

Exam Questions AWS-Certified-Data-Analytics-Specialty

AWS Certified Data Analytics - Specialty

<https://www.2passeasy.com/dumps/AWS-Certified-Data-Analytics-Specialty/>



NEW QUESTION 1

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

Which solution meets these requirements?

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

Answer: C

NEW QUESTION 2

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

What should the data analyst do to achieve this?

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

Answer: B

Explanation:

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

NEW QUESTION 3

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 4

A global company has different sub-organizations, and each sub-organization sells its products and services in various countries. The company's senior leadership wants to quickly identify which sub-organization is the strongest performer in each country. All sales data is stored in Amazon S3 in Parquet format.

Which approach can provide the visuals that senior leadership requested with the least amount of effort?

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

Answer: A

NEW QUESTION 5

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

Which solution will meet the company's requirements?

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

Answer: B

NEW QUESTION 6

An advertising company has a data lake that is built on Amazon S3. The company uses AWS Glue Data Catalog to maintain the metadata. The data lake is several years old and its overall size has increased exponentially as additional data sources and metadata are stored in the data lake. The data lake administrator wants to implement a mechanism to simplify permissions management between Amazon S3 and the Data Catalog to keep them in sync

Which solution will simplify permissions management with minimal development effort?

- A. Set AWS Identity and Access Management (IAM) permissions for AWS Glue
- B. Use AWS Lake Formation permissions
- C. Manage AWS Glue and S3 permissions by using bucket policies
- D. Use Amazon Cognito user pools.

Answer: B

NEW QUESTION 7

An online retail company is migrating its reporting system to AWS. The company's legacy system runs data processing on online transactions using a complex series of nested Apache Hive queries. Transactional data is exported from the online system to the reporting system several times a day. Schemas in the files are stable between updates.

A data analyst wants to quickly migrate the data processing to AWS, so any code changes should be minimized. To keep storage costs low, the data analyst decides to store the data in Amazon S3. It is vital that the data from the reports and associated analytics is completely up to date based on the data in Amazon S3. Which solution meets these requirements?

- A. Create an AWS Glue Data Catalog to manage the Hive metadata
- B. Create an AWS Glue crawler over Amazon S3 that runs when data is refreshed to ensure that data changes are updated
- C. Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
- D. Create an AWS Glue Data Catalog to manage the Hive metadata
- E. Create an Amazon EMR cluster with consistent view enabled
- F. Run emrfs sync before each analytics step to ensure data changes are updated
- G. Create an EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
- H. Create an Amazon Athena table with CREATE TABLE AS SELECT (CTAS) to ensure data is refreshed from underlying queries against the raw dataset
- I. Create an AWS Glue Data Catalog to manage the Hive metadata over the CTAS table
- J. Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
- K. Use an S3 Select query to ensure that the data is properly updated
- L. Create an AWS Glue Data Catalog to manage the Hive metadata over the S3 Select table
- M. Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.

Answer: A

NEW QUESTION 8

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 9

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 10

A company has developed several AWS Glue jobs to validate and transform its data from Amazon S3 and load it into Amazon RDS for MySQL in batches once every day. The ETL jobs read the S3 data using a DynamicFrame. Currently, the ETL developers are experiencing challenges in processing only the incremental data on every run, as the AWS Glue job processes all the S3 input data on each run.

Which approach would allow the developers to solve the issue with minimal coding effort?

- A. Have the ETL jobs read the data from Amazon S3 using a DataFrame.
- B. Enable job bookmarks on the AWS Glue jobs.
- C. Create custom logic on the ETL jobs to track the processed S3 objects.
- D. Have the ETL jobs delete the processed objects or data from Amazon S3 after each run.

Answer: B

NEW QUESTION 10

A real estate company has a mission-critical application using Apache HBase in Amazon EMR. Amazon EMR is configured with a single master node. The company has over 5 TB of data stored on an Hadoop Distributed File System (HDFS). The company wants a cost-effective solution to make its HBase data highly available. Which architectural pattern meets the company's requirements?

- A. Use Spot Instances for core and task nodes and a Reserved Instance for the EMR master node. Configure the EMR cluster with multiple master node
- B. Schedule automated snapshots using Amazon EventBridge.
- C. Store the data on an EMR File System (EMRFS) instead of HDF
- D. Enable EMRFS consistent view. Create an EMR HBase cluster with multiple master node
- E. Point the HBase root directory to an Amazon S3 bucket.
- F. Store the data on an EMR File System (EMRFS) instead of HDFS and enable EMRFS consistent view. Run two separate EMR clusters in two different Availability Zone
- G. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.
- H. Store the data on an EMR File System (EMRFS) instead of HDFS and enable EMRFS consistent view. Create a primary EMR HBase cluster with multiple master node
- I. Create a secondary EMR HBase read- replica cluster in a separate Availability Zon
- J. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.

Answer: D

NEW QUESTION 12

A marketing company wants to improve its reporting and business intelligence capabilities. During the planning phase, the company interviewed the relevant stakeholders and discovered that:

- The operations team reports are run hourly for the current month's data.
- The sales team wants to use multiple Amazon QuickSight dashboards to show a rolling view of the last 30 days based on several categories.
- The sales team also wants to view the data as soon as it reaches the reporting backend.
- The finance team's reports are run daily for last month's data and once a month for the last 24 months of data.

Currently, there is 400 TB of data in the system with an expected additional 100 TB added every month. The company is looking for a solution that is as cost-effective as possible.

Which solution meets the company's requirements?

- A. Store the last 24 months of data in Amazon Redshif
- B. Configure Amazon QuickSight with Amazon Redshift as the data source.
- C. Store the last 2 months of data in Amazon Redshift and the rest of the months in Amazon S3. Set up an external schema and table for Amazon Redshift Spectru
- D. Configure Amazon QuickSight with Amazon Redshift as the data source.
- E. Store the last 24 months of data in Amazon S3 and query it using Amazon Redshift Spectrum. Configure Amazon QuickSight with Amazon Redshift Spectrum as the data source.
- F. Store the last 2 months of data in Amazon Redshift and the rest of the months in Amazon S3. Use a long- running Amazon EMR with Apache Spark cluster to query the data as neede
- G. Configure Amazon QuickSight with Amazon EMR as the data source.

Answer: B

NEW QUESTION 14

A marketing company is using Amazon EMR clusters for its workloads. The company manually installs third- party libraries on the clusters by logging in to the master nodes. A data analyst needs to create an automated solution to replace the manual process.

Which options can fulfill these requirements? (Choose two.)

- A. Place the required installation scripts in Amazon S3 and execute them using custom bootstrap actions.
- B. Place the required installation scripts in Amazon S3 and execute them through Apache Spark in Amazon EMR.
- C. Install the required third-party libraries in the existing EMR master nod
- D. Create an AMI out of that master node and use that custom AMI to re-create the EMR cluster.
- E. Use an Amazon DynamoDB table to store the list of required application
- F. Trigger an AWS Lambda function with DynamoDB Streams to install the software.
- G. Launch an Amazon EC2 instance with Amazon Linux and install the required third-party libraries on the instanc
- H. Create an AMI and use that AMI to create the EMR cluster.

Answer: AE

Explanation:

<https://aws.amazon.com/about-aws/whats-new/2017/07/amazon-emr-now-supports-launching-clusters-with-cust>
https://docs.aws.amazon.com/de_de/emr/latest/ManagementGuide/emr-plan-bootstrap.html

NEW QUESTION 15

A mortgage company has a microservice for accepting payments. This microservice uses the Amazon DynamoDB encryption client with AWS KMS managed keys to encrypt the sensitive data before writing the data to DynamoDB. The finance team should be able to load this data into Amazon Redshift and aggregate the values within the sensitive fields. The Amazon Redshift cluster is shared with other data analysts from different business units.

Which steps should a data analyst take to accomplish this task efficiently and securely?

- A. Create an AWS Lambda function to process the DynamoDB strea
- B. Decrypt the sensitive data using the same KMS ke
- C. Save the output to a restricted S3 bucket for the finance tea
- D. Create a finance table in Amazon Redshift that is accessible to the finance team onl
- E. Use the COPY command to load the data from Amazon S3 to the finance table.
- F. Create an AWS Lambda function to process the DynamoDB strea
- G. Save the output to a restricted S3 bucket for the finance tea
- H. Create a finance table in Amazon Redshift that is accessible to the finance team onl
- I. Use the COPY command with the IAM role that has access to the KMS key to load the data from S3 to the finance table.
- J. Create an Amazon EMR cluster with an EMR_EC2_DefaultRole role that has access to the KMS key. Create Apache Hive tables that reference the data stored in DynamoDB and the finance table in Amazon Redshif
- K. In Hive, select the data from DynamoDB and then insert the output to the finance table in Amazon Redshift.

- L. Create an Amazon EMR cluste
- M. Create Apache Hive tables that reference the data stored in DynamoD
- N. Insert the output to the restricted Amazon S3 bucket for the finance tea
- O. Use the COPY command with the IAM role that has access to the KMS key to load the data from Amazon S3 to the finance table in Amazon Redshift.

Answer: B

NEW QUESTION 20

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

Which solution meets these requirements?

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

Answer: D

Explanation:

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

NEW QUESTION 23

A company operates toll services for highways across the country and collects data that is used to understand usage patterns. Analysts have requested the ability to run traffic reports in near-real time. The company is interested in building an ingestion pipeline that loads all the data into an Amazon Redshift cluster and alerts operations personnel when toll traffic for a particular toll station does not meet a specified threshold. Station data and the corresponding threshold values are stored in Amazon S3.

Which approach is the MOST efficient way to meet these requirements?

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

Answer: D

NEW QUESTION 25

A company uses Amazon Redshift for its data warehousing needs. ETL jobs run every night to load data, apply business rules, and create aggregate tables for reporting. The company's data analysis, data science, and business intelligence teams use the data warehouse during regular business hours. The workload management is set to auto, and separate queues exist for each team with the priority set to NORMAL.

Recently, a sudden spike of read queries from the data analysis team has occurred at least twice daily, and queries wait in line for cluster resources. The company needs a solution that enables the data analysis team to avoid query queuing without impacting latency and the query times of other teams.

Which solution meets these requirements?

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

Answer: D

NEW QUESTION 27

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 31

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 33

A media company wants to perform machine learning and analytics on the data residing in its Amazon S3 data lake. There are two data transformation requirements that will enable the consumers within the company to create reports:

- Daily transformations of 300 GB of data with different file formats landing in Amazon S3 at a scheduled time.
- One-time transformations of terabytes of archived data residing in the S3 data lake.

Which combination of solutions cost-effectively meets the company's requirements for transforming the data? (Choose three.)

- A. For daily incoming data, use AWS Glue crawlers to scan and identify the schema.
- B. For daily incoming data, use Amazon Athena to scan and identify the schema.
- C. For daily incoming data, use Amazon Redshift to perform transformations.
- D. For daily incoming data, use AWS Glue workflows with AWS Glue jobs to perform transformations.
- E. For archived data, use Amazon EMR to perform data transformations.
- F. For archived data, use Amazon SageMaker to perform data transformations.

Answer: ADE

NEW QUESTION 36

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

Which solution meets these requirements?

- A. Ingest data into Amazon S3 using AWS DataSync and use Apache Spark 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 DM.
- F. Use AWS Glue to perform data curation and store the data in Amazon S3 for ML processing.
- G. Take a full backup of the data store and ship the backup files using AWS 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 40

A company needs to collect streaming data from several sources and store the data in the AWS Cloud. The dataset is heavily structured, but analysts need to perform several complex SQL queries and need consistent performance. Some of the data is queried more frequently than the rest. The company wants a solution that meets its performance requirements in a cost-effective manner.

Which solution meets these requirements?

- A. Use Amazon Managed Streaming for Apache Kafka to ingest the data to save it to Amazon S3. Use Amazon Athena to perform SQL queries over the ingested data.
- B. Use Amazon Managed Streaming for Apache Kafka to ingest the data to save it to Amazon Redshift. Enable Amazon Redshift workload management (WLM) to prioritize workloads.
- C. Use Amazon Kinesis Data Firehose to ingest the data to save it to Amazon Redshift.
- D. Enable Amazon Redshift workload management (WLM) to prioritize workloads.
- E. Use Amazon Kinesis Data Firehose to ingest the data to save it to Amazon S3. Load frequently queried data to Amazon Redshift using the COPY command.
- F. Use Amazon Redshift Spectrum for less frequently queried data.

Answer: B

NEW QUESTION 45

A global pharmaceutical company receives test results for new drugs from various testing facilities worldwide. The results are sent in millions of 1 KB-sized JSON

objects to an Amazon S3 bucket owned by the company. The data engineering team needs to process those files, convert them into Apache Parquet format, and load them into Amazon Redshift for data analysts to perform dashboard reporting. The engineering team uses AWS Glue to process the objects, AWS Step Functions for process orchestration, and Amazon CloudWatch for job scheduling. More testing facilities were recently added, and the time to process files is increasing. What will MOST efficiently decrease the data processing time?

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

Answer: A

NEW QUESTION 48

A company developed a new elections reporting website that uses Amazon Kinesis Data Firehose to deliver full logs from AWS WAF to an Amazon S3 bucket. The company is now seeking a low-cost option to perform this infrequent data analysis with visualizations of logs in a way that requires minimal development effort. Which solution meets these requirements?

- A. Use an AWS Glue crawler to create and update a table in the Glue data catalog from the log
- B. Use Athena to perform ad-hoc analyses and use Amazon QuickSight to develop data visualizations.
- C. Create a second Kinesis Data Firehose delivery stream to deliver the log files to Amazon Elasticsearch Service (Amazon ES). Use Amazon ES to perform text-based searches of the logs for ad-hoc analyses and use Kibana for data visualizations.
- D. Create an AWS Lambda function to convert the logs into .csv format
- E. Then add the function to the Kinesis Data Firehose transformation configuration
- F. Use Amazon Redshift to perform ad-hoc analyses of the logs using SQL queries and use Amazon QuickSight to develop data visualizations.
- G. Create an Amazon EMR cluster and use Amazon S3 as the data source
- H. Create an Apache Spark job to perform ad-hoc analyses and use Amazon QuickSight to develop data visualizations.

Answer: A

Explanation:

<https://aws.amazon.com/blogs/big-data/analyzing-aws-waf-logs-with-amazon-es-amazon-athena-and-amazon-qu>

NEW QUESTION 53

An ecommerce company ingests a large set of clickstream data in JSON format and stores the data in Amazon S3. Business analysts from multiple product divisions need to use Amazon Athena to analyze the data. The company's analytics team must design a solution to monitor the daily data usage for Athena by each product division. The solution also must produce a warning when a division exceeds its quota. Which solution will meet these requirements with the LEAST operational overhead?

- A. Use a CREATE TABLE AS SELECT (CTAS) statement to create separate tables for each product division. Use AWS Budgets to track Athena usage. Configure a threshold for the budget. Use Amazon Simple Notification Service (Amazon SNS) to send notifications when thresholds are breached.
- B. Create an AWS account for each division. Provide cross-account access to an AWS Glue Data Catalog to all the accounts.
- C. Set an Amazon CloudWatch alarm to monitor Athena usage.
- D. Use Amazon Simple Notification Service (Amazon SNS) to send notifications.
- E. Create an Athena workgroup for each division. Configure a data usage control for each workgroup and a time period of 1 day. Configure an action to send notifications to an Amazon Simple Notification Service (Amazon SNS) topic.
- F. Create an AWS account for each division. Configure an AWS Glue Data Catalog in each account. Set an Amazon CloudWatch alarm to monitor Athena usage. Use Amazon Simple Notification Service (Amazon SNS) to send notifications.

Answer: C

NEW QUESTION 57

A media company has been performing analytics on log data generated by its applications. There has been a recent increase in the number of concurrent analytics jobs running, and the overall performance of existing jobs is decreasing as the number of new jobs is increasing. The partitioned data is stored in Amazon S3 One Zone-Infrequent Access (S3 One Zone-IA) and the analytic processing is performed on Amazon EMR clusters using the EMR File System (EMRFS) with consistent view enabled. A data analyst has determined that it is taking longer for the EMR task nodes to list objects in Amazon S3. Which action would MOST likely increase the performance of accessing log data in Amazon S3?

- A. Use a hash function to create a random string and add that to the beginning of the object prefixes when storing the log data in Amazon S3.
- B. Use a lifecycle policy to change the S3 storage class to S3 Standard for the log data.
- C. Increase the read capacity units (RCUs) for the shared Amazon DynamoDB table.
- D. Redeploy the EMR clusters that are running slowly to a different Availability Zone.

Answer: C

Explanation:

<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emrfs-metadata.html>

NEW QUESTION 60

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

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

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

Answer: A

Explanation:

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

NEW QUESTION 65

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 68

A central government organization is collecting events from various internal applications using Amazon Managed Streaming for Apache Kafka (Amazon MSK). The organization has configured a separate Kafka topic for each application to separate the data. For security reasons, the Kafka cluster has been configured to only allow TLS encrypted data and it encrypts the data at rest.

A recent application update showed that one of the applications was configured incorrectly, resulting in writing data to a Kafka topic that belongs to another application. This resulted in multiple errors in the analytics pipeline as data from different applications appeared on the same topic. After this incident, the organization wants to prevent applications from writing to a topic different than the one they should write to.

Which solution meets these requirements with the least amount of effort?

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

Answer: B

NEW QUESTION 73

A data analyst runs a large number of data manipulation language (DML) queries by using Amazon Athena with the JDBC driver. Recently, a query failed after it ran for 30 minutes. The query returned the following message: `Java.sql.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 76

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 81

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 83

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

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

Answer: B

Explanation:

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

NEW QUESTION 86

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 91

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 93

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 97

A company wants to improve user satisfaction for its smart home system by adding more features to its recommendation engine. Each sensor asynchronously pushes its nested JSON data into Amazon Kinesis Data Streams using the Kinesis Producer Library (KPL) in Java. Statistics from a set of failed sensors showed that, when a sensor is malfunctioning, its recorded data is not always sent to the cloud.

The company needs a solution that offers near-real-time analytics on the data from the most updated sensors. Which solution enables the company to meet these requirements?

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

Answer: B

Explanation:

<https://docs.aws.amazon.com/streams/latest/dev/developing-producers-with-kpl.html>

The KPL can incur an additional processing delay of up to RecordMaxBufferedTime within the library (user-configurable). Larger values of RecordMaxBufferedTime results in higher packing efficiencies and better performance. Applications that cannot tolerate this additional delay may need to use the AWS SDK directly.

NEW QUESTION 100

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 102

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 107

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 108

A company has a business unit uploading .csv files to an Amazon S3 bucket. The company's data platform team has set up an AWS Glue crawler to do discovery, and create tables and schemas. An AWS Glue job writes processed data from the created tables to an Amazon Redshift database. The AWS Glue job handles column mapping and creating the Amazon Redshift table appropriately. When the AWS Glue job is rerun for any reason in a day, duplicate records are introduced into the Amazon Redshift table.

Which solution will update the Redshift table without duplicates when jobs are rerun?

- A. Modify the AWS Glue job to copy the rows into a staging tabl
- B. Add SQL commands to replace the existing rows in the main table as postactions in the DynamicFrameWriter class.
- C. Load the previously inserted data into a MySQL database in the AWS Glue jo
- D. Perform an upsert operation in MySQL, and copy the results to the Amazon Redshift table.
- E. Use Apache Spark's DataFrame dropDuplicates() API to eliminate duplicates and then write the data to Amazon Redshift.
- F. Use the AWS Glue ResolveChoice built-in transform to select the most recent value of the column.

Answer: A

Explanation:

<https://aws.amazon.com/premiumsupport/knowledge-center/sql-commands-redshift-glue-job/> See the section Merge an Amazon Redshift table in AWS Glue (upsert)

NEW QUESTION 110

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

Which solution meets these requirements?

- A. Use an AWS Glue Python shell job to process incoming data from Kinesis Data Streams Use the Boto3 library to write data to Amazon Redshift
- B. Use AWS Glue streaming and Scale to process incoming data from Kinesis Data Streams Use the AWS Glue connector to write data to Amazon Redshift
- C. Use Amazon Kinesis Data Analytics to create an in-application table based upon the reference data Process incoming data from Kinesis Data Streams Use a data stream to write results to Amazon Redshift
- D. Use Amazon Kinesis Data Analytics to create an in-application table based upon the reference dataProcess incoming data from Kinesis Data Streams Use an Amazon Kinesis Data Firehose delivery stream to write results to Amazon Redshift

Answer: D

NEW QUESTION 115

A company using Amazon QuickSight Enterprise edition has thousands of dashboards analyses and datasets. The company struggles to manage and assign permissions for granting users access to various items within QuickSight. The company wants to make it easier to implement sharing and permissions management.

Which solution should the company implement to simplify permissions management?

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

Answer: C

NEW QUESTION 120

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

What should the data analyst do reduce this latency?

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

Answer: D

NEW QUESTION 125

A power utility company is deploying thousands of smart meters to obtain real-time updates about power consumption. The company is using Amazon Kinesis Data Streams to collect the data streams from smart meters. The consumer application uses the Kinesis Client Library (KCL) to retrieve the stream data. The company has only one consumer application.

The company observes an average of 1 second of latency from the moment that a record is written to the stream until the record is read by a consumer application. The company must reduce this latency to 500 milliseconds.

Which solution meets these requirements?

- A. Use enhanced fan-out in Kinesis Data Streams.
- B. Increase the number of shards for the Kinesis data stream.
- C. Reduce the propagation delay by overriding the KCL default settings.
- D. Develop consumers by using Amazon Kinesis Data Firehose.

Answer: C

Explanation:

The KCL defaults are set to follow the best practice of polling every 1 second. This default results in average propagation delays that are typically below 1 second.

NEW QUESTION 130

A company wants to collect and process events data from different departments in near-real time. Before storing the data in Amazon S3, the company needs to clean the data by standardizing the format of the address and timestamp columns. The data varies in size based on the overall load at each particular point in time. A single data record can be 100 KB-10 MB.

How should a data analytics specialist design the solution for data ingestion?

- A. Use Amazon Kinesis Data Stream
- B. Configure a stream for the raw data
- C. Use a Kinesis Agent to write data to the stream
- D. Create an Amazon Kinesis Data Analytics application that reads data from the raw stream, cleanses it, and stores the output to Amazon S3.
- E. Use Amazon Kinesis Data Firehose
- F. Configure a Firehose delivery stream with a preprocessing AWS Lambda function for data cleansing
- G. Use a Kinesis Agent to write data to the delivery stream
- H. Configure Kinesis Data Firehose to deliver the data to Amazon S3.
- I. Use Amazon Managed Streaming for Apache Kafka
- J. Configure a topic for the raw data
- K. Use a Kafka producer to write data to the topic
- L. Create an application on Amazon EC2 that reads data from the topic by using the Apache Kafka consumer API, cleanses the data, and writes to Amazon S3.
- M. Use Amazon Simple Queue Service (Amazon SQS). Configure an AWS Lambda function to read events from the SQS queue and upload the events to Amazon S3.

Answer: B

NEW QUESTION 133

A financial company hosts a data lake in Amazon S3 and a data warehouse on an Amazon Redshift cluster. The company uses Amazon QuickSight to build dashboards and wants to secure access from its on-premises Active Directory to Amazon QuickSight.

How should the data be secured?

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

Answer: A

Explanation:

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

NEW QUESTION 137

A data analytics specialist is building an automated ETL ingestion pipeline using AWS Glue to ingest compressed files that have been uploaded to an Amazon S3 bucket. The ingestion pipeline should support incremental data processing.

Which AWS Glue feature should the data analytics specialist use to meet this requirement?

- A. Workflows
- B. Triggers
- C. Job bookmarks
- D. Classifiers

Answer: C

NEW QUESTION 142

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 146

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

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

Answer: AB

Explanation:

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

NEW QUESTION 148

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 149

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 154

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 158

A company is reading data from various customer databases that run on Amazon RDS. The databases contain many inconsistent fields. For example, a customer record field that is place_id in one database is location_id in another database. The company wants to link customer records across different databases, even when many customer record fields do not match exactly. Which solution will meet these requirements with the LEAST operational overhead?

- A. Create an Amazon EMR cluster to process and analyze data in the databases. Connect to the Apache Zeppelin notebook, and use the FindMatches transform to find duplicate records in the data.
- B. Create an AWS Glue crawler to crawl the database.
- C. Use the FindMatches transform to find duplicate records in the data. Evaluate and tune the transform by evaluating performance and results of finding matches.
- D. Create an AWS Glue crawler to crawl the data in the databases. Use Amazon SageMaker to construct Apache Spark ML pipelines to find duplicate records in the data.
- E. Create an Amazon EMR cluster to process and analyze data in the database.
- F. Connect to the Apache Zeppelin notebook, and use Apache Spark ML to find duplicate records in the data.
- G. Evaluate and tune the model by evaluating performance and results of finding duplicates.

Answer: B

NEW QUESTION 161

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 164

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 165

.....

THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual AWS-Certified-Data-Analytics-Specialty Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the AWS-Certified-Data-Analytics-Specialty Product From:

<https://www.2passeasy.com/dumps/AWS-Certified-Data-Analytics-Specialty/>

Money Back Guarantee

AWS-Certified-Data-Analytics-Specialty Practice Exam Features:

- * AWS-Certified-Data-Analytics-Specialty Questions and Answers Updated Frequently
- * AWS-Certified-Data-Analytics-Specialty Practice Questions Verified by Expert Senior Certified Staff
- * AWS-Certified-Data-Analytics-Specialty Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- * AWS-Certified-Data-Analytics-Specialty Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year