

Amazon-Web-Services

Exam Questions DAS-C01

AWS Certified Data Analytics - Specialty



NEW QUESTION 1

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 2

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

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

Answer: D

Explanation:

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

NEW QUESTION 3

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

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

Answer: C

NEW QUESTION 4

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

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

application monitoring Use Amazon QuickSight for the dashboards

E. Use Amazon Kinesis Data Analytics as the data transport layer for logging data and to uncover NRT monitoring metrics Use Amazon Kinesis Data Firehose to deliver log data to Amazon OpenSearch Service (Amazon Elasticsearch Service) for search, log analytics, and application monitoring Use OpenSearch Dashboards (Kibana) in Amazon OpenSearch Service (Amazon Elasticsearch Service) for the dashboards

F. Use Amazon Kinesis Data Firehose as the data transport layer for logging data Use Amazon Kinesis Data Analytics to uncover NRT monitoring metrics Use Amazon Kinesis Data Streams to deliver logdata to Amazon OpenSearch Service (Amazon Elasticsearch Service) for search, log analytics, and application monitoring Use Amazon QuickSight for the dashboards.

Answer: C

NEW QUESTION 5

A company uses Amazon Kinesis Data Streams to ingest and process customer behavior information from application users each day. A data analytics specialist notices that its data stream is throttling. The specialist has turned on enhanced monitoring for the Kinesis data stream and has verified that the data stream did not exceed the data limits. The specialist discovers that there are hot shards

Which solution will resolve this issue?

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

Answer: A

NEW QUESTION 6

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 7

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

How should the data analyst resolve the issue?

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

Answer: B

NEW QUESTION 8

A 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 data
- 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 data
- 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 9

A company wants to optimize the cost of its data and analytics platform. The company is ingesting a number of .csv 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 format
- B. Use Athena to query the processed dataset
- 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 format
- E. Use Athena to query the processed dataset
- 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 creation
- 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 format
- I. Use Athena to query the processed dataset
- 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 accessed
- K. 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.
- L. Use an AWS Glue ETL job to partition and convert the data into a row-based data format
- M. Use Athena to query the processed dataset
- 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 accessed
- 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 10

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

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

Answer: B

NEW QUESTION 10

A large ride-sharing company has thousands of drivers globally serving millions of unique customers every day. The company has decided to migrate an existing data mart to Amazon Redshift. The existing schema includes the following tables.

A trips fact table for information on completed rides. A drivers dimension table for driver profiles. A customers fact table holding customer profile information. The company analyzes trip details by date and destination to examine profitability by region. The drivers data rarely changes. The customers data frequently changes.

What table design provides optimal query performance?

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

Answer: C

Explanation:

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

NEW QUESTION 14

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

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

Which storage service will provide the MOST performant solution?

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

Answer: B

NEW QUESTION 16

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

Which solution should the data engineer to meet this compliance requirement with LEAST amount of effort?

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

Answer: C

Explanation:

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

NEW QUESTION 20

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

Which method should the company use to collect and analyze the logs?

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

Answer: D

NEW QUESTION 25

A company has 1 million scanned documents stored as image files in Amazon S3. The documents contain typewritten application forms with information including the applicant first name, applicant last name, application date, application type, and application text. The company has developed a machine learning algorithm to extract the metadata values from the scanned documents. The company wants to allow internal data analysts to analyze and find applications using the applicant name, application date, or application text. The original images should also be downloadable. Cost control is secondary to query performance.

Which solution organizes the images and metadata to drive insights while meeting the requirements?

- A. For each image, use object tags to add the metadata
- B. Use Amazon S3 Select to retrieve the files based on the applicant name and application date.
- C. Index the metadata and the Amazon S3 location of the image file in Amazon Elasticsearch Service. Allow the data analysts to use Kibana to submit queries to the Elasticsearch cluster.
- D. Store the metadata and the Amazon S3 location of the image file in an Amazon Redshift table
- E. Allow the data analysts to run ad-hoc queries on the table.
- F. Store the metadata and the Amazon S3 location of the image files in an Apache Parquet file in Amazon S3, and define a table in the AWS Glue Data Catalog
- G. Allow data analysts to use Amazon Athena to submit custom queries.

Answer: B

Explanation:

<https://aws.amazon.com/blogs/machine-learning/automatically-extract-text-and-structured-data-from-documents>

NEW QUESTION 29

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

Which solution meets these requirements?

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

Answer: B

NEW QUESTION 30

A large financial company is running its ETL process. Part of this process is to move data from Amazon S3 into an Amazon Redshift cluster. The company wants to use the most cost-efficient method to load the dataset into Amazon Redshift. Which combination of steps would meet these requirements? (Choose two.)

- A. Use the COPY command with the manifest file to load data into Amazon Redshift.
- B. Use S3DistCp to load files into Amazon Redshift.
- C. Use temporary staging tables during the loading process.
- D. Use the UNLOAD command to upload data into Amazon Redshift.
- E. Use Amazon Redshift Spectrum to query files from Amazon S3.

Answer: AC

NEW QUESTION 33

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 35

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

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

Answer: D

NEW QUESTION 36

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 to 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 41

A company wants to provide its data analysts with uninterrupted access to the data in its Amazon Redshift cluster. All data is streamed to an Amazon S3 bucket with Amazon Kinesis Data Firehose. An AWS Glue job that is scheduled to run every 5 minutes issues a COPY command to move the data into Amazon Redshift. The amount of data delivered is uneven throughout the day, and cluster utilization is high during certain periods. The COPY command usually completes within a couple of seconds. However, when a load spike occurs, locks can exist and data can be missed. Currently, the AWS Glue job is configured to run without retries, with a timeout at 5 minutes and concurrency at 1. How should a data analytics specialist configure the AWS Glue job to optimize fault tolerance and improve data availability in the Amazon Redshift cluster?

- A. Increase the number of retries.
- B. Decrease the timeout value.
- C. Increase the job concurrency.
- D. Keep the number of retries at 0. Decrease the timeout value.
- E. Increase the job concurrency.
- F. Keep the number of retries at 0. Decrease the timeout value.
- G. Keep the job concurrency at 1.
- H. Keep the number of retries at 0. Increase the timeout value.

I. Keep the job concurrency at 1.

Answer: B

NEW QUESTION 44

A company is migrating from an on-premises Apache Hadoop cluster to an Amazon EMR cluster. The cluster runs only during business hours. Due to a company requirement to avoid intraday cluster failures, the EMR cluster must be highly available. When the cluster is terminated at the end of each business day, the data must persist.

Which configurations would enable the EMR cluster to meet these requirements? (Choose three.)

- A. EMR File System (EMRFS) for storage
- B. Hadoop Distributed File System (HDFS) for storage
- C. AWS Glue Data Catalog as the metastore for Apache Hive
- D. MySQL database on the master node as the metastore for Apache Hive
- E. Multiple master nodes in a single Availability Zone
- F. Multiple master nodes in multiple Availability Zones

Answer: ACE

Explanation:

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

NEW QUESTION 45

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 Service
- 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 46

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 47

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 Firehose
- B. Use AWS Lambda to transform the data in the Kinesis Data Firehose delivery stream and enrich it with the data in the DynamoDB table
- 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 CLI
- 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 data
- 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 source
- 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 CLI
- L. Use Apache Spark SQL on Amazon EMR to read the logs from Amazon S3 and enrich the records with the data from DynamoDB
- M. Store the enriched data in Amazon S3.

Answer: A

Explanation:

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

NEW QUESTION 49

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

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

These weather stations were placed by onsite subject-matter experts.

Each sensor has a unique ID. The data collected from each sensor will be collected using Amazon Kinesis Data Streams.

Based on the total incoming and outgoing data throughput, a single Amazon Kinesis data stream with two shards is created. Two partition keys are created based on the station names. During testing, there is a bottleneck on data coming from Station A, but not from Station B. Upon review, it is confirmed that the total stream throughput is still less than the allocated Kinesis Data Streams throughput.

How can this bottleneck be resolved without increasing the overall cost and complexity of the solution, while retaining the data collection quality requirements?

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

Answer: C

Explanation:

<https://docs.aws.amazon.com/streams/latest/dev/kinesis-using-sdk-java-resharding.html>

"Splitting increases the number of shards in your stream and therefore increases the data capacity of the stream. Because you are charged on a per-shard basis, splitting increases the cost of your stream"

NEW QUESTION 52

A company wants to research user turnover by analyzing the past 3 months of user activities. With millions of users, 1.5 TB of uncompressed data is generated each day. A 30-node Amazon Redshift cluster with 2.56 TB of solid state drive (SSD) storage for each node is required to meet the query performance goals. The company wants to run an additional analysis on a year's worth of historical data to examine trends indicating which features are most popular. This analysis will be done once a week.

What is the MOST cost-effective solution?

- A. Increase the size of the Amazon Redshift cluster to 120 nodes so it has enough storage capacity to hold 1 year of data
- B. Then use Amazon Redshift for the additional analysis.
- C. Keep the data from the last 90 days in Amazon Redshift
- D. Move data older than 90 days to Amazon S3 and store it in Apache Parquet format partitioned by date
- E. Then use Amazon Redshift Spectrum for the additional analysis.
- F. Keep the data from the last 90 days in Amazon Redshift
- G. Move data older than 90 days to Amazon S3 and store it in Apache Parquet format partitioned by date
- H. Then provision a persistent Amazon EMR cluster and use Apache Presto for the additional analysis.
- I. Resize the cluster node type to the dense storage node type (DS2) for an additional 16 TB storage capacity on each individual node in the Amazon Redshift cluster
- J. Then use Amazon Redshift for the additional analysis.

Answer: B

NEW QUESTION 57

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 Routing
- B. Enable VPC Flow Logs to monitor traffic.
- C. Allow access to the Amazon Redshift database using AWS IAM only
- 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 62

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 63

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

- A. Increase the memory of the Amazon ES master nodes.
- B. Decrease the number of Amazon ES data nodes.
- C. Decrease the number of Amazon ES shards for the index.
- D. Increase the number of Amazon ES shards for the index.

Answer: C

Explanation:

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

NEW QUESTION 65

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

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

Answer: A

Explanation:

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

NEW QUESTION 66

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 PHI
- 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 70

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 through 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 cluster
- 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 CustomAmiId property in the CloudFormation template.
- G. Use a CloudFormation template to launch an EMR cluster
- 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 75

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

time.

Which solution will allow the company to collect data for processing while meeting these requirements?

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

Answer: D

Explanation:

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

NEW QUESTION 79

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 80

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 Parquet
- B. Use AWS Glue crawlers to discover the schema and build the AWS Glue Data Catalog
- C. Use Amazon Athena to create a table with a subset of columns
- 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 Redshift
- 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 Parquet
- H. Use AWS Glue crawlers to discover the schema and build the AWS Glue Data Catalog
- 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 calls
- K. Connect Amazon QuickSight to Kinesis Data Analytics to visualize the anomaly scores.

Answer: A

NEW QUESTION 83

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.SQLException: 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 87

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 daily 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 Glue.
- 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 Sqoop.
- 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 89

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 jobs.
- 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 92

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

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

Answer: B

Explanation:

<https://docs.aws.amazon.com/athena/latest/ug/user-created-workgroups.html>

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

NEW QUESTION 93

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 94

A bank is using Amazon Managed Streaming for Apache Kafka (Amazon MSK) to populate real-time data into a data lake. The data lake is built on Amazon S3, and data must be accessible from the data lake within 24 hours. Different microservices produce messages to different topics in the cluster. The cluster is created with 8 TB of Amazon Elastic Block Store (Amazon EBS) storage and a retention period of 7 days. The customer transaction volume has tripled recently and disk monitoring has provided an alert that the cluster is almost out of storage capacity. What should a data analytics specialist do to prevent the cluster from running out of disk space?

- A. Use the Amazon MSK console to triple the broker storage and restart the cluster.
- B. Create an Amazon CloudWatch alarm that monitors the `KafkaDataLogsDiskUsed` metric. Automatically flush the oldest messages when the value of this metric exceeds 85%.
- C. Create a custom Amazon MSK configuration. Set the log retention hours parameter to 48. Update the cluster with the new configuration file.
- D. Triple the number of consumers to ensure that data is consumed as soon as it is added to a topic.

Answer: B

NEW QUESTION 98

A marketing company has data in Salesforce, MySQL, and Amazon S3. The company wants to use data from these three locations and create mobile dashboards for its users. The company is unsure how it should create the dashboards and needs a solution with the least possible customization and coding. Which solution meets these requirements?

- A. Use Amazon Athena federated queries to join the data source
- B. Use Amazon QuickSight to generate the mobile dashboards.
- C. Use AWS Lake Formation to migrate the data sources into Amazon S3. Use Amazon QuickSight to generate the mobile dashboards.
- D. Use Amazon Redshift federated queries to join the data source
- E. Use Amazon QuickSight to generate the mobile dashboards.
- F. Use Amazon QuickSight to connect to the data sources and generate the mobile dashboards.

Answer: C

NEW QUESTION 100

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

Which solution achieves these required access patterns to minimize costs and administrative tasks?

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

Answer: C

Explanation:

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

NEW QUESTION 105

A company uses an Amazon EMR cluster with 50 nodes to process operational data and make the data available for data analysts. These jobs run nightly use Apache Hive with the Apache Tez framework as a processing model and write results to Hadoop Distributed File System (HDFS). In the last few weeks, jobs are failing and are producing the following error message: "File could only be replicated to 0 nodes instead of 1"

A data analytics specialist checks the DataNode logs, the NameNode logs, and network connectivity for potential issues that could have prevented HDFS from replicating data. The data analytics specialist rules out these factors as causes for the issue.

Which solution will prevent the jobs from failing?

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

Answer: C

NEW QUESTION 106

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 107

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 111

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 dataset
- 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 dataset
- 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 112

A company has an application that uses the Amazon Kinesis Client Library (KCL) to read records from a Kinesis data stream.

After a successful marketing campaign, the application experienced a significant increase in usage. As a result, a data analyst had to split some shards in the data stream. When the shards were split, the application started throwing an `ExpiredIteratorExceptions` error sporadically. What should the data analyst do to resolve this?

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

Answer: C

NEW QUESTION 113

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

Which program modification will accelerate the COPY process?

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

Answer: B

NEW QUESTION 118

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 class
- 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 class
- 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 class
- 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 class
- H. Apply a lifecycle configuration that changes the storage class to Amazon S3 Standard-Infrequent Access (S3 Standard-IA) 1 year after data creation.
- I. Store the daily roll-up data initially in the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class
- J. Apply a lifecycle configuration that changes the storage class to Amazon S3 Glacier 1 year after data creation.

Answer: AD

NEW QUESTION 119

A manufacturing company has many IoT devices in different facilities across the world. The company is using Amazon Kinesis Data Streams to collect the data from the devices.

The company's operations team has started to observe many `WroteThroughputExceeded` exceptions. The operations team determines that the reason is the number of records that are being written to certain shards. The data contains device ID, capture date, measurement type, measurement value, and facility ID. The facility ID is used as the partition key.

Which action will resolve this issue?

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

Answer: B

NEW QUESTION 120

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 the archived 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 123

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 steps, including mapping, dropping null fields, resolving choices, 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 scripts to curate the data in an Amazon EMR cluster.
- B. Store the curated data in Amazon S3 for ML processing.
- C. Create custom ETL jobs on-premises to curate the data.
- D. Use AWS DMS to ingest data into Amazon S3 for ML processing.
- E. Ingest data into Amazon S3 using AWS Data Pipeline.
- 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 Snowball.
- 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 128

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

Which actions should the data analyst take?

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

D. Enable job bookmarks in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the num-executors job parameter.

Answer: B

NEW QUESTION 132

A marketing company collects clickstream data. The company sends the data to Amazon Kinesis Data Firehose and stores the data in Amazon S3. The company wants to build a series of dashboards that will be used by hundreds of users across different departments. The company will use Amazon QuickSight to develop these dashboards. The company has limited resources and wants a solution that could scale and provide daily updates about clickstream activity. Which combination of options will provide the MOST cost-effective solution? (Select TWO.)

- A. Use Amazon Redshift to store and query the clickstream data.
- B. Use QuickSight with a direct SQL query.
- C. Use Amazon Athena to query the clickstream data in Amazon S3.
- D. Use S3 analytics to query the clickstream data.
- E. Use the QuickSight SPICE engine with a daily refresh.

Answer: BD

NEW QUESTION 136

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

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

Answer: A

Explanation:

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

NEW QUESTION 141

A company has developed an Apache Hive script to batch process data stored 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 step.
- B. Set `KeepJobFlowAliveWhenNoSteps` to false and disable the termination protection flag.
- 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, Hive, and Apache Oozie.
- E. Hive, and Apache Oozie.
- F. Set the termination protection flag to true and use Spot Instances for the core nodes of the cluster.
- 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 operation.
- 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 142

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 simultaneously.
- 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 value.
- 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 the 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 value.
- F. Use AWS Lambda to publish an Amazon Simple Notification Service (Amazon SNS) notification if the threshold is not met.
- 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 Redshift.
- 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 simultaneously.
- 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 147

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 149

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 table.
- B. Load the columns that contain non-sensitive data into a second table.
- 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 IAM 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 150

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

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

Answer: A

NEW QUESTION 155

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 Redshift.
- 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 Spectrum.
- 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 needed.
- G. Configure Amazon QuickSight with Amazon EMR as the data source.

Answer: B

NEW QUESTION 156

An ecommerce company stores customer purchase data in Amazon RDS. The company wants a solution to store and analyze historical data. The most recent 6 months of data will be queried frequently for analytics workloads. This data is several terabytes large. Once a month, historical data for the last 5 years must be

accessible and will be joined with the more recent data. The company wants to optimize performance and cost. Which storage solution will meet these requirements?

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

Answer: D

NEW QUESTION 161

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 164

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 165

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 QuickSight 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 166

A reseller that has thousands of AWS accounts receives AWS Cost and Usage Reports in an Amazon S3 bucket. The reports are delivered to the S3 bucket in the following format:

<example-report-prefix>/<example-report-name>/yyyymmdd-yyyymmdd/<example-report-name>.parquet

An AWS Glue crawler crawls the S3 bucket and populates an AWS Glue Data Catalog with a table. Business analysts use Amazon Athena to query the table and create monthly summary reports for the AWS accounts.

The business analysts are experiencing slow queries because of the accumulation of reports from the last 5 years. The business analysts want the operations team to make changes to improve query performance.

Which action should the operations team take to meet these requirements?

- A. Change the file format to csv.zip.
- B. Partition the data by date and account ID.
- C. Partition the data by month and account ID.
- D. Partition the data by account ID, year, and month.

Answer: B

NEW QUESTION 167

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

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

Answer: D

NEW QUESTION 168

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 Redshift
- 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 171

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 of 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 172

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 nodes
- B. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch CapacityRemainingGB metric
- 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 nodes
- E. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch YARNMemoryAvailablePercentage metric
- 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 nodes
- H. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch CapacityRemainingGB metric
- I. Create an automatic scaling policy to scale in the instance groups based on the CloudWatch CapacityRemainingGB metric.
- J. Create instance group configurations for core and task nodes
- K. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch YARNMemoryAvailablePercentage metric
- 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 173

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 178

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 catalog
- B. Create AWS Lambda functions that will connect and gather the metadata information from multiple sources and update the data catalog in Aurora
- C. Schedule the Lambda functions periodically.
- D. Use the AWS Glue Data Catalog as the central metadata repository
- 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 catalog
- H. Create AWS Lambda functions that will connect and gather the metadata information from multiple sources and update the DynamoDB catalog
- I. Schedule the Lambda functions periodically.
- J. Use the AWS Glue Data Catalog as the central metadata repository
- K. Extract the schema for RDS and Amazon Redshift sources and build the Data Catalog
- L. Use AWS crawlers for data stored in Amazon S3 to infer the schema and automatically update the Data Catalog.

Answer: D

NEW QUESTION 181

A company is sending historical datasets to Amazon S3 for storage. A data engineer at the company wants to make these datasets available for analysis using Amazon Athena. The engineer also wants to encrypt the Athena query results in an S3 results location by using AWS solutions for encryption. The requirements for encrypting the query results are as follows:

Use custom keys for encryption of the primary dataset query results. Use generic encryption for all other query results.

Provide an audit trail for the primary dataset queries that shows when the keys were used and by whom.

Which solution meets these requirements?

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

Answer: A

NEW QUESTION 185

A utility company wants to visualize data for energy usage on a daily basis in Amazon QuickSight. A data analytics specialist at the company has built a data pipeline to collect and ingest the data into Amazon S3. Each day the data is stored in an individual CSV file in an S3 bucket. This is an example of the naming structure: 20210707_data.csv, 20210708_data.csv.

To allow for data querying in QuickSight through Amazon Athena, the specialist used an AWS Glue crawler to create a table with the path "s3://powertransformer/20210707_data.csv". However, when the data is queried, it returns zero rows.

How can this issue be resolved?

- A. Modify the IAM policy for the AWS Glue crawler to access Amazon S3.
- B. Ingest the files again.
- C. Store the files in Apache Parquet format.
- D. Update the table path to "s3://powertransformer/".

Answer: D

NEW QUESTION 188

A human resources company maintains a 10-node Amazon Redshift cluster to run analytics queries on the company's data. The Amazon Redshift cluster contains a product table and a transactions table, and both tables have a product_sku column. The tables are over 100 GB in size. The majority of queries run on both tables.

Which distribution style should the company use for the two tables to achieve optimal query performance?

- A. An EVEN distribution style for both tables
- B. A KEY distribution style for both tables
- C. An ALL distribution style for the product table and an EVEN distribution style for the transactions table
- D. An EVEN distribution style for the product table and a KEY distribution style for the transactions table

Answer: B

NEW QUESTION 192

.....

Thank You for Trying Our Product

We offer two products:

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

2nd - Questions 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

100% Actual & Verified — Instant Download, Please Click
[Order The DAS-C01 Practice Test Here](#)