

# 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 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 an KEY distribution style for the transactions table

**Answer:** B

#### NEW QUESTION 2

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

Which approach should the data analytics team take to allow product owners to view only their products in the dashboard?

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

**Answer:** D

#### Explanation:

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

#### NEW QUESTION 3

A company needs to store objects containing log data in JSON format. The objects are generated by eight applications running in AWS. Six of the applications generate a total of 500 KiB of data per second, and two of the applications can generate up to 2 MiB of data per second.

A data engineer wants to implement a scalable solution to capture and store usage data in an Amazon S3

bucket. The usage data objects need to be reformatted, converted to .csv format, and then compressed before they are stored in Amazon S3. The company requires the solution to include the least custom code possible and has authorized the data engineer to request a service quota increase if needed.

Which solution meets these requirements?

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

**Answer:** A

#### NEW QUESTION 4

A transportation company uses IoT sensors attached to trucks to collect vehicle data for its global delivery fleet. The company currently sends the sensor data in small .csv files to Amazon S3. The files are then loaded into a 10-node Amazon Redshift cluster with two slices per node and queried using both Amazon Athena and Amazon Redshift. The company wants to optimize the files to reduce the cost of querying and also improve the speed of data loading into the Amazon Redshift cluster.

Which solution meets these requirements?

- A. Use AWS Glue to convert all the files from .csv to a single large Apache Parquet fil
- B. COPY the file into Amazon Redshift and query the file with Athena from Amazon S3.
- C. Use Amazon EMR to convert each .csv file to Apache Avr
- D. COPY the files into Amazon Redshift and query the file with Athena from Amazon S3.
- E. Use AWS Glue to convert the files from .csv to a single large Apache ORC fil
- F. COPY the file into Amazon Redshift and query the file with Athena from Amazon S3.
- G. Use AWS Glue to convert the files from .csv to Apache Parquet to create 20 Parquet file
- H. COPY the files into Amazon Redshift and query the files with Athena from Amazon S3.

**Answer:** D

#### NEW QUESTION 5

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

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

- A. Use Amazon QuickSight with Amazon Athena as the data 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 6

A financial company uses Apache Hive on Amazon EMR for ad-hoc queries. Users are complaining of sluggish performance. A data analyst notes the following:

- > Approximately 90% of queries are submitted 1 hour after the market opens.
- > Hadoop Distributed File System (HDFS) utilization never exceeds 10%.

Which solution would help address the performance issues?

- A. Create instance fleet configurations for core and task node
- B. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch CapacityRemainingGB 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 node
- 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 node
- 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 node
- 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 7

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 8

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 9

An airline has been collecting metrics on flight activities for analytics. A recently completed proof of concept demonstrates how the company provides insights to

data analysts to improve on-time departures. The proof of concept used objects in Amazon S3, which contained the metrics in .csv format, and used Amazon Athena for querying the data. As the amount of data increases, the data analyst wants to optimize the storage solution to improve query performance. Which options should the data analyst use to improve performance as the data lake grows? (Choose three.)

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

**Answer:** CDF

**Explanation:**

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

#### NEW QUESTION 10

A company is building a service to monitor fleets of vehicles. The company collects IoT data from a device in each vehicle and loads the data into Amazon Redshift in near-real time. Fleet owners upload .csv files containing vehicle reference data into Amazon S3 at different times throughout the day. A nightly process loads the vehicle reference data from Amazon S3 into Amazon Redshift. The company joins the IoT data from the device and the vehicle reference data to power reporting and dashboards. Fleet owners are frustrated by waiting a day for the dashboards to update.

Which solution would provide the SHORTEST delay between uploading reference data to Amazon S3 and the change showing up in the owners' dashboards?

- A. Use S3 event notifications to trigger an AWS Lambda function to copy the vehicle reference data into Amazon Redshift immediately when the reference data is uploaded to Amazon S3.
- B. Create and schedule an AWS Glue Spark job to run every 5 minute
- C. The job inserts reference data into Amazon Redshift.
- D. Send reference data to Amazon Kinesis Data Stream
- E. Configure the Kinesis data stream to directly load the reference data into Amazon Redshift in real time.
- F. Send the reference data to an Amazon Kinesis Data Firehose delivery stream
- G. Configure Kinesis with a buffer interval of 60 seconds and to directly load the data into Amazon Redshift.

**Answer:** A

#### NEW QUESTION 10

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

Which solution meets these requirements?

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

**Answer:** B

#### NEW QUESTION 12

A financial services company needs to aggregate daily stock trade data from the exchanges into a data store. The company requires that data be streamed directly into the data store, but also occasionally allows data to be modified using SQL. The solution should integrate complex, analytic queries running with minimal latency. The solution must provide a business intelligence dashboard that enables viewing of the top contributors to anomalies in stock prices.

Which solution meets the company's requirements?

- A. Use Amazon Kinesis Data Firehose to stream data to Amazon S3. Use Amazon Athena as a data source for Amazon QuickSight to create a business intelligence dashboard.
- B. Use Amazon Kinesis Data Streams to stream data to Amazon Redshift
- C. Use Amazon Redshift as a data source for Amazon QuickSight to create a business intelligence dashboard.
- D. Use Amazon Kinesis Data Firehose to stream data to Amazon Redshift
- E. Use Amazon Redshift as a data source for Amazon QuickSight to create a business intelligence dashboard.
- F. Use Amazon Kinesis Data Streams to stream data to Amazon S3. Use Amazon Athena as a data source for Amazon QuickSight to create a business intelligence dashboard.

**Answer:** C

#### NEW QUESTION 15

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 update
- 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 metadat
- E. Create an Amazon EMR cluster with consistent view enable
- F. Run emrfs sync before each analytics step to ensure data changes are update
- 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 datase
- I. Create an AWS Glue Data Catalog to manage the Hive metadata over the CTAS tabl
- 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 update
- L. Create an AWS Glue Data Catalog to manage the Hive metadata over the S3 Select tabl
- 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 16

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 19

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

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

**Answer:** D

#### NEW QUESTION 23

An analytics software as a service (SaaS) provider wants to offer its customers business intelligence (BI) reporting capabilities that are self-service. The provider is using Amazon QuickSight to build these reports. The data for the reports resides in a multi-tenant database, but each customer should only be able to access their own data.

The provider wants to give customers two user role options:

- Read-only users for individuals who only need to view dashboards
  - Power users for individuals who are allowed to create and share new dashboards with other users
- Which QuickSight feature allows the provider to meet these requirements'?

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

**Answer:** A

#### NEW QUESTION 26

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 28

A large retailer has successfully migrated to an Amazon S3 data lake architecture. The company's marketing team is using Amazon Redshift and Amazon QuickSight to analyze data, and derive and visualize insights. To ensure the marketing team has the most up-to-date actionable information, a data analyst implements nightly refreshes of Amazon Redshift using terabytes of updates from the previous day.

After the first nightly refresh, users report that half of the most popular dashboards that had been running correctly before the refresh are now running much

slower. Amazon CloudWatch does not show any alerts.  
What is the MOST likely cause for the performance degradation?

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

**Answer:** D

**Explanation:**

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

**NEW QUESTION 29**

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 32**

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 node
- 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 applications
- 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 instance
- 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-custom-bootstrap-actions/>  
[https://docs.aws.amazon.com/de\\_de/emr/latest/ManagementGuide/emr-plan-bootstrap.html](https://docs.aws.amazon.com/de_de/emr/latest/ManagementGuide/emr-plan-bootstrap.html)

**NEW QUESTION 35**

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 38**

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 43

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 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 45

A company uses the Amazon Kinesis SDK to write data to Kinesis Data Streams. Compliance requirements state that the data must be encrypted at rest using a key that can be rotated. The company wants to meet this encryption requirement with minimal coding effort. How can these requirements be met?

- A. Create a customer master key (CMK) in AWS KMS
- B. Assign the CMK an alias
- C. Use the AWS Encryption SDK, providing it with the key alias to encrypt and decrypt the data.
- D. Create a customer master key (CMK) in AWS KMS
- E. Assign the CMK an alias
- F. Enable server-side encryption on the Kinesis data stream using the CMK alias as the KMS master key.
- G. Create a customer master key (CMK) in AWS KMS
- H. Create an AWS Lambda function to encrypt and decrypt the data
- I. Set the KMS key ID in the function's environment variables.
- J. Enable server-side encryption on the Kinesis data stream using the default KMS key for Kinesis Data Streams.

**Answer: B**

#### NEW QUESTION 47

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 50

A marketing company is storing its campaign response data in Amazon S3. A consistent set of sources has generated the data for each campaign. The data is saved into Amazon S3 as .csv files. A business analyst will use Amazon Athena to analyze each campaign's data. The company needs the cost of ongoing data analysis with Athena to be minimized.

Which combination of actions should a data analytics specialist take to meet these requirements? (Choose two.)

- A. Convert the .csv files to Apache Parquet.
- B. Convert the .csv files to Apache Avro.
- C. Partition the data by campaign.

- D. Partition the data by source.
- E. Compress the .csv files.

**Answer:** AC

**Explanation:**

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

**NEW QUESTION 51**

A banking company is currently using an Amazon Redshift cluster with dense storage (DS) nodes to store sensitive data. An audit found that the cluster is unencrypted. Compliance requirements state that a database with sensitive data must be encrypted through a hardware security module (HSM) with automated key rotation.

Which combination of steps is required to achieve compliance? (Choose two.)

- A. Set up a trusted connection with HSM using a client and server certificate with automatic key rotation.
- B. Modify the cluster with an HSM encryption option and automatic key rotation.
- C. Create a new HSM-encrypted Amazon Redshift cluster and migrate the data to the new cluster.
- D. Enable HSM with key rotation through the AWS CLI.
- E. Enable Elliptic Curve Diffie-Hellman Ephemeral (ECDHE) encryption in the HSM.

**Answer:** BD

**NEW QUESTION 52**

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

Which solution improves the efficiency of the data processing jobs and is well architected?

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

**Answer:** D

**Explanation:**

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

**NEW QUESTION 57**

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

Which solution meets these requirements?

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

**Answer:** B

**NEW QUESTION 62**

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 forma
- E. Then add the function to the Kinesis Data Firehose transformation configuratio
- 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 sourc
- 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 67**

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 70**

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 71**

A data analyst is designing an Amazon QuickSight dashboard using centralized sales data that resides in Amazon Redshift. The dashboard must be restricted so that a salesperson in Sydney, Australia, can see only the Australia view and that a salesperson in New York can see only United States (US) data. What should the data analyst do to ensure the appropriate data security is in place?

- A. Place the data sources for Australia and the US into separate SPICE capacity pools.
- B. Set up an Amazon Redshift VPC security group for Australia and the US.
- C. Deploy QuickSight Enterprise edition to implement row-level security (RLS) to the sales table.
- D. Deploy QuickSight Enterprise edition and set up different VPC security groups for Australia and the US.

**Answer: D**

**NEW QUESTION 72**

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 CloudWatch.
- 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 CloudWatch.
- 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 CloudWatch.
- 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 CloudWatch.
- 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](https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/CWL_ES_Stream.html)

**NEW QUESTION 76**

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 77

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

"File could only be replicated to 0 nodes instead of 1"

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

Which solution will prevent the jobs from failing?

- A. Monitor the HDFSUtilization 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 80

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 82

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 86

A data analytics specialist is setting up workload management in manual mode for an Amazon Redshift environment. The data analytics specialist is defining query monitoring rules to manage system performance and user experience of an Amazon Redshift cluster.

Which elements must each query monitoring rule include?

- A. A unique rule name, a query runtime condition, and an AWS Lambda function to resubmit any failed queries in off hours.
- B. A queue name, a unique rule name, and a predicate-based stop condition.
- C. A unique rule name, one to three predicates, and an action.
- D. A workload name, a unique rule name, and a query runtime-based condition.

**Answer: C**

#### NEW QUESTION 87

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 91

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 95

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

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

**Answer: BD**

#### NEW QUESTION 99

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

#### NEW QUESTION 103

An education provider's learning management system (LMS) is hosted in a 100 TB data lake that is built on Amazon S3. The provider's LMS supports hundreds of schools. The provider wants to build an advanced analytics reporting platform using Amazon Redshift to handle complex queries with optimal performance. System users will query the most recent 4 months of data 95% of the time while 5% of the queries will leverage data from the previous 12 months. Which solution meets these requirements in the MOST cost-effective way?

- A. Store the most recent 4 months of data in the Amazon Redshift cluster.
- B. Use Amazon Redshift Spectrum to query data in the data lake.
- C. Use S3 lifecycle management rules to store data from the previous 12 months in Amazon S3 Glacier storage.
- D. Leverage DS2 nodes for the Amazon Redshift cluster.
- E. Migrate all data from Amazon S3 to Amazon Redshift.
- F. Decommission the data lake.
- G. Store the most recent 4 months of data in the Amazon Redshift cluster.
- H. Use Amazon Redshift Spectrum to query data in the data lake.
- I. Ensure the S3 Standard storage class is in use with objects in the data lake.
- J. Store the most recent 4 months of data in the Amazon Redshift cluster.
- K. Use Amazon Redshift federated queries to join cluster data with the data lake to reduce cost.
- L. Ensure the S3 Standard storage class is in use with objects in the data lake.

**Answer: C**

#### NEW QUESTION 104

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

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

**Answer:** A

#### NEW QUESTION 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 109

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

How should a data analytics specialist configure the AWS Glue job to optimize fault tolerance and improve data availability in the Amazon Redshift cluster?

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

**Answer:** B

#### NEW QUESTION 113

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 115

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 116

A company is hosting an enterprise reporting solution with Amazon Redshift. The application provides reporting capabilities to three main groups: an executive

group to access financial reports, a data analyst group to run long-running ad-hoc queries, and a data engineering group to run stored procedures and ETL processes. The executive team requires queries to run with optimal performance. The data engineering team expects queries to take minutes. Which Amazon Redshift feature meets the requirements for this task?

- A. Concurrency scaling
- B. Short query acceleration (SQA)
- C. Workload management (WLM)
- D. Materialized views

**Answer:** D

**Explanation:**

Materialized views:

**NEW QUESTION 118**

A company has an encrypted Amazon Redshift cluster. The company recently enabled Amazon Redshift audit logs and needs to ensure that the audit logs are also encrypted at rest. The logs are retained for 1 year. The auditor queries the logs once a month. What is the MOST cost-effective way to meet these requirements?

- A. Encrypt the Amazon S3 bucket where the logs are stored by using AWS Key Management Service (AWS KMS). Copy the data into the Amazon Redshift cluster from Amazon S3 on a daily basis
- B. Query the data as required.
- C. Disable encryption on the Amazon Redshift cluster, configure audit logging, and encrypt the Amazon Redshift cluster
- D. Use Amazon Redshift Spectrum to query the data as required.
- E. Enable default encryption on the Amazon S3 bucket where the logs are stored by using AES-256 encryption
- F. Copy the data into the Amazon Redshift cluster from Amazon S3 on a daily basis
- G. Query the data as required.
- H. Enable default encryption on the Amazon S3 bucket where the logs are stored by using AES-256 encryption
- I. Use Amazon Redshift Spectrum to query the data as required.

**Answer:** A

**NEW QUESTION 122**

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 127**

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

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

**Answer:** D

**NEW QUESTION 130**

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 133

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 136

A university intends to use Amazon Kinesis Data Firehose to collect JSON-formatted batches of water quality readings in Amazon S3. The readings are from 50 sensors scattered across a local lake. Students will query the stored data using Amazon Athena to observe changes in a captured metric over time, such as water temperature or acidity. Interest has grown in the study, prompting the university to reconsider how data will be stored.

Which data format and partitioning choices will MOST significantly reduce costs? (Choose two.)

- A. Store the data in Apache Avro format using Snappy compression.
- B. Partition the data by year, month, and day.
- C. Store the data in Apache ORC format using no compression.
- D. Store the data in Apache Parquet format using Snappy compression.
- E. Partition the data by sensor, year, month, and day.

**Answer:** CD

#### NEW QUESTION 138

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 139

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 142

A healthcare company uses AWS data and analytics tools to collect, ingest, and store electronic health record (EHR) data about its patients. The raw EHR data is stored in Amazon S3 in JSON format partitioned by hour, day, and year and is updated every hour. The company wants to maintain the data catalog and metadata in an AWS Glue Data Catalog to be able to access the data using Amazon Athena or Amazon Redshift Spectrum for analytics.

When defining tables in the Data Catalog, the company has the following requirements:

Choose the catalog table name and do not rely on the catalog table naming algorithm. Keep the table updated with new partitions loaded in the respective S3 bucket prefixes.

Which solution meets these requirements with minimal effort?

- A. Run an AWS Glue crawler that connects to one or more data stores, determines the data structures, and writes tables in the Data Catalog.
- B. Use the AWS Glue console to manually create a table in the Data Catalog and schedule an AWS Lambda function to update the table partitions hourly.
- C. Use the AWS Glue API CreateTable operation to create a table in the Data Catalog.
- D. Create an AWS Glue crawler and specify the table as the source.
- E. Create an Apache Hive catalog in Amazon EMR with the table schema definition in Amazon S3, and update the table partition with a scheduled job.
- F. Migrate the Hive catalog to the Data Catalog.

**Answer: C**

**Explanation:**

Updating Manually Created Data Catalog Tables Using Crawlers: To do this, when you define a crawler, instead of specifying one or more data stores as the source of a crawl, you specify one or more existing Data Catalog tables. The crawler then crawls the data stores specified by the catalog tables. In this case, no new tables are created; instead, your manually created tables are updated.

**NEW QUESTION 147**

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 152**

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 155**

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 158**

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 log data to Amazon OpenSearch Service (Amazon Elasticsearch Service) for search, log analytics, and application monitoring. Use Amazon QuickSight for the dashboards.

**Answer: C**

**NEW QUESTION 159**

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 160**

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 164**

A hospital uses an electronic health records (EHR) system to collect two types of data

- Patient information, which includes a patient's name and address
- Diagnostic tests conducted and the results of these tests

Patient information is expected to change periodically Existing diagnostic test data never changes and only new records are added

The hospital runs an Amazon Redshift cluster with four dc2.large nodes and wants to automate the ingestion of the patient information and diagnostic test data into respective Amazon Redshift tables for analysis The EHR system exports data as CSV files to an Amazon S3 bucket on a daily basis Two sets of CSV files are generated One set of files is for patient information with updates, deletes, and inserts The other set of files is for new diagnostic test data only

What is the MOST cost-effective solution to meet these requirements?

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

**Answer: B**

**NEW QUESTION 169**

.....

## 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