



# Amazon-Web-Services

## Exam Questions DAS-C01

AWS Certified Data Analytics - Specialty

## About ExamBible

### *Your Partner of IT Exam*

## Found in 1998

ExamBible is a company specialized on providing high quality IT exam practice study materials, especially Cisco CCNA, CCDA, CCNP, CCIE, Checkpoint CCSE, CompTIA A+, Network+ certification practice exams and so on. We guarantee that the candidates will not only pass any IT exam at the first attempt but also get profound understanding about the certificates they have got. There are so many alike companies in this industry, however, ExamBible has its unique advantages that other companies could not achieve.

## Our Advances

### \* 99.9% Uptime

All examinations will be up to date.

### \* 24/7 Quality Support

We will provide service round the clock.

### \* 100% Pass Rate

Our guarantee that you will pass the exam.

### \* Unique Gurantee

If you do not pass the exam at the first time, we will not only arrange FULL REFUND for you, but also provide you another exam of your claim, ABSOLUTELY FREE!

#### NEW QUESTION 1

A market data company aggregates external data sources to create a detailed view of product consumption in different countries. The company wants to sell this data to external parties through a subscription. To achieve this goal, the company needs to make its data securely available to external parties who are also AWS users.

What should the company do to meet these requirements with the LEAST operational overhead?

- A. Store the data in Amazon S3. Share the data by using presigned URLs for security.
- B. Store the data in Amazon S3. Share the data by using S3 bucket ACLs.
- C. Upload the data to AWS Data Exchange for storage.
- D. Share the data by using presigned URLs for security.
- E. Upload the data to AWS Data Exchange for storage.
- F. Share the data by using the AWS Data Exchange sharing wizard.

**Answer:** A

#### NEW QUESTION 2

A company hosts an Apache Flink application on premises. The application processes data from several Apache Kafka clusters. The data originates from a variety of sources, such as web applications, mobile apps, and operational databases. The company has migrated some of these sources to AWS and now wants to migrate the Flink application. The company must ensure that data that resides in databases within the VPC does not traverse the internet. The application must be able to process all the data that comes from the company's AWS solution, on-premises resources, and the public internet.

Which solution will meet these requirements with the LEAST operational overhead?

- A. Implement Flink on Amazon EC2 within the company's VPC. Create Amazon Managed Streaming for Apache Kafka (Amazon MSK) clusters in the VPC to collect data that comes from applications and databases within the VPC. Use Amazon Kinesis Data Streams to collect data that comes from the public internet. Configure Flink to have sources from Kinesis Data Streams, Amazon MSK, and any on-premises Kafka clusters by using AWS Client VPN or AWS Direct Connect.
- B. Implement Flink on Amazon EC2 within the company's VPC. Use Amazon Kinesis Data Streams to collect data that comes from applications and databases within the VPC and the public internet. Configure Flink to have sources from Kinesis Data Streams and any on-premises Kafka clusters by using AWS Client VPN or AWS Direct Connect.
- C. Create an Amazon Kinesis Data Analytics application by uploading the compiled Flink JAR file. Use Amazon Kinesis Data Streams to collect data that comes from applications and databases within the VPC and the public internet. Configure the Kinesis Data Analytics application to have sources from Kinesis Data Streams and any on-premises Kafka clusters by using AWS Client VPN or AWS Direct Connect.
- D. Create an Amazon Kinesis Data Analytics application by uploading the compiled Flink JAR file. Create Amazon Managed Streaming for Apache Kafka (Amazon MSK) clusters in the company's VPC to collect data that comes from applications and databases within the VPC. Use Amazon Kinesis Data Streams to collect data that comes from the public internet. Configure the Kinesis Data Analytics application to have sources from Kinesis Data Streams, Amazon MSK, and any on-premises Kafka clusters by using AWS Client VPN or AWS Direct Connect.
- E. Amazon MSK and any on-premises Kafka clusters by using AWS Client VPN or AWS Direct Connect.

**Answer:** D

#### NEW QUESTION 3

A company has several Amazon EC2 instances sitting behind an Application Load Balancer (ALB). The company wants its IT Infrastructure team to analyze the IP addresses coming into the company's ALB. The ALB is configured to store access logs in Amazon S3. The access logs create about 1 TB of data each day, and access to the data will be infrequent. The company needs a solution that is scalable, cost-effective, and has minimal maintenance requirements.

Which solution meets these requirements?

- A. Copy the data into Amazon Redshift and query the data.
- B. Use Amazon EMR and Apache Hive to query the S3 data.
- C. Use Amazon Athena to query the S3 data.
- D. Use Amazon Redshift Spectrum to query the S3 data.

**Answer:** D

#### NEW QUESTION 4

A 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 5

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

Which steps will satisfy the security requirements?

- A. For the EMR cluster Amazon EC2 instances, create a service role that grants no access to Amazon S3. Create three additional IAM roles, each granting access to each team's specific bucket.

- visit - <https://www.exambible.com>

**Answer:** B

#### NEW QUESTION 9

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

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

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

**Answer:** C

#### Explanation:

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

#### NEW QUESTION 10

A bank wants to migrate a Teradata data warehouse to the AWS Cloud. The bank needs a solution for reading large amounts of data and requires the highest possible performance. The solution also must maintain the separation of storage and compute.

Which solution meets these requirements?

- A. Use Amazon Athena to query the data in Amazon S3.
- B. Use Amazon Redshift with dense compute nodes to query the data in Amazon Redshift managed storage.
- C. Use Amazon Redshift with RA3 nodes to query the data in Amazon Redshift managed storage.
- D. Use PrestoDB on Amazon EMR to query the data in Amazon S3.

**Answer:** C

#### NEW QUESTION 10

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

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

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

**Answer:** D

#### NEW QUESTION 15

A company currently uses Amazon Athena to query its global datasets. The regional data is stored in Amazon S3 in the us-east-1 and us-west-2 Regions. The data is not encrypted. To simplify the query process and manage it centrally, the company wants to use Athena in us-west-2 to query data from Amazon S3 in both Regions. The solution should be as low-cost as possible.

What should the company do to achieve this goal?

- A. Use AWS DMS to migrate the AWS Glue Data Catalog from us-east-1 to us-west-2. Run Athena queries in us-west-2.
- B. Run the AWS Glue crawler in us-west-2 to catalog datasets in all Region.
- C. Once the data is crawled, run Athena queries in us-west-2.
- D. Enable cross-Region replication for the S3 buckets in us-east-1 to replicate data in us-west-2. Once the data is replicated in us-west-2, run the AWS Glue crawler there to update the AWS Glue Data Catalog in us-west-2 and run Athena queries.
- E. Update AWS Glue resource policies to provide us-east-1 AWS Glue Data Catalog access to us-west-2. Once the catalog in us-west-2 has access to the catalog in us-east-1, run Athena queries in us-west-2.

**Answer:** B

#### NEW QUESTION 18

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 19



A media company is using Amazon QuickSight dashboards to visualize its national sales data. The dashboard is using a dataset with these fields: ID, date, time\_zone, city, state, country, longitude, latitude, sales\_volume, and number\_of\_items.

To modify ongoing campaigns, the company wants an interactive and intuitive visualization of which states across the country recorded a significantly lower sales volume compared to the national average.

Which addition to the company's QuickSight dashboard will meet this requirement?

- A. A geospatial color-coded chart of sales volume data across the country.
- B. A pivot table of sales volume data summed up at the state level.
- C. A drill-down layer for state-level sales volume data.
- D. A drill through to other dashboards containing state-level sales volume data.

**Answer: B**

#### NEW QUESTION 21

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 24

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 29

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 queries. 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 data. Process incoming data from Kinesis Data Streams. Use an Amazon Kinesis Data Firehose delivery stream to write results to Amazon Redshift.

**Answer: D**

#### NEW QUESTION 31

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 table.
- B. Add SQL commands to replace the existing rows in the main table as part of actions in the DynamicFrameWriter class.
- C. Load the previously inserted data into a MySQL database in the AWS Glue job.
- 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 34**

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 37**

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 38**

A company wants to enrich application logs in near-real-time and use the enriched dataset for further analysis. The application is running on Amazon EC2 instances across multiple Availability Zones and storing its logs using Amazon CloudWatch Logs. The enrichment source is stored in an Amazon DynamoDB table. Which solution meets the requirements for the event collection and enrichment?

- A. Use a CloudWatch Logs subscription to send the data to Amazon Kinesis Data Firehos
- B. Use AWS Lambda to transform the data in the Kinesis Data Firehose delivery stream and enrich it with the data inthe DynamoDB tabl
- C. Configure Amazon S3 as the Kinesis Data Firehose delivery destination.
- D. Export the raw logs to Amazon S3 on an hourly basis using the AWS CL
- E. Use AWS Glue crawlers to catalog the log
- F. Set up an AWS Glue connection for the DynamoDB table and set up an AWS Glue ETL job to enrich the dat
- G. Store the enriched data in Amazon S3.
- H. Configure the application to write the logs locally and use Amazon Kinesis Agent to send the data to Amazon Kinesis Data Stream
- I. Configure a Kinesis Data Analytics SQL application with the Kinesis data stream as the sourc
- J. Join the SQL application input stream with DynamoDB records, and then store the enriched output stream in Amazon S3 using Amazon Kinesis Data Firehose.
- K. Export the raw logs to Amazon S3 on an hourly basis using the AWS CL
- L. Use Apache Spark SQL on Amazon EMR to read the logs from Amazon S3 and enrich the records with the data from DynamoD
- M. Store the enriched data in Amazon S3.

**Answer:** A

**Explanation:**

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

**NEW QUESTION 42**

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

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

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

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

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

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

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

**Answer:** C

**Explanation:**

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

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

**NEW QUESTION 45**

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 49**

A retail company is building its data warehouse solution using Amazon Redshift. As a part of that effort, the company is loading hundreds of files into the fact table created in its Amazon Redshift cluster. The company wants the solution to achieve the highest throughput and optimally use cluster resources when loading data into the company's fact table.

How should the company meet these requirements?

- A. Use multiple COPY commands to load the data into the Amazon Redshift cluster.
- B. Use S3DistCp to load multiple files into the Hadoop Distributed File System (HDFS) and use an HDFS connector to ingest the data into the Amazon Redshift cluster.
- C. Use LOAD commands equal to the number of Amazon Redshift cluster nodes and load the data in parallel into each node.
- D. Use a single COPY command to load the data into the Amazon Redshift cluster.

**Answer:** D

**Explanation:**

[https://docs.aws.amazon.com/redshift/latest/dg/c\\_best-practices-single-copy-command.html](https://docs.aws.amazon.com/redshift/latest/dg/c_best-practices-single-copy-command.html)

**NEW QUESTION 50**

A company stores Apache Parquet-formatted files in Amazon S3. The company uses an AWS Glue Data Catalog to store the table metadata and Amazon Athena to query and analyze the data. The tables have a large number of partitions. The queries are only run on small subsets of data in the table. A data analyst adds new time partitions into the table as new data arrives. The data analyst has been asked to reduce the query runtime.

Which solution will provide the MOST reduction in the query runtime?

- A. Convert the Parquet files to the csv file format. Then attempt to query the data again.
- B. Convert the Parquet files to the Apache ORC file format.
- C. Then attempt to query the data again.
- D. Use partition projection to speed up the processing of the partitioned table.
- E. Add more partitions to be used over the table.
- F. Then filter over two partitions and put all columns in the WHERE clause.

**Answer:** C

**NEW QUESTION 55**

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



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

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

**Answer:** D

**Explanation:**

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

#### NEW QUESTION 59

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 60

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 62

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 67

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 72

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 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 81

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 82

A company hosts an on-premises PostgreSQL database that contains historical data. An internal legacy application uses the database for read-only activities. The company's business team wants to move the data to a data lake in Amazon S3 as soon as possible and enrich the data for analytics.

The company has set up an AWS Direct Connect connection between its VPC and its on-premises network. A data analytics specialist must design a solution that achieves the business team's goals with the least operational overhead.

Which solution meets these requirements?

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

**Answer:** B

#### NEW QUESTION 86

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 90**

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 93**

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 files.
- 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 94**

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 99**

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 102**

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 105

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 107

A company analyzes historical data and needs to query data that is stored in Amazon S3. New data is generated daily as .csv files that are stored in Amazon S3. The company's analysts are using Amazon Athena to perform SQL queries against a recent subset of the overall data. The amount of data that is ingested into Amazon S3 has increased substantially over time, and the query latency also has increased.

Which solutions could the company implement to improve query performance? (Choose two.)

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

**Answer:** BC

#### NEW QUESTION 111

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 114

A company uses Amazon Redshift as its data warehouse A new table includes some columns that contain sensitive data and some columns that contain non-sensitive data The data in the table eventually will be referenced by several existing queries that run many times each day

A data analytics specialist must ensure that only members of the company's auditing team can read the columns that contain sensitive data All other users must have read-only access to the columns that contain non-sensitive data

Which solution will meet these requirements with the LEAST operational overhead?

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



**Answer:** B

**Explanation:**

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

**NEW QUESTION 118**

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 120**

A company is planning to create a data lake in Amazon S3. The company wants to create tiered storage based on access patterns and cost objectives. The solution must include support for JDBC connections from legacy clients, metadata management that allows federation for access control, and batch-based ETL using PySpark and Scala. Operational management should be limited.

Which combination of components can meet these requirements? (Choose three.)

- A. AWS Glue Data Catalog for metadata management
- B. Amazon EMR with Apache Spark for ETL
- C. AWS Glue for Scala-based ETL
- D. Amazon EMR with Apache Hive for JDBC clients
- E. Amazon Athena for querying data in Amazon S3 using JDBC drivers
- F. Amazon EMR with Apache Hive, using an Amazon RDS with MySQL-compatible backed metastore

**Answer:** BEF

**NEW QUESTION 125**

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 127**

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 metadata
- 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 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 update
- 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 130**

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 135

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 136

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 140

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 143

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 minutes
- 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 144

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 148

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 file
- 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 Avro
- 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 file
- 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 files
- H. COPY the files into Amazon Redshift and query the files with Athena from Amazon S3.

**Answer:** D

#### NEW QUESTION 150

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

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

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

**Answer:** B

#### NEW QUESTION 154

.....

## Relate Links

**100% Pass Your DAS-C01 Exam with ExamBible Prep Materials**

<https://www.exambible.com/DAS-C01-exam/>

## Contact us

We are proud of our high-quality customer service, which serves you around the clock 24/7.

Viste - <https://www.exambible.com/>