Amazon.DAS-C01.v2024-02-15.q109

Exam Code:	DAS-C01
Exam Name:	AWS Certified Data Analytics - Specialty (DAS-C01) Exam
Certification Provider:	Amazon
Free Question Number:	109
Version:	v2024-02-15
# of views:	486
# of Questions views:	1090
https://www.dumpsdb.com/dumps/Amazon/DAS-C01/Amazon.DAS-C01.v2024-02-15.q109	

NEW QUESTION: 1
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:* event notification on the S3 bucket.
Answer: D (LEAVE A REPLY)
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: 2
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. Save the output to DynamoDB by using an AWS Lambda function.
B. Connect Amazon Kinesis Data Analytics to analyze the stream data. Save the output to DynamoDB by using the default output from Kinesis Data Analytics.
C. 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.
D. 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: (SHOW ANSWER)

NEW QUESTION: 3
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 2 months of data in Amazon Redshift and the rest of the months in Amazon S3. Set up an external schema and table for Amazon Redshift Spectrum. Configure Amazon QuickSight with Amazon Redshift as the data source.
B. 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.
C. Store the last 24 months of data in Amazon Redshift. Configure Amazon QuickSight with Amazon Redshift as the data source.

D. Store the last 2 months of data in Amazon Redshift and the rest of the months in Amazon S3. Use a long- running Amazon EMR with Apache Spark cluster to query the data as needed. Configure Amazon QuickSight with Amazon EMR as the data source.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 4
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. Use Amazon Athena to query the S3 data
B. Use Amazon EMR and Apache Hive to query the S3 data
C. Copy the data into Amazon Redshift and query the data
D. Use Amazon Redshift Spectrum to query the S3 data
Answer: D (LEAVE A REPLY)

NEW QUESTION: 5
A data analyst is using AWS Glue to organize, cleanse, validate, and format a 200 GB dataset. The data analyst triggered the job to run with the Standard worker type. After 3 hours, the AWS Glue job status is still RUNNING. Logs from the job run show no error codes. The data analyst wants to improve the job execution time without overprovisioning.
Which actions should the data analyst take?
A. Enable job metrics in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the maximum capacity job parameter.
B. Enable job metrics in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the spark.yarn.executor.memoryOverhead job parameter.
C. Enable job bookmarks in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the executor-cores job parameter.
D. Enable job bookmarks in AWS Glue to estimate the number of data processing units (DPUs). Based on the profiled metrics, increase the value of the num-executors job parameter.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 6
A company is migrating its existing on-premises ETL jobs to Amazon EMR. The code consists of a series of jobs written in Jav a. 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. Use a CloudFormation template to launch an EMR cluster. In the configuration section of the cluster, define a bootstrap action to enable TLS.
B. Create a custom AMI with encrypted root device volumes. Configure Amazon EMR to use the custom AMI using the CustomAmiId property in the CloudFormation template.
C. Use a CloudFormation template to launch an EMR cluster. In the configuration section of the cluster, define a bootstrap action to encrypt the root device volume of every node.
D. Install open-source Hadoop on Amazon EC2 instances with encrypted root device volumes. Configure the cluster in the CloudFormation template.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 7
A company has a data warehouse in Amazon Redshift that is approximately 500 TB in size. New data is imported every few hours and read-only queries are run throughout the day and evening. There is a particularly heavy load with no writes for several hours each morning on business days. During those hours, some queries are queued and take a long time to execute. The company needs to optimize query execution and avoid any downtime.
What is the MOST cost-effective solution?
A. Enable concurrency scaling in the workload management (WLM) queue.
B. Add more nodes using the AWS Management Console during peak hours. Set the distribution style to ALL.
C. Use elastic resize to quickly add nodes during peak times. Remove the nodes when they are not needed.
D. Use a snapshot, restore, and resize operation. Switch to the new target cluster.
Answer: A (LEAVE A REPLY)
https://docs.aws.amazon.com/redshift/latest/dg/cm-c-implementing-workload-management.html

NEW QUESTION: 8
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. Create a manifest file that contains the data file locations and issue a COPY command to load the data into Amazon Redshift.
B. 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.
C. Load all the data files in parallel to Amazon Aurora, and run an AWS Glue job to load the data into Amazon Redshift.

D. Use Amazon EMR to copy all the data files into one folder and issue a COPY command to load the data into Amazon Redshift.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 9
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 compressed .csv partitioned by date and sorted by source IP
B. In compressed nested JSON partitioned by source IP and sorted by date
C. In Apache Parquet partitioned by source IP and sorted by date
D. In Apache ORC partitioned by date and sorted by source IP
Answer: B (LEAVE A REPLY)

NEW QUESTION: 10
An online retail company with millions of users around the globe wants to improve its ecommerce analytics capabilities. Currently, clickstream data is uploaded directly to Amazon S3 as compressed files. Several times each day, an application running on Amazon EC2 processes the data and makes search options and reports available for visualization by editors and marketers. The company wants to make website clicks and aggregated data available to editors and marketers in minutes to enable them to connect with users more effectively.
Which options will help meet these requirements in the MOST efficient way? (Choose two.)
A. Upload clickstream records to Amazon S3 as compressed files. Then use AWS Lambda to send data to Amazon Elasticsearch Service from Amazon S3.
B. Upload clickstream records from Amazon S3 to Amazon Kinesis Data Streams and use a Kinesis Data Streams consumer to send records to Amazon Elasticsearch Service.
C. Use Amazon Kinesis Data Firehose to upload compressed and batched clickstream records to Amazon Elasticsearch Service.
D. Use Amazon Elasticsearch Service deployed on Amazon EC2 to aggregate, filter, and process the data.
Refresh content performance dashboards in near-real time.
E. Use Kibana to aggregate, filter, and visualize the data stored in Amazon Elasticsearch Service.
Refresh content performance dashboards in near-real time.
Answer: (SHOW ANSWER)

NEW QUESTION: 11

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 Streams to stream data to Amazon S3. Use Amazon Athena as a data source for Amazon QuickSight to create a business intelligence dashboard.
B. Use Amazon Kinesis Data Streams to stream data to Amazon Redshift. Use Amazon Redshift as a data source for Amazon QuickSight to create a business intelligence dashboard.
C. Use Amazon Kinesis Data Firehose to stream data to Amazon Redshift. 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 S3. Use Amazon Athena as a data source for Amazon QuickSight to create a business intelligence dashboard.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 12
A market data company aggregates external data sources to create a detailed view of product consumption in different countries. The company wants to sell this data to external parties through a subscription. To achieve this goal, the company needs to make its data securely available to external parties who are also AWS users.
What should the company do to meet these requirements with the LEAST operational overhead?
A. Store the data in Amazon S3. Share the data by using S3 bucket ACLs.
B. Store the data in Amazon S3. Share the data by using presigned URLs for security.
C. Upload the data to AWS Data Exchange for storage. Share the data by using presigned URLs for security.
D. Upload the data to AWS Data Exchange for storage. Share the data by using the AWS Data Exchange sharing wizard.
Answer: (SHOW ANSWER)

NEW QUESTION: 13
An ecommerce company is migrating its business intelligence environment from on premises to the AWS Cloud. The company will use Amazon Redshift in a public subnet and Amazon QuickSight. The tables already are loaded into Amazon Redshift and can be accessed by a SQL tool.
The company starts QuickSight for the first time. During the creation of the data source, a data analytics specialist enters all the information and tries to validate the connection. An error with the following message occurs: "Creating a connection to your data source timed out." How should the data analytics specialist resolve this error?
A. Grant the SELECT permission on Amazon Redshift tables.
B. Add the QuickSight IP address range into the Amazon Redshift security group.
C. Create an IAM role for QuickSight to access Amazon Redshift.

D. Use a QuickSight admin user for creating the dataset.

Answer: (SHOW ANSWER)

Explanation

Connection to the database times out

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

NEW QUESTION: 14

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

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

A. Place the required installation scripts in Amazon S3 and execute them using custom bootstrap actions.

B. Place the required installation scripts in Amazon S3 and execute them through Apache Spark in Amazon EMR.

C. Install the required third-party libraries in the existing EMR master node. Create an AMI out of that master node and use that custom AMI to re-create the EMR cluster.

D. Use an Amazon DynamoDB table to store the list of required applications. Trigger an AWS Lambda function with DynamoDB Streams to install the software.

E. Launch an Amazon EC2 instance with Amazon Linux and install the required third-party libraries on the instance. Create an AMI and use that AMI to create the EMR cluster.

Answer: (SHOW ANSWER)

Explanation

https://aws.amazon.com/about-aws/whats-new/2017/07/amazon-emr-now-supports-launching-clusters-with-custo

https://docs.aws.amazon.com/de_de/emr/latest/ManagementGuide/emr-plan-bootstrap.html

NEW QUESTION: 15

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. Create an AWS Glue crawler and specify the table as the source.
D. 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. Migrate the Hive catalog to the Data Catalog.
Answer: C (LEAVE A REPLY)
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: 16
An online gaming company is using an Amazon Kinesis Data Analytics SQL application with a Kinesis data stream as its source. The source sends three non-null fields to the application: player_id, score, and us_5_digit_zip_code.
A data analyst has a .csv mapping file that maps a small number of us_5_digit_zip_code values to a territory code. The data analyst needs to include the territory code, if one exists, as an additional output of the Kinesis Data Analytics application.
How should the data analyst meet this requirement while minimizing costs?
A. Store the contents of the mapping file in an Amazon DynamoDB table. Change the Kinesis Data Analytics application to send its output to an AWS Lambda function that fetches the mapping and supplements each record to include the territory code, if one exists. Forward the record from the Lambda function to the original application destination.
B. Store the mapping file in an Amazon S3 bucket and configure it as a reference data source for the Kinesis Data Analytics application. Change the SQL query in the application to include a join to the reference table and add the territory code field to the SELECT columns.
C. Store the mapping file in an Amazon S3 bucket and configure the reference data column headers for the
csv file in the Kinesis Data Analytics application. Change the SQL query in the application to include a join to the file's S3 Amazon Resource Name (ARN), and add the territory code field to the SELECT columns.
D. Store the contents of the mapping file in an Amazon DynamoDB table. Preprocess the records as they arrive in the Kinesis Data Analytics application with an AWS Lambda function that fetches the mapping and supplements each record to include the territory code, if one exists. Change the SQL query in the application to include the new field in the SELECT statement.
Answer: (SHOW ANSWER)

NEW QUESTION: 17
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 QuickSight with a direct SQL query
B. Use Amazon Athena to query the clickstream data in Amazon S3
C. Use Amazon Redshift to store and query the clickstream data
D. Use the QuickSight SPICE engine with a daily refresh
E. Use S3 analytics to query the clickstream data
Answer: (SHOW ANSWER)

NEW QUESTION: 18
A company wants to improve user satisfaction for its smart home system by adding more features to its recommendation engine. Each sensor asynchronously pushes its nested JSON data into Amazon Kinesis Data Streams using the Kinesis Producer Library (KPL) in Java. Statistics from a set of failed sensors showed that, when a sensor is malfunctioning, its recorded data is not always sent to the cloud. The company needs a solution that offers near-real-time analytics on the data from the most updated sensors.
Which solution enables the company to meet these requirements?
A. Set the RecordMaxBufferedTime property of the KPL to "1" to disable the buffering on the sensor side.
Use Kinesis Data Analytics to enrich the data based on a company-developed anomaly detection SQL script. Push the enriched data to a fleet of Kinesis data streams and enable the data transformation feature to flatten the JSON file. Instantiate a dense storage Amazon Redshift cluster and use it as the destination for the Kinesis Data Firehose delivery stream.
B. Update the sensors code to use the PutRecord/PutRecords call from the Kinesis Data Streams API with the AWS SDK for Java. Use Kinesis Data Analytics to enrich the data based on a company-developed anomaly detection SQL script. Direct the output of KDA application to a Kinesis Data Firehose delivery stream, enable the data transformation feature to flatten the JSON file, and set the Kinesis Data Firehose destination to an Amazon Elasticsearch Service cluster.

C. Set the RecordMaxBufferedTime property of the KPL to "0" to disable the buffering on the sensor side.
Connect for each stream a dedicated Kinesis Data Firehose delivery stream and enable the data transformation feature to flatten the JSON file before sending it to an Amazon S3 bucket. Load the S3 data into an Amazon Redshift cluster.
D. Update the sensors code to use the PutRecord/PutRecords call from the Kinesis Data Streams API with the AWS SDK for Java. Use AWS Glue to fetch and process data from the stream using the Kinesis Client Library (KCL). Instantiate an Amazon Elasticsearch Service cluster and use AWS Lambda to directly push data into it.
Answer: B (LEAVE A REPLY)
Explanation
https://docs.aws.amazon.com/streams/latest/dev/developing-producers-with-kpl.html The KPL can incur an additional processing delay of up to RecordMaxBufferedTime within the library (user-configurable). Larger values of RecordMaxBufferedTime results in higher packing efficiencies and better performance. Applications that cannot tolerate this additional delay may need to use the AWS SDK directly.

NEW QUESTION: 19
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: (SHOW ANSWER)
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: 20
A retail company stores order invoices in an Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster Indices on the cluster are created monthly Once a new month begins, no new writes are made to any of the indices from the previous months The company has been expanding the storage on the Amazon OpenSearch Service {Amazon Elasticsearch Service) cluster to avoid running out of space, but the company wants to reduce costs Most searches on the cluster are on the most recent 3 months of data while the audit team requires infrequent access to older data to generate periodic reports The most recent 3 months of data must be quickly available for queries, but the audit team can tolerate slower queries if the solution saves on cluster costs Which of the following is the MOST operationally efficient solution to meet these requirements?
A. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to store the indices in Amazon S3 Glacier When the audit team requires the archived data restore the archived indices back to the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster
B. Archive indices that are older than 3 months by taking manual snapshots and storing the snapshots in Amazon S3 When the audit team requires the archived data, restore the archived indices back to the Amazon OpenSearch Service (Amazon Elasticsearch Service) cluster
C. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to migrate the indices to Amazon OpenSearch Service (Amazon Elasticsearch Service) UltraWarm storage
D. Archive indices that are older than 3 months by using Index State Management (ISM) to create a policy to migrate the indices to Amazon OpenSearch Service (Amazon Elasticsearch Service) UltraWarm storage When the audit team requires the older data: migrate the indices in UltraWarm storage back to hot storage
Answer: (SHOW ANSWER)

NEW QUESTION: 21
A company analyzes historical data and needs to query data that is stored in Amazon S3. New data is generated daily as .csv files that are stored in Amazon S3. The company's analysts are using Amazon Athena to perform SQL queries against a recent subset of the overall data. The amount of data that is ingested into Amazon S3 has increased substantially over time, and the query latency also has increased.
Which solutions could the company implement to improve query performance? (Choose two.)
A. Use Athena to extract the data and store it in Apache Parquet format on a daily basis. Query the extracted data.
B. Run a daily AWS Glue ETL job to compress the data files by using the .lzo format. Query the compressed data.
C. Use MySQL Workbench on an Amazon EC2 instance, and connect to Athena by using a JDBC or ODBC connector. Run the query from MySQL Workbench instead of Athena directly.

D. Run a daily AWS Glue ETL job to convert the data files to Apache Parquet and to partition the converted files. Create a periodic AWS Glue crawler to automatically crawl the partitioned data on a daily basis.
E. Run a daily AWS Glue ETL job to compress the data files by using the .gzip format. Query the compressed data.
Answer: A,D (LEAVE A REPLY)

NEW QUESTION: 22
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: A,C,E (LEAVE A REPLY)
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: 23
A company launched a service that produces millions of messages every day and uses Amazon Kinesis Data Streams as the streaming service.
The company uses the Kinesis SDK to write data to Kinesis Data Streams. A few months after launch, a data analyst found that write performance is significantly reduced. The data analyst investigated the metrics and determined that Kinesis is throttling the write requests. The data analyst wants to address this issue without significant changes to the architecture.
Which actions should the data analyst take to resolve this issue? (Choose two.)
A. Increase the Kinesis Data Streams retention period to reduce throttling.
B. Replace the Kinesis API-based data ingestion mechanism with Kinesis Agent.
C. Increase the number of shards in the stream using the UpdateShardCount API.
D. Choose partition keys in a way that results in a uniform record distribution across shards.
E. Customize the application code to include retry logic to improve performance.
Answer: C,D (LEAVE A REPLY)
https://aws.amazon.com/blogs/big-data/under-the-hood-scaling-your-kinesis-data-streams/

NEW QUESTION: 24

A company has a business unit uploading .csv files to an Amazon S3 bucket. The company's data platform team has set up an AWS Glue crawler to do discovery, and create tables and schemas. An AWS Glue job writes processed data from the created tables to an Amazon Redshift database. The AWS Glue job handles column mapping and creating the Amazon Redshift table appropriately. When the AWS Glue job is rerun for any reason in a day, duplicate records are introduced into the Amazon Redshift table.
Which solution will update the Redshift table without duplicates when jobs are rerun?
A. Modify the AWS Glue job to copy the rows into a staging table. Add SQL commands to replace the existing rows in the main table as postactions in the DynamicFrameWriter class.
B. Load the previously inserted data into a MySQL database in the AWS Glue job. Perform an upsert operation in MySQL, and copy the results to the Amazon Redshift table.
C. Use Apache Spark's DataFrame dropDuplicates() API to eliminate duplicates and then write the data to Amazon Redshift.
D. Use the AWS Glue ResolveChoice built-in transform to select the most recent value of the column.
Answer: (SHOW ANSWER)
https://aws.amazon.com/premiumsupport/knowledge-center/sql-commands-redshift-glue-job/ See the section Merge an Amazon Redshift table in AWS Glue (upsert)

NEW QUESTION: 25
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 nodes. Schedule automated snapshots using Amazon EventBridge.
B. Store the data on an EMR File System (EMRFS) instead of HDFS. Enable EMRFS consistent view. Create an EMR HBase cluster with multiple master nodes. Point the HBase root directory to an Amazon S3 bucket.
C. 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 Zones. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.
D. 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 nodes. Create a secondary EMR HBase read- replica cluster in a separate Availability Zone. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 26
A company has developed several AWS Glue jobs to validate and transform its data from Amazon S3 and load it into Amazon RDS for MySQL in batches once every day. The ETL jobs read the S3 data using a DynamicFrame. Currently, the ETL developers are experiencing challenges in processing only the incremental data on every run, as the AWS Glue job processes all the S3 input data on each run.
Which approach would allow the developers to solve the issue with minimal coding effort?
A. Have the ETL jobs delete the processed objects or data from Amazon S3 after each run.
B. Have the ETL jobs read the data from Amazon S3 using a DataFrame.
C. Create custom logic on the ETL jobs to track the processed S3 objects.
D. Enable job bookmarks on the AWS Glue jobs.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 27
An advertising company has a data lake that is built on Amazon S3. The company uses AWS Glue Data Catalog to maintain the metadat a. 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. Use Amazon Cognito user pools.
B. Use AWS Lake Formation permissions
C. Set AWS Identity and Access Management (1AM) permissions tor AWS Glue
D. Manage AWS Glue and S3 permissions by using bucket policies
Answer: B (LEAVE A REPLY)

NEW QUESTION: 28
A company wants to enrich application logs in near-real-time and use the enriched dataset for further analysis.
The application is running on Amazon EC2 instances across multiple Availability Zones and storing its logs using Amazon CloudWatch Logs. The enrichment source is stored in an Amazon DynamoDB table.
Which solution meets the requirements for the event collection and enrichment?
A. Use a CloudWatch Logs subscription to send the data to Amazon Kinesis Data Firehose. Use AWS Lambda to transform the data in the Kinesis Data Firehose delivery stream and enrich it with the data in the DynamoDB table. Configure Amazon S3 as the Kinesis Data Firehose delivery destination.
B. Export the raw logs to Amazon S3 on an hourly basis using the AWS CLI. Use AWS Glue crawlers to catalog the logs. Set up an AWS Glue connection for the DynamoDB table and set up an AWS Glue ETL job to enrich the data. Store the enriched data in Amazon S3.
C. Configure the application to write the logs locally and use Amazon Kinesis Agent to send the data to Amazon Kinesis Data Streams. Configure a Kinesis Data Analytics SQL application with the Kinesis

data stream as the source. Join the SQL application input stream with DynamoDB records, and then store the enriched output stream in Amazon S3 using Amazon Kinesis Data Firehose.
D. Export the raw logs to Amazon S3 on an hourly basis using the AWS CLI. Use Apache Spark SQL on Amazon EMR to read the logs from Amazon S3 and enrich the records with the data from DynamoDB.
Store the enriched data in Amazon S3.
Answer: (SHOW ANSWER)
Explanation
https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/SubscriptionFilters.html#FirehoseExample

NEW QUESTION: 29
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 (LEAVE A REPLY)
Explanation
https://github.com/awsdocs/amazon-redshift-developer-guide/issues/21

NEW QUESTION: 30
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. Once Lake Formation has the data, apply permissions on Lake Formation.
B. To create the data catalog, run an AWS Glue crawler on the existing Parquet data. Register the Amazon S3 path and then apply permissions through Lake Formation to provide granular-level security.
C. Install Apache Ranger on an Amazon EC2 instance and integrate with Amazon EMR. Using Ranger policies, create role-based access control for the existing data assets in Amazon S3.

D. Create multiple IAM roles for different users and groups. Assign IAM roles to different data assets in Amazon S3 to create table-based and column-based access controls.

Answer: A (LEAVE A REPLY)

Explanation

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

NEW QUESTION: 31

A gaming company is collecting cllckstream 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. 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 ikand 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

B. 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.

C. 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.

D. 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.

Answer: (SHOW ANSWER)

NEW QUESTION: 32

A company is streaming its high-volume billing data (100 MBps) to Amazon Kinesis Data Streams. A data analyst partitioned the data on account_id to ensure that all records belonging to an account go to the same Kinesis shard and order is maintained. While building a custom consumer using the Kinesis Java SDK, the data analyst notices that, sometimes, the messages arrive out of order for account_id. Upon further investigation, the data analyst discovers the messages that are out of order seem to be arriving from different shards for the same account_id and are seen when a stream resize runs.
What is an explanation for this behavior and what is the solution?
A. The records are not being received by Kinesis Data Streams in order. The producer should use the PutRecords API call instead of the PutRecord API call with the SequenceNumberForOrdering parameter.
B. The hash key generation process for the records is not working correctly. The data analyst should generate an explicit hash key on the producer side so the records are directed to the appropriate shard accurately.
C. There are multiple shards in a stream and order needs to be maintained in the shard. The data analyst needs to make sure there is only a single shard in the stream and no stream resize runs.
D. The consumer is not processing the parent shard completely before processing the child shards after a stream resize. The data analyst should process the parent shard completely first before processing the child shards.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 33
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 PutRecordBatch API to send data to Amazon Kinesis Data Firehose. Submit a support case to enable dedicated throughput on the account.
B. Have the app use Amazon Kinesis Producer Library (KPL) to send data to Kinesis Data Firehose. Use the enhanced fan-out feature while consuming the data.
C. Have the app call the PutRecords API to send data to Amazon Kinesis Data Streams. Host the stream- processing application on Amazon EC2 with Auto Scaling.
D. Have the app call the PutRecords API to send data to Amazon Kinesis Data Streams. Use the enhanced fan-out feature while consuming the data.
Answer: (SHOW ANSWER)

NEW QUESTION: 34
A power utility company is deploying thousands of smart meters to obtain real-time updates about power consumption. The company is using Amazon Kinesis Data Streams to collect the data streams from smart meters. The consumer application uses the Kinesis Client Library (KCL) to retrieve the stream data. The company has only one consumer application.

The company observes an average of 1 second of latency from the moment that a record is written to the stream until the record is read by a consumer application. The company must reduce this latency to 500 milliseconds.
Which solution meets these requirements?
A. Use enhanced fan-out in Kinesis Data Streams.
B. Increase the number of shards for the Kinesis data stream.
C. Reduce the propagation delay by overriding the KCL default settings.
D. Develop consumers by using Amazon Kinesis Data Firehose.
Answer: C (LEAVE A REPLY)
Explanation
The KCL defaults are set to follow the best practice of polling every 1 second. This default results in average propagation delays that are typically below 1 second.

NEW QUESTION: 35
A telecommunications company is looking for an anomaly-detection solution to identify fraudulent calls. The company currently uses Amazon Kinesis to stream voice call records in a JSON format from its on-premises database to Amazon S3. The existing dataset contains voice call records with 200 columns. To detect fraudulent calls, the solution would need to look at 5 of these columns only.
The company is interested in a cost-effective solution using AWS that requires minimal effort and experience in anomaly-detection algorithms.
Which solution meets these requirements?
A. Use Kinesis Data Firehose to detect anomalies on a data stream from Kinesis by running SQL queries, which compute an anomaly score for all calls and store the output in Amazon RDS. Use Amazon Athena to build a dataset and Amazon QuickSight to visualize the results.
B. Use an AWS Glue job to transform the data from JSON to Apache Parquet. Use AWS Glue crawlers to discover the schema and build the AWS Glue Data Catalog. Use Amazon Athena to create a table with a subset of columns. Use Amazon QuickSight to visualize the data and then use Amazon QuickSight machine learning-powered anomaly detection.
C. Use an AWS Glue job to transform the data from JSON to Apache Parquet. Use AWS Glue crawlers to discover the schema and build the AWS Glue Data Catalog. Use Amazon SageMaker to build an anomaly detection model that can detect fraudulent calls by ingesting data from Amazon S3.
D. Use Kinesis Data Analytics to detect anomalies on a data stream from Kinesis by running SQL queries, which compute an anomaly score for all calls. Connect Amazon QuickSight to Kinesis Data Analytics to visualize the anomaly scores.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 36
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 step. Set KeepJobFlowAliveWhenNoSteps to false and disable the termination protection flag. Use Amazon CloudWatch Events to schedule the Lambda function to run daily.
B. Create an AWS Glue job with the Hive script to perform the batch operation. Configure the job to run once a day using a time-based schedule.
C. 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.
D. Use the AWS Management Console to spin up an Amazon EMR cluster with Python Hue. Hive, and Apache Oozie. Set the termination protection flag to true and use Spot Instances for the core nodes of the cluster. Configure an Oozie workflow in the cluster to invoke the Hive script daily.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 37
A company is planning to do a proof of concept for a machine learning (ML) project using Amazon SageMaker with a subset of existing on-premises data hosted in the company's 3 TB data warehouse. For part of the project, AWS Direct Connect is established and tested. To prepare the data for ML, data analysts are performing data curation. The data analysts want to perform multiple step, including mapping, dropping null fields, resolving choice, and splitting fields. The company needs the fastest solution to curate the data for this project.
Which solution meets these requirements?
A. Create custom ETL jobs on-premises to curate the data. Use AWS DMS to ingest data into Amazon S3 for ML processing.
B. Take a full backup of the data store and ship the backup files using AWS Snowball. Upload Snowball data into Amazon S3 and schedule data curation jobs using AWS Batch to prepare the data for ML.
C. Ingest data into Amazon S3 using AWS DMS. Use AWS Glue to perform data curation and store the data in Amazon S3 for ML processing.
D. Ingest data into Amazon S3 using AWS DataSync and use Apache Spark scrips to curate the data in an Amazon EMR cluster. Store the curated data in Amazon S3 for ML processing.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 38
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 full access to Amazon S3. Create three additional IAM roles, each granting access to each team's specific bucket. Add the service role for the EMR cluster EC2 instances to the trust polices for the base IAM roles. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.

B. For the EMR cluster Amazon EC2 instances, create a service role that grants no access to Amazon S3. Create three additional IAM roles, each granting access to each team's specific bucket. Add the service role for the EMR cluster EC2 instances to the trust policies for the additional IAM roles. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.
C. For the EMR cluster Amazon EC2 instances, create a service role that grants no access to Amazon S3. Create three additional IAM roles, each granting access to each team's specific bucket. Add the additional IAM roles to the cluster's EMR role for the EC2 trust policy. 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 full access to Amazon S3. Create three additional IAM roles, each granting access to each team's specific bucket. Add the service role for the EMR cluster EC2 instances to the trust polices for the additional IAM roles. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.
Answer: (SHOW ANSWER)

NEW QUESTION: 39
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 Streams to collect all the data from toll stations. 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. Use AWS Lambda to publish an Amazon Simple Notification Service (Amazon SNS) notification if the threshold is not met. Connect Amazon Kinesis Data Firehose to Kinesis Data Streams to deliver the data to Amazon Redshift.
B. Use Amazon Kinesis Data Firehose to collect data and deliver it to Amazon Redshift. 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.
C. Use Amazon Kinesis Data Firehose to collect data and deliver it to Amazon Redshift and Amazon Kinesis Data Analytics simultaneously. 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. Use AWS Lambda to publish an Amazon Simple Notification Service (Amazon SNS) notification if the threshold is not met.

D. Use Amazon Kinesis Data Firehose to collect data and deliver it to Amazon Redshift and Amazon Kinesis Data Analytics simultaneously. 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 (LEAVE A REPLY)

NEW QUESTION: 40
A company uses Amazon Redshift as its data warehouse. A new table has columns that contain sensitive data.
The data in the table will eventually be referenced by several existing queries that run many times a day.
A data analyst needs to load 100 billion rows of data into the new table. Before doing so, the data analyst must ensure that only members of the auditing group can read the columns containing sensitive data.
How can the data analyst meet these requirements with the lowest maintenance overhead?
A. Load all the data into the new table and grant the auditing group permission to read from the table. Load all the data except for the columns containing sensitive data into a second table. Grant the appropriate users read-only permissions to the second table.
B. Load all the data into the new table and grant the auditing group permission to read from the table. Use the GRANT SQL command to allow read-only access to a subset of columns to the appropriate users.
C. 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.
D. 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: (SHOW ANSWER)
Explanation
https://aws.amazon.com/blogs/big-data/achieve-finer-grained-data-security-with-column-level-access-control-in-

NEW QUESTION: 41
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. Set up S3 bucket policies to provide marketing employees with appropriate data access under compliance controls. Delete the bucket policies after the project.
B. Load tabular data from Amazon S3 to an Amazon EMR cluster using s3DistCp. 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 controls. Terminate the EMR cluster after the project.
C. Load tabular data from Amazon S3 to Amazon Redshift with the COPY command. Use the built-in row- level security feature in Amazon Redshift to provide marketing employees with appropriate data access under compliance controls. Delete the Amazon Redshift tables after the project.
D. Load tabular data from Amazon S3 to Amazon QuickSight Enterprise edition by directly importing it as a data source. Use the built-in row-level security feature in Amazon QuickSight to provide marketing employees with appropriate data access under compliance controls. Delete Amazon QuickSight data sources after the project is complete.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 42
A company is streaming its high-volume billing data (100 MBps) to Amazon Kinesis Data Streams. A data analyst partitioned the data on account_id to ensure that all records belonging to an account go to the same Kinesis shard and order is maintained. While building a custom consumer using the Kinesis Java SDK, the data analyst notices that, sometimes, the messages arrive out of order for account_id. Upon further investigation, the data analyst discovers the messages that are out of order seem to be arriving from different shards for the same account_id and are seen when a stream resize runs.
What is an explanation for this behavior and what is the solution?
A. There are multiple shards in a stream and order needs to be maintained in the shard. The data analyst needs to make sure there is only a single shard in the stream and no stream resize runs.
B. The hash key generation process for the records is not working correctly. The data analyst should generate an explicit hash key on the producer side so the records are directed to the appropriate shard accurately.
C. The records are not being received by Kinesis Data Streams in order. The producer should use the PutRecords API call instead of the PutRecord API call with the SequenceNumberForOrdering parameter.

D. The consumer is not processing the parent shard completely before processing the child shards after a stream resize. The data analyst should process the parent shard completely first before processing the child shards.

Answer: D (LEAVE A REPLY)

https://docs.aws.amazon.com/streams/latest/dev/kinesis-using-sdk-java-after-resharding.html the parent shards that remain after the reshard could still contain data that you haven't read yet that was added to the stream before the reshard. If you read data from the child shards before having read all data from the parent shards, you could read data for a particular hash key out of the order given by the data records' sequence numbers. Therefore, assuming that the order of the data is important, you should, after a reshard, always continue to read data from the parent shards until it is exhausted. Only then should you begin reading data from the child shards.

NEW QUESTION: 43

A manufacturing company uses Amazon Connect to manage its contact center and Salesforce to manage its customer relationship management (CRM) data. The data engineering team must build a pipeline to ingest data from the contact center and CRM system into a data lake that is built on Amazon S3.

What is the MOST efficient way to collect data in the data lake with the LEAST operational overhead?

A. Use Amazon Kinesis Data Firehose to ingest Amazon Connect data and Amazon Kinesis Data Streams to ingest Salesforce data.

B. Use Amazon Kinesis Data Firehose to ingest Amazon Connect data and Amazon AppFlow to ingest Salesforce data.

C. Use Amazon Kinesis Data Streams to ingest Amazon Connect data and Amazon AppFlow to ingest Salesforce data.

D. Use Amazon AppFlow to ingest Amazon Connect data and Amazon Kinesis Data Firehose to ingest Salesforce data.

Answer: A (LEAVE A REPLY)

NEW QUESTION: 44

An Amazon Redshift database contains sensitive user data. Logging is necessary to meet compliance requirements. The logs must contain database authentication attempts, connections, and disconnections. The logs must also contain each query run against the database and record which database user ran each query.

Which steps will create the required logs?

A. Enable audit logging for Amazon Redshift using the AWS Management Console or the AWS CLI.

B. Enable and download audit reports from AWS Artifact.

C. Allow access to the Amazon Redshift database using AWS IAM only. Log access using AWS CloudTrail.

D. Enable Amazon Redshift Enhanced VPC Routing. Enable VPC Flow Logs to monitor traffic.

Answer: A (LEAVE A REPLY)

NEW QUESTION: 45
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 (LEAVE A REPLY)
Explanation
Materialized views:

NEW QUESTION: 46
An ecommerce company ingests a large set of clickstream data in JSON format and stores the data in Amazon S3. Business analysts from multiple product divisions need to use Amazon Athena to analyze the dat a. The company's analytics team must design a solution to monitor the daily data usage for Athena by each product division. The solution also must produce a warning when a divisions exceeds its quota Which solution will meet these requirements with the LEAST operational overhead?
A. Create an AWS account for each division Provide cross-account access to an AWS Glue Data Catalog to all the accounts. Set an Amazon CloudWatch alarm to monitor Athena usage. Use Amazon Simple Notification Service (Amazon SNS) to send notifications.
B. Use a CREATE TABLE AS SELECT (CTAS) statement to create separate tables for each product division Use AWS Budgets to track Athena usage Configure a threshold for the budget Use Amazon Simple Notification Service (Amazon SNS) to send notifications when thresholds are breached.
C. Create an AWS account for each division Configure an AWS Glue Data Catalog in each account Set an Amazon CloudWatch alarm to monitor Athena usage Use Amazon Simple Notification Service (Amazon SNS) to send notifications.
D. Create an Athena workgroup for each division Configure a data usage control for each workgroup and a time period of 1 day Configure an action to send notifications to an Amazon Simple Notification Service (Amazon SNS) topic
Answer: D (LEAVE A REPLY)

NEW QUESTION: 47
A company is planning to create a data lake in Amazon S3. The company wants to create tiered storage based on access patterns and cost objectives. The solution must include support for JDBC connections from legacy clients, metadata management that allows federation for access control, and batch-based ETL using PySpark and Scala Operational management should be limited.
Which combination of components can meet these requirements? (Choose three.)
A. Amazon EMR with Apache Hive, using an Amazon RDS with MySQL-compatible backed metastore
B. Amazon EMR with Apache Spark for ETL
C. Amazon EMR with Apache Hive for JDBC clients
D. Amazon Athena for querying data in Amazon S3 using JDBC drivers
E. AWS Glue Data Catalog for metadata management
F. AWS Glue for Scala-based ETL
Answer: A,B,D (LEAVE A REPLY)

NEW QUESTION: 48
A team of data scientists plans to analyze market trend data for their company's new investment strategy. The trend data comes from five different data sources in large volumes. The team wants to utilize Amazon Kinesis to support their use case. The team uses SQL-like queries to analyze trends and wants to send notifications based on certain significant patterns in the trends. Additionally, the data scientists want to save the data to Amazon S3 for archival and historical re-processing, and use AWS managed services wherever possible. The team wants to implement the lowest-cost solution.
Which solution meets these requirements?
A. Publish data to one Kinesis data stream. Deploy Kinesis Data Analytic to the stream for analyzing trends, and configure an AWS Lambda function as an output to send notifications using Amazon SNS. Configure Kinesis Data Firehose on the Kinesis data stream to persist data to an S3 bucket.
B. Publish data to two Kinesis data streams. Deploy a custom application using the Kinesis Client Library (KCL) to the first stream for analyzing trends, and send notifications using Amazon SNS. Configure Kinesis Data Firehose on the second Kinesis data stream to persist data to an S3 bucket.
C. Publish data to two Kinesis data streams. Deploy Kinesis Data Analytics to the first stream for analyzing trends, and configure an AWS Lambda function as an output to send notifications using Amazon SNS.
Configure Kinesis Data Firehose on the second Kinesis data stream to persist data to an S3 bucket.
D. Publish data to one Kinesis data stream. Deploy a custom application using the Kinesis Client Library (KCL) for analyzing trends, and send notifications using Amazon SNS. Configure Kinesis Data Firehose on the Kinesis data stream to persist data to an S3 bucket.
Answer: (SHOW ANSWER)

NEW QUESTION: 49

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. Enable HSM with key rotation through the AWS CLI.
C. Enable Elliptic Curve Diffie-Hellman Ephemeral (ECDHE) encryption in the HSM.
D. Modify the cluster with an HSM encryption option and automatic key rotation.
E. Create a new HSM-encrypted Amazon Redshift cluster and migrate the data to the new cluster.
Answer: (SHOW ANSWER)

NEW QUESTION: 50
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 workflows with AWS Glue jobs to perform transformations.
B. For archived data, use Amazon SageMaker to perform data transformations.
C. For daily incoming data, use Amazon Redshift to perform transformations.
D. For archived data, use Amazon EMR to perform data transformations.
E. For daily incoming data, use AWS Glue crawlers to scan and identify the schema.
F. For daily incoming data, use Amazon Athena to scan and identify the schema.
Answer: A,C,F (LEAVE A REPLY)

NEW QUESTION: 51
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 data. Copy the historical data into Amazon S3. Create an AWS Glue Data Catalog of the data in Amazon S3 and Amazon RDS.

Run historical queries using Amazon Athena.
B. Use an ETL tool to incrementally load the most recent 6 months of data into an Amazon Redshift cluster. Run more frequent queries against this cluster. Create a read replica of the RDS database to run queries on the historical data.
C. Incrementally copy data from Amazon RDS to Amazon S3. Load and store the most recent 6 months of data in Amazon Redshift. Configure an Amazon Redshift Spectrum table to connect to all historical data.
D. 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.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 52
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. Store the files in Apache Parquet format.
C. Update the table path to "s3://powertransformer/".
D. Ingest the files again.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 53
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 (LEAVE A REPLY)
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: 54
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 Streams. Use the enhanced fan-out feature while consuming the data.
B. Have the app call the PutRecordBatch API to send data to Amazon Kinesis Data Firehose. Submit a support case to enable dedicated throughput on the account.
C. Have the app use Amazon Kinesis Producer Library (KPL) to send data to Kinesis Data Firehose. Use the enhanced fan-out feature while consuming the data.
D. Have the app call the PutRecords API to send data to Amazon Kinesis Data Streams. Host the stream- processing application on Amazon EC2 with Auto Scaling.
Answer: A (LEAVE A REPLY)
https://docs.aws.amazon.com/streams/latest/dev/enhanced-consumers.html

NEW QUESTION: 55
A manufacturing company has many IoT devices in different facilities across the world The company is using Amazon Kinesis Data Streams to collect the data from the devices The company's operations team has started to observe many WnteThroughputExceeded exceptions The operations team determines that the reason is the number of records that are being written to certain shards The data contains device ID capture date measurement type, measurement value and facility ID The facility ID is used as the partition key Which action will resolve this issue?
A. Archive the data on the producers' side
B. Change the partition key from facility ID to a randomly generated key
C. Change the partition key from facility ID to capture date
D. Increase the number of shards
Answer: D (LEAVE A REPLY)

NEW QUESTION: 56

A company uses the Amazon Kinesis SDK to write data to Kinesis Data Streams. Compliance requirements state that the data must be encrypted at rest using a key that can be rotated. The company wants to meet this encryption requirement with minimal coding effort.
How can these requirements be met?
A. Create a customer master key (CMK) in AWS KMS. Assign the CMK an alias. Use the AWS Encryption SDK, providing it with the key alias to encrypt and decrypt the data.
B. Create a customer master key (CMK) in AWS KMS. Assign the CMK an alias. Enable server-side encryption on the Kinesis data stream using the CMK alias as the KMS master key.
C. Create a customer master key (CMK) in AWS KMS. Create an AWS Lambda function to encrypt and decrypt the data. Set the KMS key ID in the function's environment variables.
D. Enable server-side encryption on the Kinesis data stream using the default KMS key for Kinesis Data
Answer: B (LEAVE A REPLY)
Streams.

NEW QUESTION: 57
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. 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 Zones. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.
B. Store the data on an EMR File System (EMRFS) instead of HDFS. Enable EMRFS consistent view. Create an EMR HBase cluster with multiple master nodes. Point the HBase root directory to an Amazon S3 bucket.
C. Use Spot Instances for core and task nodes and a Reserved Instance for the EMR master node. Configure the EMR cluster with multiple master nodes. Schedule automated snapshots using Amazon EventBridge.
D. 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 nodes. Create a secondary EMR HBase read- replica cluster in a separate Availability Zone. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 58
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 (LEAVE A REPLY)
https://docs.aws.amazon.com/quicksight/latest/user/restrict-access-to-a-data-set-using-row-level-security.html

NEW QUESTION: 59
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 Drivers. Ingest incremental records only using job bookmarks.
B. Use AWS Glue to connect to the data source using JDBC Drivers. Store the last updated key in an Amazon DynamoDB table and ingest the data using the updated key as a filter.
C. Use AWS Glue to connect to the data source using JDBC Drivers and ingest the entire dataset. Use appropriate Apache Spark libraries to compare the dataset, and find the delta.
D. Use AWS Glue to connect to the data source using JDBC Drivers and ingest the full data. Use AWS DataSync to ensure the delta only is written into Amazon S3.
Answer: A (LEAVE A REPLY)
Explanation
https://docs.aws.amazon.com/glue/latest/dg/monitor-continuations.html

NEW QUESTION: 60
A technology company is creating a dashboard that will visualize and analyze time-sensitive data. The data will come in through Amazon Kinesis Data Firehose with the butter interval set to 60 seconds. The dashboard must support near-real-time data.
Which visualization solution will meet these requirements?
A. Select Amazon S3 as the endpoint for Kinesis Data Firehose. Read data into an Amazon SageMaker Jupyter notebook and carry out the desired analyses and visualizations.
B. Select Amazon S3 as the endpoint for Kinesis Data Firehose. Use AWS Glue to catalog the data and Amazon Athena to query it. Connect Amazon QuickSight with SPICE to Athena to create the desired analyses and visualizations.

C. Select Amazon Elasticsearch Service (Amazon ES) as the endpoint for Kinesis Data Firehose. Set up a Kibana dashboard using the data in Amazon ES with the desired analyses and visualizations.
D. Select Amazon Redshift as the endpoint for Kinesis Data Firehose. Connect Amazon QuickSight with SPICE to Amazon Redshift to create the desired analyses and visualizations.
Answer: (SHOW ANSWER)

NEW QUESTION: 61
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 (LEAVE A REPLY)
https://docs.aws.amazon.com/redshift/latest/dg/c_loading-data-best-practices.html

NEW QUESTION: 62
A company has collected more than 100 TB of log files in the last 24 months. The files are stored as raw text in a dedicated Amazon S3 bucket. Each object has a key of the form year-month-day_log_HHmmss.txt where HHmmss represents the time the log file was initially created. A table was created in Amazon Athena that points to the S3 bucket. One-time queries are run against a subset of columns in the table several times an hour.
A data analyst must make changes to reduce the cost of running these queries. Management wants a solution with minimal maintenance overhead.
Which combination of steps should the data analyst take to meet these requirements? (Choose three.)
A. Convert the log files to Apace Avro format.
B. Add a key prefix of the form date=year-month-day/ to the S3 objects to partition the data.
C. Convert the log files to Apache Parquet format.
D. Add a key prefix of the form year-month-day/ to the S3 objects to partition the data.

E. Drop and recreate the table with the PARTITIONED BY clause. Run the ALTER TABLE ADD PARTITION statement.
F. Drop and recreate the table with the PARTITIONED BY clause. Run the MSCK REPAIR TABLE statement.
Answer: B,C,F (LEAVE A REPLY)
Section: (none)
Explanation

NEW QUESTION: 63
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 workload name, a unique rule name, and a query runtime-based condition
B. A queue name, a unique rule name, and a predicate-based stop condition
C. A unique rule name, a query runtime condition, and an AWS Lambda function to resubmit any failed queries in off hours
D. A unique rule name, one to three predicates, and an action
Answer: (SHOW ANSWER)

NEW QUESTION: 64
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. Create an AWS Glue crawler and specify the table as the source.
D. 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. Migrate the Hive catalog to the Data Catalog.
Answer: C (LEAVE A REPLY)
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: 65
A marketing company is using Amazon EMR clusters for its workloads. The company manually installs third- party libraries on the clusters by logging in to the master nodes. A data analyst needs to create an automated solution to replace the manual process.
Which options can fulfill these requirements? (Choose two.)
A. Place the required installation scripts in Amazon S3 and execute them using custom bootstrap actions.
B. Place the required installation scripts in Amazon S3 and execute them through Apache Spark in Amazon EMR.
C. Install the required third-party libraries in the existing EMR master node. Create an AMI out of that master node and use that custom AMI to re-create the EMR cluster.
D. Use an Amazon DynamoDB table to store the list of required applications. Trigger an AWS Lambda function with DynamoDB Streams to install the software.
E. Launch an Amazon EC2 instance with Amazon Linux and install the required third-party libraries on the instance. Create an AMI and use that AMI to create the EMR cluster.
Answer: A,E (LEAVE A REPLY)
https://aws.amazon.com/about-aws/whats-new/2017/07/amazon-emr-now-supports-launching-clusters-with-custom-amazon-linux-amis/
https://docs.aws.amazon.com/de_de/emr/latest/ManagementGuide/emr-plan-bootstrap.html

NEW QUESTION: 66
A data engineer is using AWS Glue ETL jobs to process data at frequent intervals The processed data is then copied into Amazon S3 The ETL jobs run every 15 minutes. The AWS Glue Data Catalog partitions need to be updated automatically after the completion of each job Which solution will meet these requirements MOST cost-effectively?
A. Use the AWS Glue Data Catalog to manage the data catalog Define an AWS Glue workflow for the ETL process Define a trigger within the workflow that can start the crawler when an ETL job run is complete
B. Use an Apache Hive metastore to manage the data catalog Update the AWS Glue ETL code to include the enableUpdateCatalog and partitionKeys arguments.
C. Use the AWS Glue Data Catalog to manage the data catalog Use AWS Glue Studio to manage ETL jobs. Use the AWS Glue Studio feature that supports updates to the AWS Glue Data Catalog during job runs.
D. Use the AWS Glue Data Catalog to manage the data catalog Update the AWS Glue ETL code to include the enableUpdateCatalog and partitionKeys arguments.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 67
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 Enterprise edition configured to perform identity federation using SAML 2.0 and the default encryption settings.
B. Amazon QuickSight Standard edition configured to perform identity federation using SAML 2.0. and the default encryption settings.
C. Amazon QuickSight Enterprise edition using AD Connector to authenticate using Active Directory. Configure Amazon QuickSight to use customer-provided keys imported into AWS KMS.
D. Amazon QuckSight Standard edition using AD Connector to authenticate using Active Directory. Configure Amazon QuickSight to use customer-provided keys imported into AWS KMS.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 68
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 (LEAVE A REPLY)
Explanation
Matching Records with AWS Lake Formation FindMatches

NEW QUESTION: 69
A large company has a central data lake to run analytics across different departments. Each department uses a separate AWS account and stores its data in an Amazon S3 bucket in that account. Each AWS account uses the AWS Glue Data Catalog as its data catalog. There are different data lake access requirements based on roles. Associate analysts should only have read access to their departmental dat a. Senior data analysts can have access in multiple departments including theirs, but for a subset of columns only.
Which solution achieves these required access patterns to minimize costs and administrative tasks?
A. Consolidate all AWS accounts into one account. Create different S3 buckets for each department and move all the data from every account to the central data lake account. Migrate the individual data

catalogs into a central data catalog and apply fine-grained permissions to give to each user the required access to tables and databases in AWS Glue and Amazon S3.
B. Keep the account structure and the individual AWS Glue catalogs on each account. Add a central data lake account and use AWS Glue to catalog data from various accounts. Configure cross-account access for AWS Glue crawlers to scan the data in each departmental S3 bucket to identify the schema and populate the catalog. Add the senior data analysts into the central account and apply highly detailed access controls in the Data Catalog and Amazon S3.
C. Set up an individual AWS account for the central data lake. Use AWS Lake Formation to catalog the cross- account locations. On each individual S3 bucket, modify the bucket policy to grant S3 permissions to the Lake Formation service-linked role. Use Lake Formation permissions to add fine-grained access controls to allow senior analysts to view specific tables and columns.
D. Set up an individual AWS account for the central data lake and configure a central S3 bucket. Use an AWS Lake Formation blueprint to move the data from the various buckets into the central S3 bucket. On each individual bucket, modify the bucket policy to grant S3 permissions to the Lake Formation service-linked role. Use Lake Formation permissions to add fine-grained access controls for both associate and senior analysts to view specific tables and columns.
Answer: C (LEAVE A REPLY)
Lake Formation provides secure and granular access to data through a new grant/revoke permissions model that augments AWS Identity and Access Management (IAM) policies. Analysts and data scientists can use the full portfolio of AWS analytics and machine learning services, such as Amazon Athena, to access the data. The configured Lake Formation security policies help ensure that users can access only the data that they are authorized to access. Source : https://docs.aws.amazon.com/lake-formation/latest/dg/how-it-works.html

NEW QUESTION: 70
A mortgage company has a microservice for accepting payments. This microservice uses the Amazon DynamoDB encryption client with AWS KMS managed keys to encrypt the sensitive data before writing the data to DynamoDB. The finance team should be able to load this data into Amazon Redshift and aggregate the values within the sensitive fields. The Amazon Redshift cluster is shared with other data analysts from different business units.
Which steps should a data analyst take to accomplish this task efficiently and securely?
A. Create an Amazon EMR cluster. Create Apache Hive tables that reference the data stored in DynamoDB. Insert the output to the restricted Amazon S3 bucket for the finance team. Use the COPY command with the IAM role that has access to the KMS key to load the data from Amazon S3 to the finance table in Amazon Redshift.
B. Create an Amazon EMR cluster with an EMR_EC2_DefaultRole role that has access to the KMS key.
Create Apache Hive tables that reference the data stored in DynamoDB and the finance table in Amazon Redshift. In Hive, select the data from DynamoDB and then insert the output to the finance table in Amazon Redshift.

C. Create an AWS Lambda function to process the DynamoDB stream. Save the output to a restricted S3 bucket for the finance team. Create a finance table in Amazon Redshift that is accessible to the finance team only. Use the COPY command with the IAM role that has access to the KMS key to load the data from S3 to the finance table.
D. Create an AWS Lambda function to process the DynamoDB stream. Decrypt the sensitive data using the same KMS key. Save the output to a restricted S3 bucket for the finance team. Create a finance table in Amazon Redshift that is accessible to the finance team only. Use the COPY command to load the data from Amazon S3 to the finance table.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 71
A regional energy company collects voltage data from sensors attached to buildings. To address any known dangerous conditions, the company wants to be alerted when a sequence of two voltage drops is detected within 10 minutes of a voltage spike at the same building. It is important to ensure that all messages are delivered as quickly as possible. The system must be fully managed and highly available. The company also needs a solution that will automatically scale up as it covers additional cites with this monitoring feature. The alerting system is subscribed to an Amazon SNS topic for remediation.
Which solution meets these requirements?
A. Create a REST-based web service using Amazon API Gateway in front of an AWS Lambda function. Create an Amazon RDS for PostgreSQL database with sufficient Provisioned IOPS (PIOPS). In the Lambda function, store incoming events in the RDS database and query the latest data to detect the known event sequence and send the SNS message.
B. Create an Amazon Kinesis data stream to capture the incoming sensor data and create another stream for alert messages. Set up AWS Application Auto Scaling on both. Create a Kinesis Data Analytics for Java application to detect the known event sequence, and add a message to the message stream. Configure an AWS Lambda function to poll the message stream and publish to the SNS topic.
C. Create an Amazon Kinesis Data Firehose delivery stream to capture the incoming sensor data. Use an AWS Lambda transformation function to detect the known event sequence and send the SNS message.
D. Create an Amazon Managed Streaming for Kafka cluster to ingest the data, and use an Apache Spark Streaming with Apache Kafka consumer API in an automatically scaled Amazon EMR cluster to process the incoming data. Use the Spark Streaming application to detect the known event sequence and send the SNS message.
Answer: (SHOW ANSWER)

NEW QUESTION: 72
A large company has a central data lake to run analytics across different departments. Each department uses a separate AWS account and stores its data in an Amazon S3 bucket in that account. Each AWS account uses the AWS Glue Data Catalog as its data catalog. There are different data lake access requirements based on roles. Associate analysts should only have read access to their

departmental data. Senior data analysts can have access in multiple departments including theirs, but for a subset of columns only.
Which solution achieves these required access patterns to minimize costs and administrative tasks?
A. Consolidate all AWS accounts into one account. Create different S3 buckets for each department and move all the data from every account to the central data lake account. Migrate the individual data catalogs into a central data catalog and apply fine-grained permissions to give to each user the required access to tables and databases in AWS Glue and Amazon S3.
B. Keep the account structure and the individual AWS Glue catalogs on each account. Add a central data lake account and use AWS Glue to catalog data from various accounts. Configure cross-account access for AWS Glue crawlers to scan the data in each departmental S3 bucket to identify the schema and populate the catalog. Add the senior data analysts into the central account and apply highly detailed access controls in the Data Catalog and Amazon S3.
C. Set up an individual AWS account for the central data lake. Use AWS Lake Formation to catalog the cross- account locations. On each individual S3 bucket, modify the bucket policy to grant S3 permissions to the Lake Formation service-linked role. Use Lake Formation permissions to add fine-grained access controls to allow senior analysts to view specific tables and columns.
D. Set up an individual AWS account for the central data lake and configure a central S3 bucket. Use an AWS Lake Formation blueprint to move the data from the various buckets into the central S3 bucket. On each individual bucket, modify the bucket policy to grant S3 permissions to the Lake Formation service-linked role. Use Lake Formation permissions to add fine-grained access controls for both associate and senior analysts to view specific tables and columns.
Answer: C (LEAVE A REPLY)
Explanation
Lake Formation provides secure and granular access to data through a new grant/revoke permissions model that augments AWS Identity and Access Management (IAM) policies. Analysts and data scientists can use the full portfolio of AWS analytics and machine learning services, such as Amazon Athena, to access the data.
The configured Lake Formation security policies help ensure that users can access only the data that they are authorized to access. Source : https://docs.aws.amazon.com/lake-formation/latest/dg/how-it-works.html

NEW QUESTION: 73
An education provider's learning management system (LMS) is hosted in a 100 TB data lake that is built on Amazon S3. The provider's LMS supports hundreds of schools. The provider wants to build an advanced analytics reporting platform using Amazon Redshift to handle complex queries with optimal performance.
System users will query the most recent 4 months of data 95% of the time while 5% of the queries will leverage data from the previous 12 months.
Which solution meets these requirements in the MOST cost-effective way?

A. Store the most recent 4 months of data in the Amazon Redshift cluster. Use Amazon Redshift Spectrum to query data in the data lake. Use S3 lifecycle management rules to store data from the previous 12 months in Amazon S3 Glacier storage.
B. Leverage DS2 nodes for the Amazon Redshift cluster. Migrate all data from Amazon S3 to Amazon Redshift. Decommission the data lake.
C. Store the most recent 4 months of data in the Amazon Redshift cluster. Use Amazon Redshift federated queries to join cluster data with the data lake to reduce costs. Ensure the S3 Standard storage class is in use with objects in the data lake.
D. Store the most recent 4 months of data in the Amazon Redshift cluster. Use Amazon Redshift Spectrum to query data in the data lake. Ensure the S3 Standard storage class is in use with objects in the data lake.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 74
A team of data scientists plans to analyze market trend data for their company's new investment strategy. The trend data comes from five different data sources in large volumes. The team wants to utilize Amazon Kinesis to support their use case. The team uses SQL-like queries to analyze trends and wants to send notifications based on certain significant patterns in the trends. Additionally, the data scientists want to save the data to Amazon S3 for archival and historical re-processing, and use AWS managed services wherever possible. The team wants to implement the lowest-cost solution.
Which solution meets these requirements?
A. Publish data to one Kinesis data stream. Deploy a custom application using the Kinesis Client Library (KCL) for analyzing trends, and send notifications using Amazon SNS. Configure Kinesis Data Firehose on the Kinesis data stream to persist data to an S3 bucket.
B. Publish data to two Kinesis data streams. Deploy a custom application using the Kinesis Client Library (KCL) to the first stream for analyzing trends, and send notifications using Amazon SNS. Configure Kinesis Data Firehose on the second Kinesis data stream to persist data to an S3 bucket.
C. Publish data to two Kinesis data streams. Deploy Kinesis Data Analytics to the first stream for analyzing trends, and configure an AWS Lambda function as an output to send notifications using Amazon SNS.
Configure Kinesis Data Firehose on the second Kinesis data stream to persist data to an S3 bucket.
D. Publish data to one Kinesis data stream. Deploy Kinesis Data Analytic to the stream for analyzing trends, and configure an AWS Lambda function as an output to send notifications using Amazon SNS. Configure Kinesis Data Firehose on the Kinesis data stream to persist data to an S3 bucket.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 75
A hospital uses wearable medical sensor devices to collect data from patients. The hospital is architecting a near-real-time solution that can ingest the data securely at scale. The solution should also be able to remove the patient's protected health information (PHI) from the streaming data and store the data in durable storage.

Which solution meets these requirements with the least operational overhead?
A. Ingest the data using Amazon Kinesis Data Streams, which invokes an AWS Lambda function using Kinesis Client Library (KCL) to remove all PHI. Write the data in Amazon S3.
B. Ingest the data using Amazon Kinesis Data Firehose to write the data to Amazon S3. Have Amazon S3 trigger an AWS Lambda function that parses the sensor data to remove all PHI in Amazon S3.
C. Ingest the data using Amazon Kinesis Data Streams to write the data to Amazon S3. Have the data stream launch an AWS Lambda function that parses the sensor data and removes all PHI in Amazon S3.
D. Ingest the data using Amazon Kinesis Data Firehose to write the data to Amazon S3. Implement a transformation AWS Lambda function that parses the sensor data to remove all PHI.
Answer: (SHOW ANSWER)
Explanation
https://aws.amazon.com/blogs/big-data/persist-streaming-data-to-amazon-s3-using-amazon-kinesis-firehose-and-

NEW QUESTION: 76
An online retail company with millions of users around the globe wants to improve its ecommerce analytics capabilities. Currently, clickstream data is uploaded directly to Amazon S3 as compressed files. Several times each day, an application running on Amazon EC2 processes the data and makes search options and reports available for visualization by editors and marketers. The company wants to make website clicks and aggregated data available to editors and marketers in minutes to enable them to connect with users more effectively.
Which options will help meet these requirements in the MOST efficient way? (Choose two.)
A. Use Amazon Kinesis Data Firehose to upload compressed and batched clickstream records to Amazon Elasticsearch Service.
B. Upload clickstream records to Amazon S3 as compressed files. Then use AWS Lambda to send data to Amazon Elasticsearch Service from Amazon S3.
C. Use Kibana to aggregate, filter, and visualize the data stored in Amazon Elasticsearch Service. Refresh content performance dashboards in near-real time.
D. Use Amazon Elasticsearch Service deployed on Amazon EC2 to aggregate, filter, and process the data.
Refresh content performance dashboards in near-real time.
E. Upload clickstream records from Amazon S3 to Amazon Kinesis Data Streams and use a Kinesis Data Streams consumer to send records to Amazon Elasticsearch Service.
Answer: D,E (LEAVE A REPLY)

NEW QUESTION: 77
A global company has different sub-organizations, and each sub-organization sells its products and services in various countries. The company's senior leadership wants to quickly identify which sub-organization is the strongest performer in each country. All sales data is stored in Amazon S3 in Parquet format.
Which approach can provide the visuals that senior leadership requested with the least amount of effort?
A. Use Amazon QuickSight with Amazon Athena as the data source. Use pivot tables as the visual type.
B. Use Amazon QuickSight with Amazon Athena as the data source. Use heat maps as the visual type.
C. Use Amazon QuickSight with Amazon S3 as the data source. Use heat maps as the visual type.
D. Use Amazon QuickSight with Amazon S3 as the data source. Use pivot tables as the visual type.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 78
A data engineering team within a shared workspace company wants to build a centralized logging system for all weblogs generated by the space reservation system. The company has a fleet of Amazon EC2 instances that process requests for shared space reservations on its website. The data engineering team wants to ingest all weblogs into a service that will provide a near-real-time search engine. The team does not want to manage the maintenance and operation of the logging system.
Which solution allows the data engineering team to efficiently set up the web logging system within AWS?
A. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis data stream to CloudWatch. Choose Amazon Elasticsearch Service as the end destination of the weblogs.
B. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis Data Firehose delivery stream to CloudWatch. Choose Amazon Elasticsearch Service as the end destination of the weblogs.
C. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis data stream to CloudWatch. Configure Splunk as the end destination of the weblogs.
D. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis Firehose delivery stream to CloudWatch. Configure Amazon DynamoDB as the end destination of the weblogs.
Answer: (SHOW ANSWER)
Explanation
https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/CWL_ES_Stream.html

NEW QUESTION: 79

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

NEW QUESTION: 80
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 delete the processed objects or data from Amazon S3 after each run.
B. Have the ETL jobs read the data from Amazon S3 using a DataFrame.
C. Create custom logic on the ETL jobs to track the processed S3 objects.
D. Enable job bookmarks on the AWS Glue jobs.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 81
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. 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.

B. 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.
C. Create a VPC endpoint from the Amazon QuickSight VPC to the Amazon Redshift VPC so Amazon QuickSight can access data from 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: (SHOW ANSWER)

NEW QUESTION: 82
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 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
B. Use Amazon Kinesis Data Analytics to create an in-application table based upon the reference data Process incoming data from Kinesis Data Streams Use an Amazon Kinesis Data Firehose delivery stream to write results to Amazon Redshift
C. 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
D. 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
Answer: B (LEAVE A REPLY)

NEW QUESTION: 83
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 Drivers and ingest the entire dataset. Use appropriate Apache Spark libraries to compare the dataset, and find the delta.
B. Use AWS Glue to connect to the data source using JDBC Drivers. Ingest incremental records only using job bookmarks.
C. Use AWS Glue to connect to the data source using JDBC Drivers and ingest the full data. Use AWS DataSync to ensure the delta only is written into Amazon S3.
D. Use AWS Glue to connect to the data source using JDBC Drivers. Store the last updated key in an Amazon DynamoDB table and ingest the data using the updated key as a filter.

Answer: D (LEAVE A REPLY)

NEW QUESTION: 84
A hospital uses wearable medical sensor devices to collect data from patients. The hospital is architecting a near-real-time solution that can ingest the data securely at scale. The solution should also be able to remove the patient's protected health information (PHI) from the streaming data and store the data in durable storage.
Which solution meets these requirements with the least operational overhead?
A. Ingest the data using Amazon Kinesis Data Streams, which invokes an AWS Lambda function using Kinesis Client Library (KCL) to remove all PHI. Write the data in Amazon S3.
B. Ingest the data using Amazon Kinesis Data Firehose to write the data to Amazon S3. Have Amazon S3 trigger an AWS Lambda function that parses the sensor data to remove all PHI in Amazon S3.
C. Ingest the data using Amazon Kinesis Data Streams to write the data to Amazon S3. Have the data stream launch an AWS Lambda function that parses the sensor data and removes all PHI in Amazon S3.
D. Ingest the data using Amazon Kinesis Data Firehose to write the data to Amazon S3. Implement a transformation AWS Lambda function that parses the sensor data to remove all PHI.
Answer: (SHOW ANSWER)
https://aws.amazon.com/blogs/big-data/persist-streaming-data-to-amazon-s3-using-amazon-kinesis-firehose-and-aws-lambda/)

NEW QUESTION: 85
A company uses Amazon Elasticsearch Service (Amazon ES) to store and analyze its website clickstream dat a. The company ingests 1 TB of data daily using Amazon Kinesis Data Firehose and stores one day's worth of data in an Amazon ES cluster.
The company has very slow query performance on the Amazon ES index and occasionally sees errors from Kinesis Data Firehose when attempting to write to the index. The Amazon ES cluster has 10 nodes running a single index and 3 dedicated master nodes. Each data node has 1.5 TB of Amazon EBS storage attached and the cluster is configured with 1,000 shards. Occasionally, JVMMemoryPressure errors are found in the cluster logs.
Which solution will improve the performance of Amazon ES?
A. Increase the memory of the Amazon ES master nodes.
B. Decrease the number of Amazon ES data nodes.
C. Decrease the number of Amazon ES shards for the index.
D. Increase the number of Amazon ES shards for the index.
Answer: C (LEAVE A REPLY)
https://aws.amazon.com/premiumsupport/knowledge-center/high-jvm-memory-pressure-elasticsearch/

NEW QUESTION: 86
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 Glue. Perform the join with AWS Glue ETL scripts.
C. Create an external table using Amazon Redshift Spectrum for the call center data and perform the join with Amazon Redshift.
D. Export the call center data from Amazon Redshift to Amazon EMR using Apache Sqoop. Perform the join with Apache Hive.
Answer: C (LEAVE A REPLY)
https://docs.aws.amazon.com/redshift/latest/dg/c-spectrum-external-tables.html

NEW QUESTION: 87
An online retail company with millions of users around the globe wants to improve its ecommerce analytics capabilities. Currently, clickstream data is uploaded directly to Amazon S3 as compressed files. Several times each day, an application running on Amazon EC2 processes the data and makes search options and reports available for visualization by editors and marketers. The company wants to make website clicks and aggregated data available to editors and marketers in minutes to enable them to connect with users more effectively.
Which options will help meet these requirements in the MOST efficient way? (Choose two.)
A. Use Kibana to aggregate, filter, and visualize the data stored in Amazon Elasticsearch Service. Refresh content performance dashboards in near-real time.
B. Use Amazon Elasticsearch Service deployed on Amazon EC2 to aggregate, filter, and process the data. Refresh content performance dashboards in near-real time.
C. Upload clickstream records from Amazon S3 to Amazon Kinesis Data Streams and use a Kinesis Data Streams consumer to send records to Amazon Elasticsearch Service.
D. Upload clickstream records to Amazon S3 as compressed files. Then use AWS Lambda to send data to Amazon Elasticsearch Service from Amazon S3.
E. Use Amazon Kinesis Data Firehose to upload compressed and batched clickstream records to Amazon Elasticsearch Service.
Answer: A,E (LEAVE A REPLY)

NEW QUESTION: 88
A company uses Amazon Elasticsearch Service (Amazon ES) to store and analyze its website clickstream data. The company ingests 1 TB of data daily using Amazon Kinesis Data Firehose and stores one day's worth of data in an Amazon ES cluster.
The company has very slow query performance on the Amazon ES index and occasionally sees errors from Kinesis Data Firehose when attempting to write to the index. The Amazon ES cluster has 10

nodes running a single index and 3 dedicated master nodes. Each data node has 1.5 TB of Amazon EBS storage attached and the cluster is configured with 1,000 shards. Occasionally, JVMMemoryPressure errors are found in the cluster logs.
Which solution will improve the performance of Amazon ES?
A. Increase the memory of the Amazon ES master nodes.
B. Decrease the number of Amazon ES data nodes.
C. Decrease the number of Amazon ES shards for the index.
D. Increase the number of Amazon ES shards for the index.
Answer: C (LEAVE A REPLY)
Explanation
https://aws.amazon.com/premiumsupport/knowledge-center/high-jvm-memory-pressure-elasticsearch/

NEW QUESTION: 89
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: (SHOW ANSWER)
Explanation
https://docs.aws.amazon.com/quicksight/latest/user/directory-integration.html

NEW QUESTION: 90
A hospital uses wearable medical sensor devices to collect data from patients. The hospital is architecting a near-real-time solution that can ingest the data securely at scale. The solution should also be able to remove the patient's protected health information (PHI) from the streaming data and store the data in durable storage.
Which solution meets these requirements with the least operational overhead?
A. Ingest the data using Amazon Kinesis Data Firehose to write the data to Amazon S3. Implement a transformation AWS Lambda function that parses the sensor data to remove all PHI.
B. Ingest the data using Amazon Kinesis Data Streams, which invokes an AWS Lambda function using Kinesis Client Library (KCL) to remove all PHI. Write the data in Amazon S3.
C. Ingest the data using Amazon Kinesis Data Firehose to write the data to Amazon S3. Have Amazon S3 trigger an AWS Lambda function that parses the sensor data to remove all PHI in Amazon S3.

D. Ingest the data using Amazon Kinesis Data Streams to write the data to Amazon S3. Have the data stream launch an AWS Lambda function that parses the sensor data and removes all PHI in Amazon S3.

Answer: D (LEAVE A REPLY)

NEW QUESTION: 91

A company wants to improve user satisfaction for its smart home system by adding more features to its recommendation engine. Each sensor asynchronously pushes its nested JSON data into Amazon Kinesis Data Streams using the Kinesis Producer Library (KPL) in Java. Statistics from a set of failed sensors showed that, when a sensor is malfunctioning, its recorded data is not always sent to the cloud. The company needs a solution that offers near-real-time analytics on the data from the most updated sensors.

Which solution enables the company to meet these requirements?

A. Update the sensors code to use the PutRecord/PutRecords call from the Kinesis Data Streams API with the AWS SDK for Java. Use Kinesis Data Analytics to enrich the data based on a company-developed anomaly detection SQL script. Direct the output of KDA application to a Kinesis Data Firehose delivery stream, enable the data transformation feature to flatten the JSON file, and set the Kinesis Data Firehose destination to an Amazon Elasticsearch Service cluster.

B. Update the sensors code to use the PutRecord/PutRecords call from the Kinesis Data Streams API with the AWS SDK for Java. Use AWS Glue to fetch and process data from the stream using the Kinesis Client Library (KCL). Instantiate an Amazon Elasticsearch Service cluster and use AWS Lambda to directly push data into it.

C. Set the RecordMaxBufferedTime property of the KPL to "0" to disable the buffering on the sensor side.

Connect for each stream a dedicated Kinesis Data Firehose delivery stream and enable the data transformation feature to flatten the JSON file before sending it to an Amazon S3 bucket. Load the S3 data into an Amazon Redshift cluster.

D. Set the RecordMaxBufferedTime property of the KPL to "-1" to disable the buffering on the sensor side.

Use Kinesis Data Analytics to enrich the data based on a company-developed anomaly detection SQL script. Push the enriched data to a fleet of Kinesis data streams and enable the data transformation feature to flatten the JSON file. Instantiate a dense storage Amazon Redshift cluster and use it as the destination for the Kinesis Data Firehose delivery stream.

Answer: D (LEAVE A REPLY)

NEW QUESTION: 92
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 daily roll-up data initially in the Amazon S3 Standard storage class. Apply a lifecycle configuration that changes the storage class to Amazon S3 Standard-Infrequent Access (S3 Standard-IA)
1 year after
data creation.
B. Store the source data initially in the Amazon S3 Glacier storage class. 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.
C. Store the daily roll-up data initially in the Amazon S3 Standard storage class. Apply a lifecycle configuration that changes the storage class to Amazon S3 Glacier Deep Archive 1 year after data creation.
D. Store the source data initially in the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class. 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.
E. Store the daily roll-up data initially in the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class. Apply a lifecycle configuration that changes the storage class to Amazon S3 Glacier 1 year after data creation.
Answer: A,D (LEAVE A REPLY)

NEW QUESTION: 93
A global pharmaceutical company receives test results for new drugs from various testing facilities worldwide.
The results are sent in millions of 1 KB-sized JSON objects to an Amazon S3 bucket owned by the company.
The data engineering team needs to process those files, convert them into Apache Parquet format, and load them into Amazon Redshift for data analysts to perform dashboard reporting. The engineering team uses AWS Glue to process the objects, AWS Step Functions for process orchestration, and Amazon CloudWatch for job scheduling.

More testing facilities were recently added, and the time to process files is increasing.
What will MOST efficiently decrease the data processing time?
A. Use Amazon EMR instead of AWS Glue to group the small input files. Process the files in Amazon EMR and load them into Amazon Redshift tables.
B. Use AWS Lambda to group the small files into larger files. Write the files back to Amazon S3. Process the files using AWS Glue and load them into Amazon Redshift tables.
C. Use the Amazon Redshift COPY command to move the files from Amazon S3 into Amazon Redshift tables directly. Process the files in Amazon Redshift.
D. Use the AWS Glue dynamic frame file grouping option while ingesting the raw input files. Process the files and load them into Amazon Redshift tables.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 94
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 Lambda consumers from standard data stream iterators to an HTTP/2 stream consumer.
B. Configure multiple Lambda functions to process the stream.
C. Migrate the validation process to Amazon Kinesis Data Firehose.
D. Increase the number of shards in the stream.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 95
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. Place Amazon QuickSight and Amazon Redshift in the security group and use an Amazon S3 endpoint to connect Amazon QuickSight to Amazon S3.
B. Use a VPC endpoint to connect to Amazon S3 from Amazon QuickSight and an IAM role to authenticate Amazon Redshift.
C. Use an Active Directory connector and single sign-on (SSO) in a corporate network environment.
D. Establish a secure connection by creating an S3 endpoint to connect Amazon QuickSight and a VPC endpoint to connect to Amazon Redshift.
Answer: (SHOW ANSWER)

NEW QUESTION: 96
A company uses Amazon Redshift as its data warehouse. A new table has columns that contain sensitive dat a. The data in the table will eventually be referenced by several existing queries that run many times a day.
A data analyst needs to load 100 billion rows of data into the new table. Before doing so, the data analyst must ensure that only members of the auditing group can read the columns containing sensitive data.
How can the data analyst meet these requirements with the lowest maintenance overhead?
A. Load all the data into the new table and grant the auditing group permission to read from the table. Load all the data except for the columns containing sensitive data into a second table. Grant the appropriate users read-only permissions to the second table.
B. Load all the data into the new table and grant the auditing group permission to read from the table. Use the GRANT SQL command to allow read-only access to a subset of columns to the appropriate users.
C. 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.
D. 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: (SHOW ANSWER)
https://aws.amazon.com/blogs/big-data/achieve-finer-grained-data-security-with-column-level-access-control-in-amazon-redshift/

NEW QUESTION: 97
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 archived data, use Amazon EMR to perform data transformations.
B. For archived data, use Amazon SageMaker to perform data transformations.
C. For daily incoming data, use Amazon Athena to scan and identify the schema.
D. For daily incoming data, use AWS Glue workflows with AWS Glue jobs to perform transformations.
E. For daily incoming data, use Amazon Redshift to perform transformations.
F. For daily incoming data, use AWS Glue crawlers to scan and identify the schema.
Answer: A,D,F (LEAVE A REPLY)

NEW QUESTION: 98
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. Add more partitions to be used over the table. Then filter over two partitions and put all columns in the WHERE clause
B. Convert the Parquet files to the Apache ORC file format. Then attempt to query the data again
C. Use partition projection to speed up the processing of the partitioned table
D. Convert the Parquet files to the csv file format..Then attempt to query the data again
Answer: C (LEAVE A REPLY)

NEW QUESTION: 99
A data analyst is using Amazon QuickSight for data visualization across multiple datasets generated by applications. Each application stores files within a separate Amazon S3 bucket. AWS Glue Data Catalog is used as a central catalog across all application data in Amazon S3. A new application stores its data within a separate S3 bucket. After updating the catalog to include the new application data source, the data analyst created a new Amazon QuickSight data source from an Amazon Athena table, but the import into SPICE failed.
How should the data analyst resolve the issue?
A. Edit the permissions for the AWS Glue Data Catalog from within the AWS Glue console.
B. Edit the permissions for the AWS Glue Data Catalog from within the Amazon QuickSight console.
C. Edit the permissions for the new S3 bucket from within the Amazon QuickSight console.
D. Edit the permissions for the new S3 bucket from within the S3 console.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 100
A mortgage company has a microservice for accepting payments. This microservice uses the Amazon DynamoDB encryption client with AWS KMS managed keys to encrypt the sensitive data before writing the data to DynamoDB. The finance team should be able to load this data into Amazon Redshift and aggregate the values within the sensitive fields. The Amazon Redshift cluster is shared with other data analysts from different business units.
Which steps should a data analyst take to accomplish this task efficiently and securely?
A. Create an Amazon EMR cluster with an EMR_EC2_DefaultRole role that has access to the KMS key.
Create Apache Hive tables that reference the data stored in DynamoDB and the finance table in Amazon Redshift. In Hive, select the data from DynamoDB and then insert the output to the finance table in Amazon Redshift.

B. Create an AWS Lambda function to process the DynamoDB stream. Decrypt the sensitive data using the same KMS key. Save the output to a restricted S3 bucket for the finance team. Create a finance table in Amazon Redshift that is accessible to the finance team only. Use the COPY command to load the data from Amazon S3 to the finance table.
C. Create an AWS Lambda function to process the DynamoDB stream. Save the output to a restricted S3 bucket for the finance team. Create a finance table in Amazon Redshift that is accessible to the finance team only. Use the COPY command with the IAM role that has access to the KMS key to load the data from S3 to the finance table.
D. Create an Amazon EMR cluster. Create Apache Hive tables that reference the data stored in DynamoDB. Insert the output to the restricted Amazon S3 bucket for the finance team. Use the COPY command with the IAM role that has access to the KMS key to load the data from Amazon S3 to the finance table in Amazon Redshift.
Answer: (SHOW ANSWER)

NEW QUESTION: 101
A data engineering team within a shared workspace company wants to build a centralized logging system for all weblogs generated by the space reservation system. The company has a fleet of Amazon EC2 instances that process requests for shared space reservations on its website. The data engineering team wants to ingest all weblogs into a service that will provide a near-real-time search engine. The team does not want to manage the maintenance and operation of the logging system.
Which solution allows the data engineering team to efficiently set up the web logging system within AWS?
A. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis data stream to CloudWatch. Choose Amazon Elasticsearch Service as the end destination of the weblogs.
B. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis Data Firehose delivery stream to CloudWatch. Choose Amazon Elasticsearch Service as the end destination of the weblogs.
C. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis data stream to CloudWatch. Configure Splunk as the end destination of the weblogs.
D. Set up the Amazon CloudWatch agent to stream weblogs to CloudWatch logs and subscribe the Amazon Kinesis Firehose delivery stream to CloudWatch. Configure Amazon DynamoDB as the end destination of the weblogs.
Answer: B (LEAVE A REPLY)
https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/CWL_ES_Stream.html

NEW QUESTION: 102
A large financial company is running its ETL process. Part of this process is to move data from Amazon S3 into an Amazon Redshift cluster. The company wants to use the most cost-efficient method to load the dataset into Amazon Redshift.
Which combination of steps would meet these requirements? (Choose two.)

A. Use temporary staging tables during the loading process.
B. Use Amazon Redshift Spectrum to query files from Amazon S3.
C. Use S3DistCp to load files into Amazon Redshift.
D. Use the UNLOAD command to upload data into Amazon Redshift.
E. Use the COPY command with the manifest file to load data into Amazon Redshift.
Answer: A,E (LEAVE A REPLY)

NEW QUESTION: 103
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. Use an S3 Select query to ensure that the data is properly updated. Create an AWS Glue Data Catalog to manage the Hive metadata over the S3 Select table. Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
B. Create an AWS Glue Data Catalog to manage the Hive metadata. Create an Amazon EMR cluster with consistent view enabled. Run emrfs sync before each analytics step to ensure data changes are updated. Create an EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
C. Create an AWS Glue Data Catalog to manage the Hive metadata. Create an AWS Glue crawler over Amazon S3 that runs when data is refreshed to ensure that data changes are updated. 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 Amazon Athena table with CREATE TABLE AS SELECT (CTAS) to ensure data is refreshed from underlying queries against the raw dataset. Create an AWS Glue Data Catalog to manage the Hive metadata over the CTAS table. Create an Amazon EMR cluster and use the metadata in the AWS Glue Data Catalog to run Hive processing queries in Amazon EMR.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 104
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. Create a VPC endpoint from the Amazon QuickSight VPC to the Amazon Redshift VPC so Amazon QuickSight can access data from Amazon Redshift.
B. 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.
C. 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.
D. 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.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 105
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. Take a snapshot of the Amazon Redshift cluster. Restore the cluster to a new cluster using dense storage nodes with additional storage capacity.
B. 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.
C. Unload all the tables in Amazon Redshift to an Amazon S3 bucket using S3 Intelligent-Tiering. 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.
D. Create a daily job in AWS Glue to UNLOAD records older than 13 months to Amazon S3 and delete those records from Amazon Redshift. Create an external table in Amazon Redshift to point to the S3 location. Use Amazon Redshift Spectrum to join to data that is older than 13 months.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 106
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 (LEAVE A REPLY)
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: 107
A company has 1 million scanned documents stored as image files in Amazon S3. The documents contain typewritten application forms with information including the applicant first name, applicant last name, application date, application type, and application text. The company has developed a machine learning algorithm to extract the metadata values from the scanned documents. The company wants to allow internal data analysts to analyze and find applications using the applicant name, application date, or application text.
The original images should also be downloadable. Cost control is secondary to query performance.
Which solution organizes the images and metadata to drive insights while meeting the requirements?
A. Store the metadata and the Amazon S3 location of the image file in an Amazon Redshift table. Allow the data analysts to run ad-hoc queries on the table.
B. For each image, use object tags to add the metadata. Use Amazon S3 Select to retrieve the files based on the applicant name and application date.
C. Index the metadata and the Amazon S3 location of the image file in Amazon Elasticsearch Service. Allow the data analysts to use Kibana to submit queries to the Elasticsearch cluster.

D. Store the metadata and the Amazon S3 location of the image files in an Apache Parquet file in Amazon S3, and define a table in the AWS Glue Data Catalog. Allow data analysts to use Amazon Athena to submit custom queries.
Answer: (SHOW ANSWER)

NEW QUESTION: 108
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 (LEAVE A REPLY)
Matching Records with AWS Lake Formation FindMatches

NEW QUESTION: 109
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. Job bookmarks
B. Triggers
C. Workflows
D. Classifiers
Answer: A (LEAVE A REPLY)