

AWS-Certified-Data-Analytics-Specialty Dumps

AWS Certified Data Analytics - Specialty

<https://www.certleader.com/AWS-Certified-Data-Analytics-Specialty-dumps.html>



NEW QUESTION 1

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 2

A retail company wants to use Amazon QuickSight to generate dashboards for web and in-store sales. A group of 50 business intelligence professionals will develop and use the dashboards. Once ready, the dashboards will be shared with a group of 1,000 users.

The sales data comes from different stores and is uploaded to Amazon S3 every 24 hours. The data is partitioned by year and month, and is stored in Apache Parquet format. The company is using the AWS Glue Data Catalog as its main data catalog and Amazon Athena for querying. The total size of the uncompressed data that the dashboards query from at any point is 200 GB.

Which configuration will provide the MOST cost-effective solution that meets these requirements?

- A. Load the data into an Amazon Redshift cluster by using the COPY command.
- B. Configure 50 author users and 1,000 reader user.
- C. Use QuickSight Enterprise edition.
- D. Configure an Amazon Redshift data source with a direct query option.
- E. Use QuickSight Standard edition.
- F. Configure 50 author users and 1,000 reader user.
- G. Configure an Athena data source with a direct query option.
- H. Use QuickSight Enterprise edition.
- I. Configure 50 author users and 1,000 reader user.
- J. Configure an Athena data source and import the data into SPICE.
- K. Automatically refresh every 24 hours.
- L. Use QuickSight Enterprise edition.
- M. Configure 1 administrator and 1,000 reader user.
- N. Configure an S3 data source and import the data into SPICE.
- O. Automatically refresh every 24 hours.

Answer: C

NEW QUESTION 3

A large telecommunications company is planning to set up a data catalog and metadata management for multiple data sources running on AWS. The catalog will be used to maintain the metadata of all the objects stored in the data stores. The data stores are composed of structured sources like Amazon RDS and Amazon Redshift, and semistructured sources like JSON and XML files stored in Amazon S3. The catalog must be updated on a regular basis, be able to detect the changes to object metadata, and require the least possible administration.

Which solution meets these requirements?

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

Answer: D

NEW QUESTION 4

An airline has been collecting metrics on flight activities for analytics. A recently completed proof of concept demonstrates how the company provides insights to data analysts to improve on-time departures. The proof of concept used objects in Amazon S3, which contained the metrics in .csv format, and used Amazon Athena for querying the data. As the amount of data increases, the data analyst wants to optimize the storage solution to improve query performance.

Which options should the data analyst use to improve performance as the data lake grows? (Choose three.)

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

Answer: CDF

Explanation:

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

NEW QUESTION 5

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 6

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

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

Answer: D

NEW QUESTION 7

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 8

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 9

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 10

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

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

Answer: BD

NEW QUESTION 10

A company has collected more than 100 TB of log files in the last 24 months. The files are stored as raw text in a dedicated Amazon S3 bucket. Each object has a key of the form year-month-day_log_HH:mm:ss.txt where HH:mm:ss represents the time the log file was initially created. A table was created in Amazon Athena that points to the S3 bucket. One-time queries are run against a subset of columns in the table several times an hour.

A data analyst must make changes to reduce the cost of running these queries. Management wants a solution with minimal maintenance overhead.

Which combination of steps should the data analyst take to meet these requirements? (Choose three.)

- A. Convert the log files to Apache Avro format.
- B. Add a key prefix of the form date=year-month-day/ to the S3 objects to partition the data.
- C. Convert the log files to Apache Parquet format.
- D. Add a key prefix of the form year-month-day/ to the S3 objects to partition the data.
- E. Drop and recreate the table with the PARTITIONED BY clause
- F. Run the ALTER TABLE ADD PARTITION statement.
- G. Drop and recreate the table with the PARTITIONED BY clause
- H. Run the MSCK REPAIR TABLE statement.

Answer: BCF

NEW QUESTION 15

A company uses Amazon Redshift as its data warehouse. A new table has columns that contain sensitive data. The data in the table will eventually be referenced by several existing queries that run many times a day.

A data analyst needs to load 100 billion rows of data into the new table. Before doing so, the data analyst must ensure that only members of the auditing group can read the columns containing sensitive data.

How can the data analyst meet these requirements with the lowest maintenance overhead?

- A. Load all the data into the new table and grant the auditing group permission to read from the table.
- B. Load all the data except for the columns containing sensitive data into a second table.
- C. Grant the appropriate users read-only permissions to the second table.
- D. Load all the data into the new table and grant the auditing group permission to read from the table.
- E. Use the GRANT SQL command to allow read-only access to a subset of columns to the appropriate users.
- F. Load all the data into the new table and grant all users read-only permissions to non-sensitive columns. Attach an IAM policy to the auditing group with explicit ALLOW access to the sensitive data columns.
- G. Load all the data into the new table and grant the auditing group permission to read from the table. Create a view of the new table that contains all the columns, except for those considered sensitive, and grant the appropriate users read-only permissions to the table.

Answer: B

Explanation:

<https://aws.amazon.com/blogs/big-data/achieve-finer-grained-data-security-with-column-level-access-control-in>

NEW QUESTION 18

A retail company stores order invoices in an Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster. Indices on the cluster are created monthly. Once a new month begins, no new writes are made to any of the indices from the previous months. The company has been expanding the storage on the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster to avoid running out of space, but the company wants to reduce costs. Most searches on the cluster are on the most recent 3 months of data, while the audit team requires infrequent access to older data to generate periodic reports. The most recent 3 months of data must be quickly available for queries, but the audit team can tolerate slower queries if the solution saves on cluster costs.

Which of the following is the MOST operationally efficient solution to meet these requirements?

- A. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to store the indices in Amazon S3 Glacier. When the audit team requires the archived data, restore the archived indices back to the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster.
- B. Archive indices that are older than 3 months by taking manual snapshots and storing the snapshots in Amazon S3. When the audit team requires the archived data, restore the archived indices back to the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster.

- C. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to migrate the indices to Amazon OpenSearch Service (Amazon Elasticsearch Service) UltraWarm storage
- D. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to migrate the indices to Amazon OpenSearch Service (Amazon Elasticsearch Service) UltraWarm storage When the audit team requires the older data: migrate the indices in UltraWarm storage back to hot storage

Answer: D

NEW QUESTION 22

An online retailer needs to deploy a product sales reporting solution. The source data is exported from an external online transaction processing (OLTP) system for reporting. Roll-up data is calculated each day for the previous day's activities. The reporting system has the following requirements:

Have the daily roll-up data readily available for 1 year.

After 1 year, archive the daily roll-up data for occasional but immediate access.

The source data exports stored in the reporting system must be retained for 5 years. Query access will be needed only for re-evaluation, which may occur within the first 90 days.

Which combination of actions will meet these requirements while keeping storage costs to a minimum? (Choose two.)

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

Answer: AD

NEW QUESTION 25

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 28

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 31

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 32

A data engineer is using AWS Glue ETL jobs to process data at frequent intervals. The processed data is then copied into Amazon S3. The ETL jobs run every 15 minutes. The AWS Glue Data Catalog partitions need to be updated automatically after the completion of each job. Which solution will meet these requirements MOST cost-effectively?

- A. Use the AWS Glue Data Catalog to manage the data catalog. Define an AWS Glue workflow for the ETL process. Define a trigger within the workflow that can start the crawler when an ETL job run is complete.
- B. Use the AWS Glue Data Catalog to manage the data catalog. Use AWS Glue Studio to manage ETL job.
- C. Use the AWS Glue Studio feature that supports updates to the AWS Glue Data Catalog during job runs.
- D. Use an Apache Hive metastore to manage the data catalog. Update the AWS Glue ETL code to include the `enableUpdateCatalog` and `partitionKeys` arguments.
- E. Use the AWS Glue Data Catalog to manage the data catalog. Update the AWS Glue ETL code to include the `enableUpdateCatalog` and `partitionKeys` arguments.

Answer: A

NEW QUESTION 36

A hospital uses wearable medical sensor devices to collect data from patients. The hospital is architecting a near-real-time solution that can ingest the data securely at scale. The solution should also be able to remove the patient's protected health information (PHI) from the streaming data and store the data in durable storage. Which solution meets these requirements with the least operational overhead?

- A. Ingest the data using Amazon Kinesis Data Streams, which invokes an AWS Lambda function using Kinesis Client Library (KCL) to remove all PHI.
- B. Write the data in Amazon S3.
- C. Ingest the data using Amazon Kinesis Data Firehose to write the data to Amazon S3. Have Amazon S3 trigger an AWS Lambda function that parses the sensor data to remove all PHI in Amazon S3.
- D. Ingest the data using Amazon Kinesis Data Streams to write the data to Amazon S3. Have the data stream launch an AWS Lambda function that parses the sensor data and removes all PHI in Amazon S3.
- E. Ingest the data using Amazon Kinesis Data Firehose to write the data to Amazon S3. Implement a transformation AWS Lambda function that parses the sensor data to remove all PHI.

Answer: D

Explanation:

<https://aws.amazon.com/blogs/big-data/persist-streaming-data-to-amazon-s3-using-amazon-kinesis-firehose-and>

NEW QUESTION 41

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 43

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

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

Answer: B

NEW QUESTION 46

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

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

Answer: A

Explanation:

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

NEW QUESTION 50

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 55

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 the job 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 57

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 60

A manufacturing company has been collecting IoT sensor data from devices on its factory floor for a year and is storing the data in Amazon Redshift for daily analysis. A data analyst has determined that, at an expected ingestion rate of about 2 TB per day, the cluster will be undersized in less than 4 months. A long-term solution is needed. The data analyst has indicated that most queries only reference the most recent 13 months of data, yet there are also quarterly reports that need to query all the data generated from the past 7 years. The chief technology officer (CTO) is concerned about the costs, administrative effort, and performance of a long-term solution.

Which solution should the data analyst use to meet these requirements?

- A. Create a daily job in AWS Glue to UNLOAD records older than 13 months to Amazon S3 and delete those records from Amazon Redshift

- B. Create an external table in Amazon Redshift to point to the S3 location.
- C. Use Amazon Redshift Spectrum to join to data that is older than 13 months.
- D. Take a snapshot of the Amazon Redshift cluster.
- E. Restore the cluster to a new cluster using dense storage nodes with additional storage capacity.
- F. Execute a CREATE TABLE AS SELECT (CTAS) statement to move records that are older than 13 months to quarterly partitioned data in Amazon Redshift Spectrum backed by Amazon S3.
- G. Unload all the tables in Amazon Redshift to an Amazon S3 bucket using S3 Intelligent-Tiering.
- H. Use AWS Glue to crawl the S3 bucket location to create external tables in an AWS Glue Data Catalog. Create an Amazon EMR cluster using Auto Scaling for any daily analytics needs, and use Amazon Athena for the quarterly reports, with both using the same AWS Glue Data Catalog.

Answer: A

NEW QUESTION 65

An operations team notices that a few AWS Glue jobs for a given ETL application are failing. The AWS Glue jobs read a large number of small JSON files from an Amazon S3 bucket and write the data to a different S3 bucket in Apache Parquet format with no major transformations. Upon initial investigation, a data engineer notices the following error message in the History tab on the AWS Glue console: “Command Failed with Exit Code 1.”

Upon further investigation, the data engineer notices that the driver memory profile of the failed jobs crosses the safe threshold of 50% usage quickly and reaches 90–95% soon after. The average memory usage across all executors continues to be less than 4%.

The data engineer also notices the following error while examining the related Amazon CloudWatch Logs. What should the data engineer do to solve the failure in the MOST cost-effective way?

- A. Change the worker type from Standard to G.2X.
- B. Modify the AWS Glue ETL code to use the ‘groupFiles’: ‘inPartition’ feature.
- C. Increase the fetch size setting by using AWS Glue dynamics frame.
- D. Modify maximum capacity to increase the total maximum data processing units (DPUs) used.

Answer: B

Explanation:

<https://docs.aws.amazon.com/glue/latest/dg/monitor-profile-debug-oom-abnormalities.html#monitor-debug-oom>

NEW QUESTION 68

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 71

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

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

Which storage service will provide the MOST performant solution?

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

Answer: B

NEW QUESTION 75

An ecommerce company is migrating its business intelligence environment from on premises to the AWS Cloud. The company will use Amazon Redshift in a public subnet and Amazon QuickSight. The tables already are loaded into Amazon Redshift and can be accessed by a SQL tool.

The company starts QuickSight for the first time. During the creation of the data source, a data analytics specialist enters all the information and tries to validate the connection. An error with the following message occurs: “Creating a connection to your data source timed out.”

How should the data analytics specialist resolve this error?

- A. Grant the SELECT permission on Amazon Redshift tables.
- B. Add the QuickSight IP address range into the Amazon Redshift security group.
- C. Create an IAM role for QuickSight to access Amazon Redshift.
- D. Use a QuickSight admin user for creating the dataset.

Answer: A

Explanation:

Connection to the database times out

Your client connection to the database appears to hang or time out when running long queries, such as a COPY command. In this case, you might observe that the Amazon Redshift console displays that the query has completed, but the client tool itself still appears to be running the query. The results of the query might be missing or incomplete depending on when the connection stopped.

NEW QUESTION 78

A company has a data lake on AWS that ingests sources of data from multiple business units and uses Amazon Athena for queries. The storage layer is Amazon S3 using the AWS Glue Data Catalog. The company wants to make the data available to its data scientists and business analysts. However, the company first needs to manage data access for Athena based on user roles and responsibilities.

What should the company do to apply these access controls with the LEAST operational overhead?

- A. Define security policy-based rules for the users and applications by role in AWS Lake Formation.
- B. Define security policy-based rules for the users and applications by role in AWS Identity and Access Management (IAM).
- C. Define security policy-based rules for the tables and columns by role in AWS Glue.
- D. Define security policy-based rules for the tables and columns by role in AWS Identity and Access Management (IAM).

Answer: D

NEW QUESTION 81

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

Which steps will satisfy the security requirements?

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

Answer: C

NEW QUESTION 86

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 88

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 92

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 Hud
- 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 96

.....

Thank You for Trying Our Product

* **100% Pass or Money Back**

All our products come with a 90-day Money Back Guarantee.

* **One year free update**

You can enjoy free update one year. 24x7 online support.

* **Trusted by Millions**

We currently serve more than 30,000,000 customers.

* **Shop Securely**

All transactions are protected by VeriSign!

100% Pass Your AWS-Certified-Data-Analytics-Specialty Exam with Our Prep Materials Via below:

<https://www.certleader.com/AWS-Certified-Data-Analytics-Specialty-dumps.html>