Exam Questions AWS-Certified-Data-Analytics-Specialty

AWS Certified Data Analytics - Specialty

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

NEW QUESTION 1
A retail company leverages Amazon Athena for ad-hoc queries against an AWS Glue Data Catalog. The data analytics team manages the data catalog and data access for the company. The data analytics team wants to separate queries and manage the cost of running those queries by different workloads and teams. Ideally, the data analysts want to group the queries run by different users within a team, store the query results in individual Amazon S3 buckets specific to each team, and enforce cost constraints on the queries run against the Data Catalog.
Which solution meets these requirements?

A. Create IAM groups and resource tags for each team within the compan
B. Set up IAM policies that controluser access and actions on the Data Catalog resources.
C. Create Athena resource groups for each team within the company and assign users to these group
D. Add S3 bucket names and other query configurations to the properties list for the resource groups.
E. Create Athena workgroups for each team within the compan
F. Set up IAM workgroup policies that control user access and actions on the workgroup resources.
G. Create Athena query groups for each team within the company and assign users to the groups.

Answer: C

Explanation:
https://aws.amazon.com/about-aws/whats-new/2019/02/athena_workgroups/

NEW QUESTION 2
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_datacsv 20210708_datacsv
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 3
A company is sending historical datasets to Amazon S3 for storage. A data engineer at the company wants to make these datasets available for analysis using Amazon Athena. The engineer also wants to encrypt the Athena query results in an S3 results location by using AWS solutions for encryption. The requirements for encrypting the query results are as follows:
Use custom keys for encryption of the primary dataset query results. Use generic encryption for all other query results.
Provide an audit trail for the primary dataset queries that shows when the keys were used and by whom.
Which solution meets these requirements?

A. Use server-side encryption with S3 managed encryption keys (SSE-S3) for the primary datase
B. Use SSE-S3 for the other datasets.
C. Use server-side encryption with customer-provided encryption keys (SSE-C) for the primary dataset.Use server-side encryption with S3 managed encryption keys (SSE-S3) for the other datasets.
D. Use server-side encryption with AWS KMS managed customer master keys (SSE-KMS CMKs) for the primary datase
E. Use server-side encryption with S3 managed encryption keys (SSE-S3) for the other datasets.
F. Use client-side encryption with AWS Key Management Service (AWS KMS) customer managed keys for the primary datase
G. Use S3 client-side encryption with client-side keys for the other datasets.

Answer: A

NEW QUESTION 4
A company needs to store objects containing log data in JSON format. The objects are generated by eight applications running in AWS. Six of the applications generate a total of 500 KiB of data per second, and two of the applications can generate up to 2 MiB of data per second.
A data engineer wants to implement a scalable solution to capture and store usage data in an Amazon S3
bucket. The usage data objects need to be reformatted, converted to .csv format, and then compressed before they are stored in Amazon S3. The company requires the solution to include the least custom code possible and has authorized the data engineer to request a service quota increase if needed.
Which solution meets these requirements?

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

Answer: A

NEW QUESTION 5
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 6
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 7
A global company has different sub-organizations, and each sub-organization sells its products and services in various countries. The company's senior leadership wants to quickly identify which sub-organization is the strongest performer in each country. All sales data is stored in Amazon S3 in Parquet format.
Which approach can provide the visuals that senior leadership requested with the least amount of effort?

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

Answer: A

NEW QUESTION 8
A financial company uses Apache Hive on Amazon EMR for ad-hoc queries. Users are complaining of sluggish performance.
A data analyst notes the following:
- Approximately 90% of queries are submitted 1 hour after the market opens.
- Hadoop Distributed File System (HDFS) utilization never exceeds 10%.

Which solution would help address the performance issues?

A. Create instance fleet configurations for core and task node
B. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch CapacityRemainingGB metri
C. Create an automatic scaling policy to scale in the instance fleet based on the CloudWatch CapacityRemainingGB metric.
D. Create instance fleet configurations for core and task node
E. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch YARNMemoryAvailablePercentage metri
F. Create an automatic scaling policy to scale in the instance fleet based on the CloudWatch YARNMemoryAvailablePercentage metric.
G. Create instance group configurations for core and task node
H. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch CapacityRemainingGB metri
I. Create anautomatic scaling policy to scale in the instance groups based on the CloudWatch CapacityRemainingGB metric.
J. Create instance group configurations for core and task node
K. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch YARNMemoryAvailablePercentage metri
L. Create an automatic scaling policy to scale in the instance groups based on the CloudWatch YARNMemoryAvailablePercentage metric.

Answer: D

Explanation:
https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html

NEW QUESTION 9
A large energy company is using Amazon QuickSight to build dashboards and report the historical usage data of its customers This data is hosted in Amazon Redshift The reports need access to all the fact tables' billions ot records to create aggregation in real time grouping by multiple dimensions
A data analyst created the dataset in QuickSight by using a SQL query and not SPICE Business users have noted that the response time is not fast enough to meet their needs
Which action would speed up the response time for the reports with the LEAST implementation effort?

A. Use QuickSight to modify the current dataset to use SPICE
B. Use AWS Glue to create an Apache Spark job that joins the fact table with the dimension
C. Load the data into a new table
D. Use Amazon Redshift to create a materialized view that joins the fact table with the dimensions
E. Use Amazon Redshift to create a stored procedure that joins the fact table with the dimensions Load the data into a new table

Answer: A

NEW QUESTION 10
An online retail company uses Amazon Redshift to store historical sales transactions. The company is required to encrypt data at rest in the clusters to comply with the Payment Card Industry Data Security Standard (PCI DSS). A corporate governance policy mandates management of encryption keys using an on-premises hardware security module (HSM).
Which solution meets these requirements?

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

Answer: B

NEW QUESTION 10
An advertising company has a data lake that is built on Amazon S3. The company uses AWS Glue Data Catalog to maintain the metadata. The data lake is several years old and its overall size has increased exponentially as additional data sources and metadata are stored in the data lake. The data lake administrator wants to implement a mechanism to simplify permissions management between Amazon S3 and the Data Catalog to keep them in sync
Which solution will simplify permissions management with minimal development effort?

A. Set AWS Identity and Access Management (1AM) permissions tor AWS Glue
B. Use AWS Lake Formation permissions
C. Manage AWS Glue and S3 permissions by using bucket policies
D. Use Amazon Cognito user pools.

Answer: B

NEW QUESTION 15
A reseller that has thousands of AWS accounts receives AWS Cost and Usage Reports in an Amazon S3 bucket The reports are delivered to the S3 bucket in the following format
<examp/e-reporT-prefix>/<examp/e-report-rtame>/yyyymmdd-yyyymmdd/<examp/e-report-name> parquet An AWS Glue crawler crawls the S3 bucket and populates an AWS Glue Data Catalog with a table Business analysts use Amazon Athena to query the table and create monthly summary reports for the AWS accounts
The business analysts are experiencing slow queries because of the accumulation of reports from the last 5 years The business analysts want the operations team to make changes to improve query performance
Which action should the operations team take to meet these requirements?

A. Change the file format to csv.zip.
B. Partition the data by date and account ID
C. Partition the data by month and account ID
D. Partition the data by account ID, year, and month

Answer: B

NEW QUESTION 20
An online retail company is migrating its reporting system to AWS. The company's legacy system runs data processing on online transactions using a complex series of nested Apache Hive queries. Transactional data is exported from the online system to the reporting system several times a day. Schemas in the files are stable between updates.
A data analyst wants to quickly migrate the data processing to AWS, so any code changes should be minimized. To keep storage costs low, the data analyst decides to store the data in Amazon S3. It is vital that the data from the reports and associated analytics is completely up to date based on the data in Amazon S3.
Which solution meets these requirements?

A. Create an AWS Glue Data Catalog to manage the Hive metadat
B. Create an AWS Glue crawler over Amazon S3 that runs when data is refreshed to ensure that data changes are update
C. Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
D. Create an AWS Glue Data Catalog to manage the Hive metadat
E. Create an Amazon EMR cluster with consistent view enable
F. Run emrfs sync before each analytics step to ensure data changes are update
G. Create an EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.

H. Create an Amazon Athena table with CREATE TABLE AS SELECT (CTAS) to ensure data is refreshed from underlying queries against the raw datase
I. Create an AWS Glue Data Catalog to manage the Hive metadata over the CTAS tabl
J. Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
K. Use an S3 Select query to ensure that the data is properly update
L. Create an AWS Glue Data Catalog to manage the Hive metadata over the S3 Select tabl
M. Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.

Answer: A

NEW QUESTION 22
A company that produces network devices has millions of users. Data is collected from the devices on an hourly basis and stored in an Amazon S3 data lake.
The company runs analyses on the last 24 hours of data flow logs for abnormality detection and to
troubleshoot and resolve user issues. The company also analyzes historical logs dating back 2 years to discover patterns and look for improvement opportunities.
The data flow logs contain many metrics, such as date, timestamp, source IP, and target IP. There are about 10 billion events every day.
How should this data be stored for optimal performance?

A. In Apache ORC partitioned by date and sorted by source IP
B. In compressed .csv partitioned by date and sorted by source IP
C. In Apache Parquet partitioned by source IP and sorted by date
D. In compressed nested JSON partitioned by source IP and sorted by date

Answer: A

NEW QUESTION 25
An analytics software as a service (SaaS) provider wants to offer its customers business intelligence <BI) reporting capabilities that are self-service The provider is using Amazon QuickSight to build these reports The data for the reports resides in a multi-tenant database, but each customer should only be able to access their own data
The provider wants to give customers two user role options
• Read-only users for individuals who only need to view dashboards
• Power users for individuals who are allowed to create and share new dashboards with other users Which QuickSight feature allows the provider to meet these requirements'?

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

Answer: A

NEW QUESTION 30
An ecommerce company stores customer purchase data in Amazon RDS. The company wants a solution to store and analyze historical data. The most recent 6 months of data will be queried frequently for analytics workloads. This data is several terabytes large. Once a month, historical data for the last 5 years must be accessible and will be joined with the more recent data. The company wants to optimize performance and cost.
Which storage solution will meet these requirements?

A. Create a read replica of the RDS database to store the most recent 6 months of dat
B. Copy the historical data into Amazon S3. Create an AWS Glue Data Catalog of the data in Amazon S3 and Amazon RD
C. Run historical queries using Amazon Athena.
D. Use an ETL tool to incrementally load the most recent 6 months of data into an Amazon Redshift cluste
E. Run more frequent queries against this cluste
F. Create a read replica of the RDS database to run queries on the historical data.
G. Incrementally copy data from Amazon RDS to Amazon S3. Create an AWS Glue Data Catalog of the data in Amazon S3. Use Amazon Athena to query the data.
H. Incrementally copy data from Amazon RDS to Amazon S3. Load and store the most recent 6 months of data in Amazon Redshif
I. Configure an Amazon Redshift Spectrum table to connect to all historical data.

Answer: D

NEW QUESTION 31
A company has developed several AWS Glue jobs to validate and transform its data from Amazon S3 and load it into Amazon RDS for MySQL in batches once every day. The ETL jobs read the S3 data using a DynamicFrame. Currently, the ETL developers are experiencing challenges in processing only the incremental data on every run, as the AWS Glue job processes all the S3 input data on each run.
Which approach would allow the developers to solve the issue with minimal coding effort?

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

Answer: B

NEW QUESTION 32
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 33
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 38
A marketing company wants to improve its reporting and business intelligence capabilities. During the planning phase, the company interviewed the relevant stakeholders and discovered that:

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

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

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

Answer: B

NEW QUESTION 42
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 be 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 one processing semantics of Apache Flink and Apache Spark Streaming included in Amazon EMR.

Answer: A

NEW QUESTION 47
An insurance company has raw data in JSON format that is sent without a predefined schedule through an Amazon Kinesis Data Firehose delivery stream to an Amazon S3 bucket. An AWS Glue crawler is scheduled to run every 8 hours to update the schema in the data catalog of the tables stored in the S3 bucket. Data analysts analyze the data using Apache Spark SQL on Amazon EMR set up with AWS Glue Data Catalog as the metastore. Data analysts say that, occasionally, the data they receive is stale. A data engineer needs to provide access to the most up-to-date data.
Which solution meets these requirements?

A. Create an external schema based on the AWS Glue Data Catalog on the existing Amazon Redshift cluster to query new data in Amazon S3 with Amazon

Redshift Spectrum.
B. Use Amazon CloudWatch Events with the rate (1 hour) expression to execute the AWS Glue crawler every hour.
C. Using the AWS CLI, modify the execution schedule of the AWS Glue crawler from 8 hours to 1 minute.
D. Run the AWS Glue crawler from an AWS Lambda function triggered by an S3:ObjectCreated:* eventnotification on the S3 bucket.

Answer: D

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

NEW QUESTION 51
A company has developed an Apache Hive script to batch process data stared 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 ste
B. Set KeepJobFlowAliveWhenNoSteps to false and disable the termination protection fla
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 Hu
E. Hive, and Apache Oozi
F. Set the termination protection flag to true and use Spot Instances for the core nodes of the cluste
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 52
A streaming application is reading data from Amazon Kinesis Data Streams and immediately writing the data to an Amazon S3 bucket every 10 seconds. The application is reading data from hundreds of shards. The batch interval cannot be changed due to a separate requirement. The data is being accessed by Amazon Athena. Users are seeing degradation in query performance as time progresses.
Which action can help improve query performance?

A. Merge the files in Amazon S3 to form larger files.
B. Increase the number of shards in Kinesis Data Streams.
C. Add more memory and CPU capacity to the streaming application.
D. Write the files to multiple S3 buckets.

Answer: A

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

NEW QUESTION 56
A marketing company is storing its campaign response data in Amazon S3. A consistent set of sources has generated the data for each campaign. The data is saved into Amazon S3 as .csv files. A business analyst will use Amazon Athena to analyze each campaign's data. The company needs the cost of ongoing data analysis with Athena to be minimized.
Which combination of actions should a data analytics specialist take to meet these requirements? (Choose two.)

A. Convert the .csv files to Apache Parquet.
B. Convert the .csv files to Apache Avro.
C. Partition the data by campaign.
D. Partition the data by source.
E. Compress the .csv files.

Answer: AC

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

NEW QUESTION 61
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 65
A banking company is currently using an Amazon Redshift cluster with dense storage (DS) nodes to store sensitive data. An audit found that the cluster is unencrypted. Compliance requirements state that a database with sensitive data must be encrypted through a hardware security module (HSM) with automated key rotation.
Which combination of steps is required to achieve compliance? (Choose two.)

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

Answer: BD

NEW QUESTION 70
A company needs to collect streaming data from several sources and store the data in the AWS Cloud. The dataset is heavily structured, but analysts need to perform several complex SQL queries and need consistent performance. Some of the data is queried more frequently than the rest. The company wants a solution that meets its performance requirements in a cost-effective manner.
Which solution meets these requirements?

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

Answer: B

NEW QUESTION 74
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 tabl
B. Load all the data except for the columns containing sensitive data into a second tabl
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 tabl
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 75
A company has an application that uses the Amazon Kinesis Client Library (KCL) to read records from a Kinesis data stream.
After a successful marketing campaign, the application experienced a significant increase in usage. As a result, a data analyst had to split some shards in the data stream. When the shards were split, the application started
throwing an ExpiredIteratorExceptions error sporadically. What should the data analyst do to resolve this?

A. Increase the number of threads that process the stream records.
B. Increase the provisioned read capacity units assigned to the stream's Amazon DynamoDB table.
C. Increase the provisioned write capacity units assigned to the stream's Amazon DynamoDB table.
D. Decrease the provisioned write capacity units assigned to the stream's Amazon DynamoDB table.

Answer: C

NEW QUESTION 80
A company is building a data lake and needs to ingest data from a relational database that has time-series data. The company wants to use managed services to accomplish this. The process needs to be scheduled daily and bring incremental data only from the source into Amazon S3.
What is the MOST cost-effective approach to meet these requirements?

A. Use AWS Glue to connect to the data source using JDBC Driver
B. Ingest incremental records only using job bookmarks.
C. Use AWS Glue to connect to the data source using JDBC Driver
D. Store the last updated key in an Amazon DynamoDB table and ingest the data using the updated key as a filter.
E. Use AWS Glue to connect to the data source using JDBC Drivers and ingest the entire datase
F. Use appropriate Apache Spark libraries to compare the dataset, and find the delta.
G. Use AWS Glue to connect to the data source using JDBC Drivers and ingest the full dat
H. Use AWS DataSync to ensure the delta only is written into Amazon S3.

Answer: A

Explanation:
https://docs.aws.amazon.com/glue/latest/dg/monitor-continuations.html

NEW QUESTION 82
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 85
A data analytics specialist is setting up workload management in manual mode for an Amazon Redshift environment. The data analytics specialist is defining query monitoring rules to manage system performance and user experience of an Amazon Redshift cluster.
Which elements must each query monitoring rule include?

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

Answer: C

NEW QUESTION 88
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 applicatio
B. Configure each security group to have access to a specific topic in the Amazon MSK cluste
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 topi
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 applicatio
H. Create an Amazon MSK cluster and Kafka topic for each applicatio
I. Configure each security group to have access to the specific cluster.

Answer: B

NEW QUESTION 91
A data analyst runs a large number of data manipulation language (DML) queries by using Amazon Athena with the JDBC driver. Recently, a query failed after It ran for 30 minutes. The query returned the following message Java.sql.SGLException: 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 92
An airline has .csv-formatted data stored in Amazon S3 with an AWS Glue Data Catalog. Data analysts want to join this data with call center data stored in Amazon Redshift as part of a dally batch process. The Amazon Redshift cluster is already under a heavy load. The solution must be managed, serverless, well-functioning, and minimize the load on the existing Amazon Redshift cluster. The solution should also require minimal effort and development activity.
Which solution meets these requirements?

A. Unload the call center data from Amazon Redshift to Amazon S3 using an AWS Lambda function.Perform the join with AWS Glue ETL scripts.
B. Export the call center data from Amazon Redshift using a Python shell in AWS Glu
C. Perform the join with AWS Glue ETL scripts.
D. Create an external table using Amazon Redshift Spectrum for the call center data and perform the join with Amazon Redshift.
E. Export the call center data from Amazon Redshift to Amazon EMR using Apache Sqoo
F. Perform the join with Apache Hive.

Answer: C

Explanation:
https://docs.aws.amazon.com/redshift/latest/dg/c-spectrum-external-tables.html

NEW QUESTION 93
An online retailer is rebuilding its inventory management system and inventory reordering system to automatically reorder products by using Amazon Kinesis Data Streams. The inventory management system uses the Kinesis Producer Library (KPL) to publish data to a stream. The inventory reordering system uses the Kinesis Client Library (KCL) to consume data from the stream. The stream has been configured to scale as needed. Just before production deployment, the retailer discovers that the inventory reordering system is receiving duplicated data.
Which factors could be causing the duplicated data? (Choose two.)

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

Answer: BD

NEW QUESTION 98
A company is migrating its existing on-premises ETL jobs to Amazon EMR. The code consists of a series of jobs written in Java. The company needs to reduce overhead for the system administrators without changing the underlying code. Due to the sensitivity of the data, compliance requires that the company use root device volume encryption on all nodes in the cluster. Corporate standards require that environments be provisioned though AWS CloudFormation when possible.
Which solution satisfies these requirements?

A. Install open-source Hadoop on Amazon EC2 instances with encrypted root device volume
B. Configure the cluster in the CloudFormation template.
C. Use a CloudFormation template to launch an EMR cluste
D. In the configuration section of the cluster, define a bootstrap action to enable TLS.
E. Create a custom AMI with encrypted root device volume
F. Configure Amazon EMR to use the custom AMI using the CustomAmiId property in the CloudFormation template.
G. Use a CloudFormation template to launch an EMR cluste
H. In the configuration section of the cluster, define a bootstrap action to encrypt the root device volume of every node.

Answer: C

NEW QUESTION 103
A company wants to improve the data load time of a sales data dashboard. Data has been collected as .csv files and stored within an Amazon S3 bucket that is partitioned by date. The data is then loaded to an Amazon Redshift data warehouse for frequent analysis. The data volume is up to 500 GB per day.
Which solution will improve the data loading performance?

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

Answer: B

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

NEW QUESTION 105
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 forma
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 tabl
F. Then filter over two partitions and put all columns in the WHERE clause

Answer: C

NEW QUESTION 106
A mobile gaming company wants to capture data from its gaming app and make the data available for analysis immediately. The data record size will be approximately 20 KB. The company is concerned about achieving optimal throughput from each device. Additionally, the company wants to develop a data stream processing application with dedicated throughput for each consumer.
Which solution would achieve this goal?

A. Have the app call the PutRecords API to send data to Amazon Kinesis Data Stream
B. Use the enhanced fan-out feature while consuming the data.
C. Have the app call the PutRecordBatch API to send data to Amazon Kinesis Data Firehos
D. Submit a support case to enable dedicated throughput on the account.
E. Have the app use Amazon Kinesis Producer Library (KPL) to send data to Kinesis Data Firehos

F. Use the enhanced fan-out feature while consuming the data.
G. Have the app call the PutRecords API to send data to Amazon Kinesis Data Stream
H. Host the stream- processing application on Amazon EC2 with Auto Scaling.

Answer: A

Explanation:
https://docs.aws.amazon.com/streams/latest/dev/enhanced-consumers.html

NEW QUESTION 107
A retail company is building its data warehouse solution using Amazon Redshift. As a part of that effort, the company is loading hundreds of files into the fact table created in its Amazon Redshift cluster. The company wants the solution to achieve the highest throughput and optimally use cluster resources when loading data into the company's fact table.
How should the company meet these requirements?

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

Answer: D

Explanation:
https://docs.aws.amazon.com/redshift/latest/dg/c_best-practices-single-copy-command.html

NEW QUESTION 111
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 116
A company wants to enrich application logs in near-real-time and use the enriched dataset for further analysis. The application is running on Amazon EC2 instances across multiple Availability Zones and storing its logs using Amazon CloudWatch Logs. The enrichment source is stored in an Amazon DynamoDB table.
Which solution meets the requirements for the event collection and enrichment?

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

Answer: A

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

NEW QUESTION 118
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 outers, 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 120
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 122
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 retrie
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 126
A company using Amazon QuickSight Enterprise edition has thousands of dashboards analyses and datasets. The company struggles to manage and assign permissions for granting users access to various items within QuickSight. The company wants to make it easier to implement sharing and permissions management.
Which solution should the company implement to simplify permissions management?

A. Use QuickSight folders to organize dashboards, analyses, and datasets Assign individual users permissions to these folders
B. Use QuickSight folders to organize dashboards analyses, and datasets Assign group permissions by using these folders.
C. Use AWS 1AM resource-based policies to assign group permissions to QuickSight items
D. Use QuickSight user management APIs to provision group permissions based on dashboard naming conventions

Answer: C

NEW QUESTION 128
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 Redshif
B. Create an external table in Amazon Redshift to point to the S3 locatio
C. Use Amazon Redshift Spectrum to join to data that is older than 13 months.
D. Take a snapshot of the Amazon Redshift cluste
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-Tierin
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 129
A company is hosting an enterprise reporting solution with Amazon Redshift. The application provides reporting capabilities to three main groups: an executive group to access financial reports, a data analyst group to run long-running ad-hoc queries, and a data engineering group to run stored procedures and ETL processes. The executive team requires queries to run with optimal performance. The data engineering team expects queries to take minutes.
Which Amazon Redshift feature meets the requirements for this task?

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

Answer: D

Explanation:
Materialized views:

NEW QUESTION 130
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 135
A retail company has 15 stores across 6 cities in the United States. Once a month, the sales team requests a visualization in Amazon QuickSight that provides the ability to easily identify revenue trends across cities and stores. The visualization also helps identify outliers that need to be examined with further analysis.
Which visual type in QuickSight meets the sales team's requirements?

A. Geospatial chart
B. Line chart
C. Heat map
D. Tree map

Answer: A

NEW QUESTION 136
A financial company hosts a data lake in Amazon S3 and a data warehouse on an Amazon Redshift cluster. The company uses Amazon QuickSight to build dashboards and wants to secure access from its on-premises Active Directory to Amazon QuickSight.
How should the data be secured?

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

Answer: A

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

NEW QUESTION 139
A data analytics specialist is building an automated ETL ingestion pipeline using AWS Glue to ingest compressed files that have been uploaded to an Amazon S3 bucket. The ingestion pipeline should support incremental data processing.
Which AWS Glue feature should the data analytics specialist use to meet this requirement?

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

Answer: C

NEW QUESTION 143
A manufacturing company wants to create an operational analytics dashboard to visualize metrics from equipment in near-real time. The company uses Amazon Kinesis Data Streams to stream the data to other applications. The dashboard must automatically refresh every 5 seconds. A data analytics specialist must design a solution that requires the least possible implementation effort.
Which solution meets these requirements?

A. Use Amazon Kinesis Data Firehose to store the data in Amazon S3. Use Amazon QuickSight to build the dashboard.
B. Use Apache Spark Streaming on Amazon EMR to read the data in near-real tim
C. Develop a custom application for the dashboard by using D3.js.
D. Use Amazon Kinesis Data Firehose to push the data into an Amazon Elasticsearch Service (Amazon ES) cluste
E. Visualize the data by using a Kibana dashboard.
F. Use AWS Glue streaming ETL to store the data in Amazon S3. Use Amazon QuickSight to build the dashboard.

Answer: B

NEW QUESTION 148
A bank wants to migrate a Teradata data warehouse to the AWS Cloud The bank needs a solution for reading large amounts of data and requires the highest possible performance. The solution also must maintain the separation of storage and compute
Which solution meets these requirements?

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

Answer: C

NEW QUESTION 150
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 accounts CloudWatch Logs to target the Kinesis data stream in the test account as its destination In the test account create an 1AM 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 1AM 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 1AM 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 1AM 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 1AM 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 1AM 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 accounts CloudWatch Logs to target the Kinesis data stream in the test account as its destination

Answer: D

NEW QUESTION 155
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 Catalo
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 jo
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 157
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 159
A large ride-sharing company has thousands of drivers globally serving millions of unique customers every day. The company has decided to migrate an existing data mart to Amazon Redshift. The existing schema includes the following tables.
A trips fact table for information on completed rides. A drivers dimension table for driver profiles. A customers fact table holding customer profile information.
The company analyzes trip details by date and destination to examine profitability by region. The drivers data rarely changes. The customers data frequently changes.
What table design provides optimal query performance?

A. Use DISTSTYLE KEY (destination) for the trips table and sort by dat
B. Use DISTSTYLE ALL for the drivers and customers tables.
C. Use DISTSTYLE EVEN for the trips table and sort by dat
D. Use DISTSTYLE ALL for the drivers table.Use DISTSTYLE EVEN for the customers table.
E. Use DISTSTYLE KEY (destination) for the trips table and sort by dat
F. Use DISTSTYLE ALL for the drivers tabl
G. Use DISTSTYLE EVEN for the customers table.
H. Use DISTSTYLE EVEN for the drivers table and sort by dat
I. Use DISTSTYLE ALL for both fact tables.

Answer: C

Explanation:
https://www.matillion.com/resources/blog/aws-redshift-performance-choosing-the-right-distribution-styles/#:~:t
https://docs.aws.amazon.com/redshift/latest/dg/c_best-practices-best-dist-key.html

NEW QUESTION 160
A company wants to optimize the cost of its data and analytics platform. The company is ingesting a number of .c sv 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 forma
B. Use Athena to query the processed datase
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 forma
E. Use Athena to query the processed datase
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 creatio
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 forma
I. Use Athena to query the processed datase
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 accesse
K. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier forlong-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 forma
M. Use Athena to query the processed datase
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 accesse
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 163
A company is building an analytical solution that includes Amazon S3 as data lake storage and Amazon Redshift for data warehousing. The company wants to use Amazon Redshift Spectrum to query the data that is stored in Amazon S3.
Which steps should the company take to improve performance when the company uses Amazon Redshift Spectrum to query the S3 data files? (Select THREE)
Use gzip compression with individual file sizes of 1-5 GB

A. Use a columnar storage file format
B. Partition the data based on the most common query predicates
C. Split the data into KB-sized files.
D. Keep all files about the same size.
E. Use file formats that are not splittable

Answer: BCD

NEW QUESTION 168
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 170
A software company wants to use instrumentation data to detect and resolve errors to improve application recovery time. The company requires API usage anomalies, like error rate and response time spikes, to be detected in near-real time (NRT) The company also requires that data analysts have access to dashboards for log analysis in NRT
Which solution meets these requirements'?

A. Use Amazon Kinesis Data Firehose as the data transport layer for logging data Use Amazon Kinesis Data Analytics to uncover the NRT API usage anomalies Use Kinesis Data Firehose to deliver log data to Amazon OpenSearch Service (Amazon Elasticsearch Service) for search, log analytics, and application monitoring Use OpenSearch Dashboards (Kibana) in Amazon OpenSearch Service (Amazon Elasticsearch Service) for the dashboards.
B. Use Amazon Kinesis Data Analytics as the data transport layer for logging dat
C. Use Amazon Kinesis Data Streams to uncover NRT monitoring metric
D. Use Amazon Kinesis Data Firehose to deliver log data to Amazon OpenSearch Service (Amazon Elasticsearch Service) for search, log analytics, and application monitoring Use Amazon QuickSight for the dashboards
E. Use Amazon Kinesis Data Analytics as the data transport layer for logging data and to uncover NRT monitoring metrics Use Amazon Kinesis Data Firehose to deliver log data to Amazon OpenSearch Service (Amazon Elasticsearch Service) for search, log analytics, and application monitoring Use OpenSearch Dashboards (Kibana) in Amazon OpenSearch Service (Amazon Elasticsearch Service) for the dashboards
F. Use Amazon Kinesis Data Firehose as the data transport layer for logging data Use Amazon Kinesis Data Analytics to uncover NRT monitoring metrics Use Amazon Kinesis Data Streams to deliver logdata to Amazon OpenSearch Service (Amazon Elasticsearch Service) for search, log analytics, and application monitoring Use Amazon QuickSight for the dashboards.

Answer: C

NEW QUESTION 171
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 teams 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 bucke
B. Add the additional IAM roles to the cluster’s EMR role for the EC2 trust polic
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 bucke
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 bucke
H. Add the service role for the EMR cluster EC2 instances to the trust polices 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 bucke
K. Add the service role for the EMR cluster EC2 instances to the trust polices 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 173
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 177
A company has several Amazon EC2 instances sitting behind an Application Load Balancer (ALB) The company wants its IT Infrastructure team to analyze the IP addresses coming into the company's ALB The ALB is configured to store access logs in Amazon S3 The access logs create about 1 TB of data each day, and access to the data will be infrequent The company needs a solution that is scalable, cost-effective and has minimal maintenance requirements
Which solution meets these requirements?

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

Answer: D

NEW QUESTION 181
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 182
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 183
.....

THANKS FOR TRYING THE DEMO OF OUR PRODUCT

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

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

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

Money Back Guarantee

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

* AWS-Certified-Data-Analytics-Specialty Questions and Answers Updated Frequently

* AWS-Certified-Data-Analytics-Specialty Practice Questions Verified by Expert Senior Certified Staff

* AWS-Certified-Data-Analytics-Specialty Most Realistic Questions that Guarantee you a Pass on Your FirstTry

* AWS-Certified-Data-Analytics-Specialty Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year