Amazon-Web-Services

Exam Questions DAS-C01

AWS Certified Data Analytics - Specialty

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

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

Answer: B

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

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

Answer: AD

NEW QUESTION 3
A 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 calculatethe average per secon
E. Send the results to an AWS Lambda function that sends the alarm to Amazon SNS.

Answer: D

NEW QUESTION 4
An airline has been collecting metrics on flight activities for analytics. A recently completed proof of concept demonstrates how the company provides insights to data analysts to improve on-time departures. The proof of concept used objects in Amazon S3, which contained the metrics in .csv format, and used Amazon Athena for querying the data. As the amount of data increases, the data analyst wants to optimize the storage solution to improve query performance.
Which options should the data analyst use to improve performance as the data lake grows? (Choose three.)

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

Answer: CDF

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

NEW QUESTION 5
A global company has different sub-organizations, and each sub-organization sells its products and services in various countries. The company's senior leadership

wants to quickly identify which sub-organization is the strongest performer in each country. All sales data is stored in Amazon S3 in Parquet format.
Which approach can provide the visuals that senior leadership requested with the least amount of effort?

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

Answer: A

NEW QUESTION 6
An insurance company has raw data in JSON format that is sent without a predefined schedule through an Amazon Kinesis Data Firehose delivery stream to an Amazon S3 bucket. An AWS Glue crawler is scheduled to run every 8 hours to update the schema in the data catalog of the tables stored in the S3 bucket. Data analysts analyze the data using Apache Spark SQL on Amazon EMR set up with AWS Glue Data Catalog as the metastore. Data analysts say that, occasionally, the data they receive is stale. A data engineer needs to provide access to the most up-to-date data.
Which solution meets these requirements?

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

Answer: D

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

NEW QUESTION 7
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 8
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 9
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 destinationof the weblog

Answer: B

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

NEW QUESTION 10
A company's data analyst needs to ensure that queries executed in Amazon Athena cannot scan more than a prescribed amount of data for cost control purposes. Queries that exceed the prescribed threshold must be canceled immediately.
What should the data analyst do to achieve this?

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

Answer: B

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

NEW QUESTION 10
A central government organization is collecting events from various internal applications using Amazon Managed Streaming for Apache Kafka (Amazon MSK). The organization has configured a separate Kafka topic for each application to separate the data. For security reasons, the Kafka cluster has been configured to only allow TLS encrypted data and it encrypts the data at rest.
A recent application update showed that one of the applications was configured incorrectly, resulting in writing data to a Kafka topic that belongs to another application. This resulted in multiple errors in the analytics pipeline as data from different applications appeared on the same topic. After this incident, the organization wants to prevent applications from writing to a topic different than the one they should write to.
Which solution meets these requirements with the least amount of effort?

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

Answer: B

NEW QUESTION 14
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 17
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 21
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 22
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 27
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 32
A company has developed an Apache Hive script to batch process data stared in Amazon S3. The script needs to run once every day and store the output in Amazon S3. The company tested the script, and it completes within 30 minutes on a small local three-node cluster.
Which solution is the MOST cost-effective for scheduling and executing the script?

A. Create an AWS Lambda function to spin up an Amazon EMR cluster with a Hive execution ste
B. Set KeepJobFlowAliveWhenNoSteps to false and disable the termination protection fla
C. Use Amazon CloudWatch Events to schedule the Lambda function to run daily.
D. Use the AWS Management Console to spin up an Amazon EMR cluster with Python Hu
E. Hive, and Apache Oozi

F. Set the termination protection flag to true and use Spot Instances for the core nodes of the cluste
G. Configure an Oozie workflow in the cluster to invoke the Hive script daily.
H. Create an AWS Glue job with the Hive script to perform the batch operatio
I. Configure the job to run once a day using a time-based schedule.
J. Use AWS Lambda layers and load the Hive runtime to AWS Lambda and copy the Hive script.Schedule the Lambda function to run daily by creating a workflow using AWS Step Functions.

Answer: C

NEW QUESTION 34
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 36
An online gaming company is using an Amazon Kinesis Data Analytics SQL application with a Kinesis data stream as its source. The source sends three non-null fields to the application: player_id, score, and us_5_digit_zip_code.
A data analyst has a .csv mapping file that maps a small number of us_5_digit_zip_code values to a territory code. The data analyst needs to include the territory code, if one exists, as an additional output of the Kinesis Data Analytics application.
How should the data analyst meet this requirement while minimizing costs?

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

Answer: C

NEW QUESTION 41
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 46
A company has an application that ingests streaming data. The company needs to analyze this stream over a 5-minute timeframe to evaluate the stream for anomalies with Random Cut Forest (RCF) and summarize the current count of status codes. The source and summarized data should be persisted for future use.
Which approach would enable the desired outcome while keeping data persistence costs low?

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

application evaluate the stream over a 1-minute window using the RCF function and summarize the count of status code
J. Persist the results to Amazon S3 through a Kinesis Data Analytics output to an AWS Lambda integration.

Answer: B

NEW QUESTION 49
A smart home automation company must efficiently ingest and process messages from various connected devices and sensors. The majority of these messages are comprised of a large number of small files. These messages are ingested using Amazon Kinesis Data Streams and sent to Amazon S3 using a Kinesis data stream consumer application. The Amazon S3 message data is then passed through a processing pipeline built on Amazon EMR running scheduled PySpark jobs. The data platform team manages data processing and is concerned about the efficiency and cost of downstream data processing. They want to continue to use PySpark.
Which solution improves the efficiency of the data processing jobs and is well architected?

A. Send the sensor and devices data directly to a Kinesis Data Firehose delivery stream to send the data to Amazon S3 with Apache Parquet record format conversion enable
B. Use Amazon EMR running PySpark to process the data in Amazon S3.
C. Set up an AWS Lambda function with a Python runtime environmen
D. Process individual Kinesis data stream messages from the connected devices and sensors using Lambda.
E. Launch an Amazon Redshift cluste
F. Copy the collected data from Amazon S3 to Amazon Redshift and move the data processing jobs from Amazon EMR to Amazon Redshift.
G. Set up AWS Glue Python jobs to merge the small data files in Amazon S3 into larger files and transform them to Apache Parquet forma
H. Migrate the downstream PySpark jobs from Amazon EMR to AWS Glue.

Answer: D

Explanation:
https://aws.amazon.com/it/about-aws/whats-new/2020/04/aws-glue-now-supports-serverless-streaming-etl/

NEW QUESTION 50
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 55
A company wants to use an automatic machine learning (ML) Random Cut Forest (RCF) algorithm to visualize complex real-world scenarios, such as detecting seasonality and trends, excluding outers, and imputing missing values.
The team working on this project is non-technical and is looking for an out-of-the-box solution that will require the LEAST amount of management overhead.
Which solution will meet these requirements?

A. Use an AWS Glue ML transform to create a forecast and then use Amazon QuickSight to visualize the data.
B. Use Amazon QuickSight to visualize the data and then use ML-powered forecasting to forecast the key business metrics.
C. Use a pre-build ML AMI from the AWS Marketplace to create forecasts and then use Amazon QuickSight to visualize the data.
D. Use calculated fields to create a new forecast and then use Amazon QuickSight to visualize the data.

Answer: A

NEW QUESTION 57
An airline has .csv-formatted data stored in Amazon S3 with an AWS Glue Data Catalog. Data analysts want to join this data with call center data stored in Amazon Redshift as part of a dally batch process. The Amazon Redshift cluster is already under a heavy load. The solution must be managed, serverless, well-functioning, and minimize the load on the existing Amazon Redshift cluster. The solution should also require minimal effort and development activity.
Which solution meets these requirements?

A. Unload the call center data from Amazon Redshift to Amazon S3 using an AWS Lambda function.Perform the join with AWS Glue ETL scripts.
B. Export the call center data from Amazon Redshift using a Python shell in AWS Glu
C. Perform the join with AWS Glue ETL scripts.
D. Create an external table using Amazon Redshift Spectrum for the call center data and perform the join with Amazon Redshift.
E. Export the call center data from Amazon Redshift to Amazon EMR using Apache Sqoo
F. Perform the join with Apache Hive.

Answer: C

Explanation:
https://docs.aws.amazon.com/redshift/latest/dg/c-spectrum-external-tables.html

NEW QUESTION 59
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 62
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 66
A company that produces network devices has millions of users. Data is collected from the devices on an hourly basis and stored in an Amazon S3 data lake.
The company runs analyses on the last 24 hours of data flow logs for abnormality detection and to troubleshoot and resolve user issues. The company also analyzes historical logs dating back 2 years to discover patterns and look for improvement opportunities.
The data flow logs contain many metrics, such as date, timestamp, source IP, and target IP. There are about 10 billion events every day.
How should this data be stored for optimal performance?

A. In Apache ORC partitioned by date and sorted by source IP
B. In compressed .csv partitioned by date and sorted by source IP
C. In Apache Parquet partitioned by source IP and sorted by date
D. In compressed nested JSON partitioned by source IP and sorted by date

Answer: A

NEW QUESTION 68
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 69
A company is planning to do a proof of concept for a machine learning (ML) project using Amazon SageMaker with a subset of existing on-premises data hosted in the company's 3 TB data warehouse. For part of the project, AWS Direct Connect is established and tested. To prepare the data for ML, data analysts are performing data curation. The data analysts want to perform multiple step, including mapping, dropping null fields, resolving choice, and splitting fields. The company needs the fastest solution to curate the data for this project.
Which solution meets these requirements?

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

H. Upload Snowball data into Amazon S3 and schedule data curation jobs using AWS Batch to prepare the data for ML.

Answer: C

NEW QUESTION 72
A company wants to improve the data load time of a sales data dashboard. Data has been collected as .csv files and stored within an Amazon S3 bucket that is partitioned by date. The data is then loaded to an Amazon Redshift data warehouse for frequent analysis. The data volume is up to 500 GB per day.
Which solution will improve the data loading performance?

A. Compress .csv files and use an INSERT statement to ingest data into Amazon Redshift.
B. Split large .csv files, then use a COPY command to load data into Amazon Redshift.
C. Use Amazon Kinesis Data Firehose to ingest data into Amazon Redshift.
D. Load the .csv files in an unsorted key order and vacuum the table in Amazon Redshift.

Answer: B

Explanation:
https://docs.aws.amazon.com/redshift/latest/dg/c_loading-data-best-practices.html

NEW QUESTION 74
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 78
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 82
A data analyst is designing a solution to interactively query datasets with SQL using a JDBC connection. Users will join data stored in Amazon S3 in Apache ORC format with data stored in Amazon Elasticsearch Service (Amazon ES) and Amazon Aurora MySQL.
Which solution will provide the MOST up-to-date results?

A. Use AWS Glue jobs to ETL data from Amazon ES and Aurora MySQL to Amazon S3. Query the data with Amazon Athena.
B. Use Amazon DMS to stream data from Amazon ES and Aurora MySQL to Amazon Redshif
C. Query the data with Amazon Redshift.
D. Query all the datasets in place with Apache Spark SQL running on an AWS Glue developer endpoint.
E. Query all the datasets in place with Apache Presto running on Amazon EMR.

Answer: C

NEW QUESTION 87
.....

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