In the configuration section of the cluster, define a bootstrap action to enable TLS.
E. Create a custom AMI with encrypted root device volume
F. Configure Amazon EMR to use the custom AMI using the CustomAmild property in the CloudFormation template.
G. Use a CloudFormation template to launch an EMR cluste
H. In the configuration section of the cluster, define a bootstrap action to encrypt the root device volume of every node.

Answer: C

NEW QUESTION 62

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

Which factors could be causing the duplicated data? (Choose two.)

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

Answer: BD

NEW QUESTION 66

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 70

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

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

Answer: A

NEW QUESTION 73

A data analyst runs a large number of data manipulation language (DML) queries by using Amazon Athena with the JDBC driver. Recently, a query failed after It ran for 30 minutes. The query returned the following message Java.sql.SGLException: Query timeout
The data analyst does not immediately need the query results However, the data analyst needs a long-term solution for this problem
Which solution will meet these requirements?

A. Split the query into smaller queries to search smaller subsets of data.
B. In the settings for Athena, adjust the DML query timeout limit
C. In the Service Quotas console, request an increase for the DML query timeout
D. Save the tables as compressed .csv files

Answer: A

NEW QUESTION 78

A data engineer is using AWS Glue ETL jobs to process data at frequent intervals The processed data is then copied into Amazon S3 The ETL jobs run every 15 minutes. The AWS Glue Data Catalog partitions need to be updated automatically after the completion of each job
Which solution will meet these requirements MOST cost-effectively?

A. Use the AWS Glue Data Catalog to manage the data catalog Define an AWS Glue workflow for the ETL process Define a trigger within the workflow that can start the crawler when an ETL job run is complete
B. Use the AWS Glue Data Catalog to manage the data catalog Use AWS Glue Studio to manage ETL job
C. Use the AWS Glue Studio feature that supports updates to the AWS Glue Data Catalog during job runs.
D. Use an Apache Hive metastore to manage the data catalog Update the AWS Glue ETL code to include the enableUpdateCatalog and partitionKeys arguments.
E. Use the AWS Glue Data Catalog to manage the data catalog Update the AWS Glue ETL code to include the enableUpdateCatalog and partitionKeys

arguments.

Answer: A

NEW QUESTION 82
A company leverages Amazon Athena for ad-hoc queries against data stored in Amazon S3. The company wants to implement additional controls to separate query execution and query history among users, teams, or applications running in the same AWS account to comply with internal security policies.
Which solution meets these requirements?

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

Answer: B

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

NEW QUESTION 87
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 90
An IoT company wants to release a new device that will collect data to track sleep overnight on an intelligent mattress. Sensors will send data that will be uploaded to an Amazon S3 bucket. About 2 MB of data is generated each night for each bed. Data must be processed and summarized for each user, and the results need to be available as soon as possible. Part of the process consists of time windowing and other functions. Based on tests with a Python script, every run will require about 1 GB of memory and will complete within a couple of minutes.
Which solution will run the script in the MOST cost-effective way?

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

Answer: A

NEW QUESTION 95
A US-based sneaker retail company launched its global website. All the transaction data is stored in Amazon RDS and curated historic transaction data is stored in Amazon Redshift in the us-east-1 Region. The business intelligence (BI) team wants to enhance the user experience by providing a dashboard for sneaker trends. The BI team decides to use Amazon QuickSight to render the website dashboards. During development, a team in Japan provisioned Amazon QuickSight in ap-northeast-1. The team is having difficulty connecting Amazon QuickSight from ap-northeast-1 to Amazon Redshift in us-east-1.
Which solution will solve this issue and meet the requirements?

A. In the Amazon Redshift console, choose to configure cross-Region snapshots and set the destination Region as ap-northeast-1. Restore the Amazon Redshift Cluster from the snapshot and connect to Amazon QuickSight launched in ap-northeast-1.
B. Create a VPC endpoint from the Amazon QuickSight VPC to the Amazon Redshift VPC so Amazon QuickSight can access data from Amazon Redshift.
C. Create an Amazon Redshift endpoint connection string with Region information in the string and use this connection string in Amazon QuickSight to connect to Amazon Redshift.
D. Create a new security group for Amazon Redshift in us-east-1 with an inbound rule authorizing access from the appropriate IP address range for the Amazon QuickSight servers in ap-northeast-1.

Answer: B

NEW QUESTION 98
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 is building a data lake and needs to ingest data from a relational database that has time-series data. The company wants to use managed services to accomplish this. The process needs to be scheduled daily and bring incremental data only from the source into Amazon S3.
What is the MOST cost-effective approach to meet these requirements?

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

Answer: A

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

NEW QUESTION 101
A company hosts an on-premises PostgreSQL database that contains historical data. An internal legacy application uses the database for read-only activities. The company's business team wants to move the data to a data lake in Amazon S3 as soon as possible and enrich the data for analytics.
The company has set up an AWS Direct Connect connection between its VPC and its on-premises network. A data analytics specialist must design a solution that achieves the business team's goals with the least operational overhead.
Which solution meets these requirements?

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

Answer: B

NEW QUESTION 105
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 109
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 112
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 115
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 118
A company has collected more than 100 TB of log files in the last 24 months. The files are stored as raw text in a dedicated Amazon S3 bucket. Each object has a key of the form year-month-day_log_HHmmss.txt where HHmmss represents the time the log file was initially created. A table was created in Amazon Athena that points to the S3 bucket. One-time queries are run against a subset of columns in the table several times an hour.
A data analyst must make changes to reduce the cost of running these queries. Management wants a solution with minimal maintenance overhead.
Which combination of steps should the data analyst take to meet these requirements? (Choose three.)

A. Convert the log files to Apace Avro format.
B. Add a key prefix of the form date=year-month-day/ to the S3 objects to partition the data.
C. Convert the log files to Apache Parquet format.
D. Add a key prefix of the form year-month-day/ to the S3 objects to partition the data.
E. Drop and recreate the table with the PARTITIONED BY claus
F. Run the ALTER TABLE ADD PARTITION statement.
G. Drop and recreate the table with the PARTITIONED BY claus
H. Run the MSCK REPAIR TABLE statement.

Answer: BCF

NEW QUESTION 120

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 123
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 128
A company analyzes historical data and needs to query data that is stored in Amazon S3. New data is generated daily as .csv files that are stored in Amazon S3. The company's analysts are using Amazon Athena to perform SQL queries against a recent subset of the overall data. The amount of data that is ingested into Amazon S3 has increased substantially over time, and the query latency also has increased.
Which solutions could the company implement to improve query performance? (Choose two.)

A. Use MySQL Workbench on an Amazon EC2 instance, and connect to Athena by using a JDBC or ODBC connecto
B. Run the query from MySQL Workbench instead of Athena directly.
C. Use Athena to extract the data and store it in Apache Parquet format on a daily basi
D. Query the extracted data.
E. Run a daily AWS Glue ETL job to convert the data files to Apache Parquet and to partition the converted file
F. Create a periodic AWS Glue crawler to automatically crawl the partitioned data on a daily basis.
G. Run a daily AWS Glue ETL job to compress the data files by using the .gzip forma
H. Query the compressed data.
I. Run a daily AWS Glue ETL job to compress the data files by using the .lzo forma
J. Query the compressed data.

Answer: BC

NEW QUESTION 131
A company uses Amazon Redshift for its data warehousing needs. ETL jobs run every night to load data, apply business rules, and create aggregate tables for reporting. The company's data analysis, data science, and business intelligence teams use the data warehouse during regular business hours. The workload management is set to auto, and separate queues exist for each team with the priority set to NORMAL.
Recently, a sudden spike of read queries from the data analysis team has occurred at least twice daily, and queries wait in line for cluster resources. The company needs a solution that enables the data analysis team to avoid query queuing without impacting latency and the query times of other teams.
Which solution meets these requirements?

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

Answer: D

NEW QUESTION 135
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 138
A manufacturing company uses Amazon Connect to manage its contact center and Salesforce to manage its customer relationship management (CRM) data. The data engineering team must build a pipeline to ingest data from the contact center and CRM system into a data lake that is built on Amazon S3.
What is the MOST efficient way to collect data in the data lake with the LEAST operational overhead?

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

Answer: B

NEW QUESTION 141
A company uses Amazon Redshift as its data warehouse A new table includes some columns that contain sensitive data and some columns that contain non-sensitive data The data in the table eventually will be referenced by several existing queries that run many times each day
A data analytics specialist must ensure that only members of the company's auditing team can read the columns that contain sensitive data All other users must have read-only access to the columns that contain non-sensitive data
Which solution will meet these requirements with the LEAST operational overhead?

A. Grant the auditing team permission to read from the tabl
B. Load the columns that contain non-sensitive data into a second tabl
C. Grant the appropriate users read-only permissions to the second table.
D. Grant all users read-only permissions to the columns that contain non-sensitive data Use the GRANT SELECT command to allow the auditing team to access the columns that contain sensitive data
E. Grant all users read-only permissions to the columns that contain non-sensitive data Attach an 1AM policy to the auditing team with an explicit Allow action that grants access to the columns that contain sensitive data
F. Grant the auditing team permission to read from the table Create a view of the table that includes the columns that contain non-sensitive data Grant the appropriate users read-only permissions to that view

Answer: B

Explanation:
https://aws.amazon.com/jp/about-aws/whats-new/2020/03/announcing-column-level-access-control-for-amazon

NEW QUESTION 146
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 150
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 152
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 153
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 157
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 161
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 164

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 165

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 167

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 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 168

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

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

Answer: CDF

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

NEW QUESTION 173

A large energy company is using Amazon QuickSight to build dashboards and report the historical usage data of its customers This data is hosted in Amazon Redshift The reports need access to all the fact tables' billions ot records to create aggregation in real time grouping by multiple dimensions
A data analyst created the dataset in QuickSight by using a SQL query and not SPICE Business users have noted that the response time is not fast enough to meet their needs
Which action would speed up the response time for the reports with the LEAST implementation effort?

A. Use QuickSight to modify the current dataset to use SPICE
B. Use AWS Glue to create an Apache Spark job that joins the fact table with the dimension
C. Load the data into a new table

D. Use Amazon Redshift to create a materialized view that joins the fact table with the dimensions
E. Use Amazon Redshift to create a stored procedure that joins the fact table with the dimensions Load the data into a new table

Answer: A

NEW QUESTION 176
Once a month, a company receives a 100 MB .csv file compressed with gzip. The file contains 50,000 property listing records and is stored in Amazon S3 Glacier. The company needs its data analyst to query a subset of the data for a specific vendor.
What is the most cost-effective solution?

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

Answer: A

NEW QUESTION 181
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 182
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 187
A retail company's data analytics team recently created multiple product sales analysis dashboards for the average selling price per product using Amazon QuickSight. The dashboards were created from .csv files uploaded to Amazon S3. The team is now planning to share the dashboards with the respective external product owners by creating individual users in Amazon QuickSight. For compliance and governance reasons, restricting access is a key requirement. The product owners should view only their respective product analysis in the dashboard reports.
Which approach should the data analytics team take to allow product owners to view only their products in the dashboard?

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

Answer: D

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

---

NEW QUESTION 192
......

Thank You for Trying Our Product

* 100% Pass or Money Back

    All our products come with a 90-day Money Back Guarantee.

* One year free update

    You can enjoy free update one year. 24x7 online support.

* Trusted by Millions

    We currently serve more than 30,000,000 customers.

* Shop Securely

    All transactions are protected by VeriSign!

100% Pass Your DAS-C01 Exam with Our Prep Materials Via below:

https://www.certleader.com/DAS-C01-dumps.html