

Amazon

Exam Questions AWS-Certified-Data-Analytics-Specialty

AWS Certified Data Analytics - Specialty



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

Which solution meets these requirements?

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

Answer: C

NEW QUESTION 3

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

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

How can this issue be resolved?

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

Answer: D

NEW QUESTION 4

A manufacturing company uses Amazon S3 to store its data. The company wants to use AWS Lake Formation to provide granular-level security on those data assets. The data is in Apache Parquet format. The company has set a deadline for a consultant to build a data lake.

How should the consultant create the MOST cost-effective solution that meets these requirements?

- A. Run Lake Formation blueprints to move the data to Lake Formation.
- B. Once Lake Formation has the data, apply permissions on Lake Formation.
- C. To create the data catalog, run an AWS Glue crawler on the existing Parquet data.
- D. Register the Amazon S3 path and then apply permissions through Lake Formation to provide granular-level security.
- E. Install Apache Ranger on an Amazon EC2 instance and integrate with Amazon EMR.
- F. Using Ranger policies, create role-based access control for the existing data assets in Amazon S3.
- G. Create multiple IAM roles for different users and groups.
- H. Assign IAM roles to different data assets in Amazon S3 to create table-based and column-based access controls.

Answer: A

Explanation:

<https://aws.amazon.com/blogs/big-data/building-securing-and-managing-data-lakes-with-aws-lake-formation/>

NEW QUESTION 5

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

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

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

Which solution meets these requirements?

- A. Configure an Amazon Kinesis Data Firehose delivery stream for each application.
- B. Write AWS Lambda functions to read log data objects from the stream for each application.
- C. Have the function perform reformatting and .csv conversion.
- D. Enable compression on all the delivery streams.

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

Answer: A

NEW QUESTION 6

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

Answer: D

NEW QUESTION 7

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 8

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 forpredicate

Answer: CDF

Explanation:

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

NEW QUESTION 9

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 Redshif
- 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 Redshif
- 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 10

A company wants to run analytics on its Elastic Load Balancing logs stored in Amazon S3. A data analyst needs to be able to query all data from a desired year, month, or day. The data analyst should also be able to query a subset of the columns. The company requires minimal operational overhead and the most cost-effective solution.

Which approach meets these requirements for optimizing and querying the log data?

- A. Use an AWS Glue job nightly to transform new log files into .csv format and partition by year, month, and da
- B. Use AWS Glue crawlers to detect new partition
- C. Use Amazon Athena to query data.
- D. Launch a long-running Amazon EMR cluster that continuously transforms new log files from Amazon S3 into its Hadoop Distributed File System (HDFS) storage and partitions by year, month, and da
- E. Use Apache Presto to query the optimized format.
- F. Launch a transient Amazon EMR cluster nightly to transform new log files into Apache ORC format and partition by year, month, and da
- G. Use Amazon Redshift Spectrum to query the data.
- H. Use an AWS Glue job nightly to transform new log files into Apache Parquet format and partition by year, month, and da
- I. Use AWS Glue crawlers to detect new partition
- J. Use Amazon Athena to querydata.

Answer: C

NEW QUESTION 10

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 14

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 HDF
- D. Enable EMRFS consistent view. Create an EMR HBase cluster with multiple master node
- E. Point the HBase root directory to an Amazon S3 bucket.
- F. Store the data on an EMR File System (EMRFS) instead of HDFS and enable EMRFS consistent view. Run two separate EMR clusters in two different Availability Zone
- G. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.
- H. Store the data on an EMR File System (EMRFS) instead of HDFS and enable EMRFS consistent view. Create a primary EMR HBase cluster with multiple master node
- I. Create a secondary EMR HBase read- replica cluster in a separate Availability Zon
- J. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.

Answer: D

NEW QUESTION 17

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

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

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

Answer: AE

Explanation:

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

NEW QUESTION 21

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 23

A bank operates in a regulated environment. The compliance requirements for the country in which the bank operates say that customer data for each state should only be accessible by the bank's employees located in the same state. Bank employees in one state should NOT be able to access data for customers who have provided a home address in a different state.

The bank's marketing team has hired a data analyst to gather insights from customer data for a new campaign being launched in certain states. Currently, data linking each customer account to its home state is stored in a tabular .csv file within a single Amazon S3 folder in a private S3 bucket. The total size of the S3 folder is 2 GB uncompressed. Due to the country's compliance requirements, the marketing team is not able to access this folder.

The data analyst is responsible for ensuring that the marketing team gets one-time access to customer data for their campaign analytics project, while being subject to all the compliance requirements and controls.

Which solution should the data analyst implement to meet the desired requirements with the LEAST amount of setup effort?

- A. Re-arrange data in Amazon S3 to store customer data about each state in a different S3 folder within the same bucket
- B. Set up S3 bucket policies to provide marketing employees with appropriate data access under compliance control
- C. Delete the bucket policies after the project.
- D. Load tabular data from Amazon S3 to an Amazon EMR cluster using s3DistC
- E. Implement a custom Hadoop-based row-level security solution on the Hadoop Distributed File System (HDFS) to provide marketing employees with appropriate data access under compliance control
- F. Terminate the EMR cluster after the project.
- G. Load tabular data from Amazon S3 to Amazon Redshift with the COPY command
- H. Use the built-in row-level security feature in Amazon Redshift to provide marketing employees with appropriate data access under compliance control
- I. Delete the Amazon Redshift tables after the project.
- J. Load tabular data from Amazon S3 to Amazon QuickSight Enterprise edition by directly importing it as a data source
- K. Use the built-in row-level security feature in Amazon QuickSight to provide marketing employees with appropriate data access under compliance control
- L. Delete Amazon QuickSight data sources after the project is complete.

Answer: C

NEW QUESTION 27

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

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

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

Answer: D

NEW QUESTION 28

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 operatio
- 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 33

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

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

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

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

Answer: ADE

NEW QUESTION 34

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

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

Answer: A

Explanation:

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

NEW QUESTION 38

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 clas
- 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 clas
- 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 clas
- 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 clas
- 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 clas
- J. Apply a lifecycle configuration that changes the storage class to Amazon S3 Glacier 1 year after data creation.

Answer: AD

NEW QUESTION 39

A US-based sneaker retail company launched its global website. All the transaction data is stored in Amazon RDS and curated historic transaction data is stored in Amazon Redshift in the us-east-1 Region. The business intelligence (BI) team wants to enhance the user experience by providing a dashboard for sneaker trends. The BI team decides to use Amazon QuickSight to render the website dashboards. During development, a team in Japan provisioned Amazon QuickSight in ap-northeast-1. The team is having difficulty connecting Amazon QuickSight from ap-northeast-1 to Amazon Redshift in us-east-1. Which solution will solve this issue and meet the requirements?

- A. In the Amazon Redshift console, choose to configure cross-Region snapshots and set the destination Region as ap-northeast-1. Restore the Amazon Redshift Cluster from the snapshot and connect to Amazon QuickSight launched in ap-northeast-1.
- B. Create a VPC endpoint from the Amazon QuickSight VPC to the Amazon Redshift VPC so Amazon QuickSight can access data from Amazon Redshift.
- C. Create an Amazon Redshift endpoint connection string with Region information in the string and use this connection string in Amazon QuickSight to connect to

Amazon Redshift.

D. Create a new security group for Amazon Redshift in us-east-1 with an inbound rule authorizing access from the appropriate IP address range for the Amazon QuickSight servers in ap-northeast-1.

Answer: B

NEW QUESTION 44

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 46

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

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

Answer: B

Explanation:

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

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

NEW QUESTION 48

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 51

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 55

A company receives data from its vendor in JSON format with a timestamp in the file name. The vendor uploads the data to an Amazon S3 bucket, and the data is registered into the company's data lake for analysis and reporting. The company has configured an S3 Lifecycle policy to archive all files to S3 Glacier after 5 days.

The company wants to ensure that its AWS Glue crawler catalogs data only from S3 Standard storage and ignores the archived files. A data analytics specialist must implement a solution to achieve this goal without changing the current S3 bucket configuration.

Which solution meets these requirements?

- A. Use the exclude patterns feature of AWS Glue to identify the S3 Glacier files for the crawler to exclude.
- B. Schedule an automation job that uses AWS Lambda to move files from the original S3 bucket to a new S3 bucket for S3 Glacier storage.
- C. Use the excludeStorageClasses property in the AWS Glue Data Catalog table to exclude files on S3 Glacier storage
- D. Use the include patterns feature of AWS Glue to identify the S3 Standard files for the crawler to include.

Answer: A

NEW QUESTION 57

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 60

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 64

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

Which solution meets these requirements?

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

Answer: D

NEW QUESTION 68

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

What should the data analyst do reduce this latency?

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

D. Configure multiple Lambda functions to process the stream.

Answer: D

NEW QUESTION 73

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 77

A hospital is building a research data lake to ingest data from electronic health records (EHR) systems from multiple hospitals and clinics. The EHR systems are independent of each other and do not have a common patient identifier. The data engineering team is not experienced in machine learning (ML) and has been asked to generate a unique patient identifier for the ingested records.

Which solution will accomplish this task?

- A. An AWS Glue ETL job with the FindMatches transform
- B. Amazon Kendra
- C. Amazon SageMaker Ground Truth
- D. An AWS Glue ETL job with the ResolveChoice transform

Answer: A

Explanation:

Matching Records with AWS Lake Formation FindMatches

NEW QUESTION 81

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

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

Answer: D

NEW QUESTION 82

A company recently created a test AWS account to use for a development environment. The company also created a production AWS account in another AWS Region. As part of its security testing, the company wants to send log data from Amazon CloudWatch Logs in its production account to an Amazon Kinesis data stream in its test account.

Which solution will allow the company to accomplish this goal?

- A. Create a subscription filter in the production account's CloudWatch Logs to target the Kinesis data stream in the test account as its destination. In the test account, create an IAM role that grants access to the Kinesis data stream and the CloudWatch Logs resources in the production account.
- B. In the test account, create an IAM role that grants access to the Kinesis data stream and the CloudWatch Logs resources in the production account. Create a destination data stream in Kinesis Data Streams in the test account with an IAM role and a trust policy that allow CloudWatch Logs in the production account to write to the test account.
- C. In the test account, create an IAM role that grants access to the Kinesis data stream and the CloudWatch Logs resources in the production account. Create a destination data stream in Kinesis Data Streams in the test account with an IAM role and a trust policy that allow CloudWatch Logs in the production account to write to the test account.
- D. Create a destination data stream in Kinesis Data Streams in the test account with an IAM role and a trust policy that allow CloudWatch Logs in the production account to write to the test account. Create a subscription filter in the production account's CloudWatch Logs to target the Kinesis data stream in the test account as its destination.

Answer: D

NEW QUESTION 83

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

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

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

Which solution meets these requirements with minimal effort?

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

Answer: C

Explanation:

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

NEW QUESTION 88

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 90

A company wants to optimize the cost of its data and analytics platform. The company is ingesting a number of .csv and JSON files in Amazon S3 from various data sources. Incoming data is expected to be 50 GB each day. The company is using Amazon Athena to query the raw data in Amazon S3 directly. Most queries aggregate data from the past 12 months, and data that is older than 5 years is infrequently queried. The typical query scans about 500 MB of data and is expected to return results in less than 1 minute. The raw data must be retained indefinitely for compliance requirements.

Which solution meets the company's requirements?

- A. Use an AWS Glue ETL job to compress, partition, and convert the data into a columnar data format.
- B. Use Athena to query the processed dataset.
- C. Configure a lifecycle policy to move the processed data into the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class 5 years after object creation. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after object creation.
- D. Use an AWS Glue ETL job to partition and convert the data into a row-based data format.
- E. Use Athena to query the processed dataset.
- F. Configure a lifecycle policy to move the data into the Amazon S3 Standard- Infrequent Access (S3 Standard-IA) storage class 5 years after object creation.
- G. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after object creation.
- H. Use an AWS Glue ETL job to compress, partition, and convert the data into a columnar data format.
- I. Use Athena to query the processed dataset.
- J. Configure a lifecycle policy to move the processed data into the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class 5 years after the object was last accessed.
- K. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after the last date the object was accessed.
- L. Use an AWS Glue ETL job to partition and convert the data into a row-based data format.
- M. Use Athena to query the processed dataset.
- N. Configure a lifecycle policy to move the data into the Amazon S3 Standard- Infrequent Access (S3 Standard-IA) storage class 5 years after the object was last accessed.
- O. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after the last date the object was accessed.

Answer: A

NEW QUESTION 92

A company owns facilities with IoT devices installed across the world. The company is using Amazon Kinesis Data Streams to stream data from the devices to Amazon S3. The company's operations team wants to get insights from the IoT data to monitor data quality at ingestion. The insights need to be derived in near-real time, and the output must be logged to Amazon DynamoDB for further analysis.

Which solution meets these requirements?

- A. Connect Amazon Kinesis Data Analytics to analyze the stream data.
- B. Save the output to DynamoDB by using the default output from Kinesis Data Analytics.

- C. Connect Amazon Kinesis Data Analytics to analyze the stream data.
- D. Save the output to DynamoDB by using an AWS Lambda function.
- E. Connect Amazon Kinesis Data Firehose to analyze the stream data by using an AWS Lambda function. Save the output to DynamoDB by using the default output from Kinesis Data Firehose.
- F. Connect Amazon Kinesis Data Firehose to analyze the stream data by using an AWS Lambda function. Save the data to Amazon S3. Then run an AWS Glue job on schedule to ingest the data into DynamoDB.

Answer: C

NEW QUESTION 96

.....

Thank You for Trying Our Product

We offer two products:

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

2nd - Questions and Answers in PDF Format

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

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

100% Actual & Verified — Instant Download, Please Click
[Order The AWS-Certified-Data-Analytics-Specialty Practice Test Here](#)