Amazon-Web-Services

Exam Questions DAS-C01

AWS Certified Data Analytics - Specialty

NEW QUESTION 1
A market data company aggregates external data sources to create a detailed view of product consumption in different countries. The company wants to sell this data to external parties through a subscription. To achieve this goal, the company needs to make its data securely available to external parties who are also AWS users.
What should the company do to meet these requirements with the LEAST operational overhead?

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

Answer: A

NEW QUESTION 2
A company has several Amazon EC2 instances sitting behind an Application Load Balancer (ALB) The company wants its IT Infrastructure team to analyze the IP addresses coming into the company's ALB The ALB is configured to store access logs in Amazon S3 The access logs create about 1 TB of data each day, and access to the data will be infrequent The company needs a solution that is scalable, cost-effective and has minimal maintenance requirements
Which solution meets these requirements?

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

Answer: D

NEW QUESTION 3
Three teams of data analysts use Apache Hive on an Amazon EMR cluster with the EMR File System (EMRFS) to query data stored within each teams Amazon S3 bucket. The EMR cluster has Kerberos enabled and is configured to authenticate users from the corporate Active Directory. The data is highly sensitive, so access must be limited to the members of each team.
Which steps will satisfy the security requirements?

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

Answer: C

NEW QUESTION 4
A retail company has 15 stores across 6 cities in the United States. Once a month, the sales team requests a visualization in Amazon QuickSight that provides the ability to easily identify revenue trends across cities and stores. The visualization also helps identify outliers that need to be examined with further analysis.
Which visual type in QuickSight meets the sales team's requirements?

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

Answer: A

NEW QUESTION 5
A media company is using Amazon QuickSight dashboards to visualize its national sales data. The dashboard is using a dataset with these fields: ID, date, time_zone, city, state, country, longitude, latitude, sales_volume, and number_of_items.
To modify ongoing campaigns, the company wants an interactive and intuitive visualization of which states across the country recorded a significantly lower sales volume compared to the national average.
Which addition to the company’s QuickSight dashboard will meet this requirement?

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

Answer: B

NEW QUESTION 6
A company is hosting an enterprise reporting solution with Amazon Redshift. The application provides reporting capabilities to three main groups: an executive group to access financial reports, a data analyst group to run long-running ad-hoc queries, and a data engineering group to run stored procedures and ETL processes. The executive team requires queries to run with optimal performance. The data engineering team expects queries to take minutes.
Which Amazon Redshift feature meets the requirements for this task?

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

Answer: D

Explanation:
Materialized views:

NEW QUESTION 7
A company with a video streaming website wants to analyze user behavior to make recommendations to users in real time Clickstream data is being sent to Amazon Kinesis Data Streams and reference data is stored in Amazon S3 The company wants a solution that can use standard SQL quenes The solution must also provide a way to look up pre-calculated reference data while making recommendations
Which solution meets these requirements?

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

Answer: D

NEW QUESTION 8
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 tabl
B. Add SQL commands to replace the existing rows in the main table as postactions in the DynamicFrameWriter class.
C. Load the previously inserted data into a MySQL database in the AWS Glue jo
D. Perform an upsert operation in MySQL, and copy the results to the Amazon Redshift table.
E. Use Apache Spark's DataFrame dropDuplicates() API to eliminate duplicates and then write the data to Amazon Redshift.
F. Use the AWS Glue ResolveChoice built-in transform to select the most recent value of the column.

Answer: A

Explanation:
https://aws.amazon.com/premiumsupport/knowledge-center/sql-commands-redshift-glue-job/ See the section Merge an Amazon Redshift table in AWS Glue (upsert)

NEW QUESTION 9
A company wants to provide its data analysts with uninterrupted access to the data in its Amazon Redshift cluster. All data is streamed to an Amazon S3 bucket with Amazon Kinesis Data Firehose. An AWS Glue job that is scheduled to run every 5 minutes issues a COPY command to move the data into Amazon Redshift. The amount of data delivered is uneven throughout the day, and cluster utilization is high during certain periods. The COPY command usually completes within a couple of seconds. However, when load spike occurs, locks can exist and data can be missed. Currently, the AWS Glue job is configured to run without retries, with timeout at 5 minutes and concurrency at 1.
How should a data analytics specialist configure the AWS Glue job to optimize fault tolerance and improve data availability in the Amazon Redshift cluster?

A. Increase the number of retrie
B. Decrease the timeout valu
C. Increase the job concurrency.
D. Keep the number of retries at 0. Decrease the timeout valu
E. Increase the job concurrency.
F. Keep the number of retries at 0. Decrease the timeout valu
G. Keep the job concurrency at 1.
H. Keep the number of retries at 0. Increase the timeout valu
I. Keep the job concurrency at 1.

Answer: B

NEW QUESTION 10
A company is migrating from an on-premises Apache Hadoop cluster to an Amazon EMR cluster. The cluster runs only during business hours. Due to a company requirement to avoid intraday cluster failures, the EMR cluster must be highly available. When the cluster is terminated at the end of each business day, the data must persist.
Which configurations would enable the EMR cluster to meet these requirements? (Choose three.)

A. EMR File System (EMRFS) for storage
B. Hadoop Distributed File System (HDFS) for storage
C. AWS Glue Data Catalog as the metastore for Apache Hive
D. MySQL database on the master node as the metastore for Apache Hive
E. Multiple master nodes in a single Availability Zone
F. Multiple master nodes in multiple Availability Zones

Answer: ACE

Explanation:
https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-ha.html "Note : The cluster can reside only in one Availability Zone or subnet."

NEW QUESTION 10
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 scrip
B. Push the enriched data to a fleet of Kinesis data streams and enable the data transformation feature to flatten the JSON fil
C. Instantiate a dense storage Amazon Redshift cluster and use it as the destination for the Kinesis Data Firehose delivery stream.
D. Update the sensors code to use the PutRecord/PutRecords call from the Kinesis Data Streams API with the AWS SDK for Jav
E. Use Kinesis Data Analytics to enrich the data based on a company-developed anomaly detection SQL scrip
F. Direct the output of KDA application to a Kinesis Data Firehose delivery stream, enable the data transformation feature to flatten the JSON file, and set the Kinesis Data Firehose destination to an Amazon Elasticsearch Service cluster.
G. Set the RecordMaxBufferedTime property of the KPL to "0" to disable the buffering on the sensor side.Connect for each stream a dedicated Kinesis Data Firehose delivery stream and enable the data transformation feature to flatten the JSON file before sending it to an Amazon S3 bucke
H. Load the S3 data into an Amazon Redshift cluster.
I. Update the sensors code to use the PutRecord/PutRecords call from the Kinesis Data Streams API with the AWS SDK for Jav
J. Use AWS Glue to fetch and process data from the stream using the Kinesis Client Library (KCL). Instantiate an Amazon Elasticsearch Service cluster and use AWS Lambda to directly push data into it.

Answer: B

Explanation:
https://docs.aws.amazon.com/streams/latest/dev/developing-producers-with-kpl.html
The KPL can incur an additional processing delay of up to RecordMaxBufferedTime within the library (user-configurable). Larger values of RecordMaxBufferedTime results in higher packing efficiencies and better performance. Applications that cannot tolerate this additional delay may need to use the AWS SDK directly.

NEW QUESTION 13
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 cluste
B. Use Amazon Redshift Spectrum to query data in the data lak
C. Use S3 lifecycle management rules to store data from the previous 12 months in Amazon S3 Glacier storage.
D. Leverage DS2 nodes for the Amazon Redshift cluste
E. Migrate all data from Amazon S3 to Amazon Redshif
F. Decommission the data lake.
G. Store the most recent 4 months of data in the Amazon Redshift cluste
H. Use Amazon Redshift Spectrum to query data in the data lak
I. Ensure the S3 Standard storage class is in use with objects in the data lake.
J. Store the most recent 4 months of data in the Amazon Redshift cluste
K. Use Amazon Redshift federated queries to join cluster data with the data lake to reduce cost
L. Ensure the S3 Standard storage class is in use with objects in the data lake.

Answer: C

NEW QUESTION 15
A retail company is building its data warehouse solution using Amazon Redshift. As a part of that effort, the company is loading hundreds of files into the fact table created in its Amazon Redshift cluster. The company wants the solution to achieve the highest throughput and optimally use cluster resources when loading data into the company’s fact table.
How should the company meet these requirements?

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

Answer: D

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

NEW QUESTION 16
An online retailer is rebuilding its inventory management system and inventory reordering system to automatically reorder products by using Amazon Kinesis Data Streams. The inventory management system uses the Kinesis Producer Library (KPL) to publish data to a stream. The inventory reordering system uses the Kinesis Client Library (KCL) to consume data from the stream. The stream has been configured to scale as needed. Just before production deployment, the retailer discovers that the inventory reordering system is receiving duplicated data.
Which factors could be causing the duplicated data? (Choose two.)

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

Answer: BD

NEW QUESTION 19
An airline has .csv-formatted data stored in Amazon S3 with an AWS Glue Data Catalog. Data analysts want to join this data with call center data stored in Amazon Redshift as part of a dally batch process. The Amazon Redshift cluster is already under a heavy load. The solution must be managed, serverless, well-functioning, and minimize the load on the existing Amazon Redshift cluster. The solution should also require minimal effort and development activity.
Which solution meets these requirements?

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

Answer: C

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

NEW QUESTION 22
A central government organization is collecting events from various internal applications using Amazon Managed Streaming for Apache Kafka (Amazon MSK). The organization has configured a separate Kafka topic for each application to separate the data. For security reasons, the Kafka cluster has been configured to only allow TLS encrypted data and it encrypts the data at rest.
A recent application update showed that one of the applications was configured incorrectly, resulting in writing data to a Kafka topic that belongs to another application. This resulted in multiple errors in the analytics pipeline as data from different applications appeared on the same topic. After this incident, the organization wants to prevent applications from writing to a topic different than the one they should write to.
Which solution meets these requirements with the least amount of effort?

A. Create a different Amazon EC2 security group for each applicatio
B. Configure each security group to have access to a specific topic in the Amazon MSK cluste
C. Attach the security group to each application based on the topic that the applications should read and write to.
D. Install Kafka Connect on each application instance and configure each Kafka Connect instance to write to a specific topic only.
E. Use Kafka ACLs and configure read and write permissions for each topi
F. Use the distinguished name of the clients' TLS certificates as the principal of the ACL.
G. Create a different Amazon EC2 security group for each applicatio
H. Create an Amazon MSK cluster and Kafka topic for each applicatio
I. Configure each security group to have access to the specific cluster.

Answer: B

NEW QUESTION 27
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 hour
C. Set the distribution style to ALL.
D. Use elastic resize to quickly add nodes during peak time
E. Remove the nodes when they are not needed.
F. Use a snapshot, restore, and resize operatio
G. Switch to the new target cluster.

Answer: A

Explanation:
https://docs.aws.amazon.com/redshift/latest/dg/cm-c-implementing-workload-management.html

NEW QUESTION 30
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 accoun
B. Create different S3 buckets for each department and move all the data from every account to the central data lake accoun
C. 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.
D. Keep the account structure and the individual AWS Glue catalogs on each accoun
E. Add a central data lake account and use AWS Glue to catalog data from various account
F. Configure cross-account access for AWS Glue crawlers to scan the data in each departmental S3 bucket to identify the schema and populate the catalo
G. Add the senior data analysts into the central account and apply highly detailed access controls in the Data Catalog and Amazon S3.
H. Set up an individual AWS account for the central data lak
I. Use AWS Lake Formation to catalog the cross- account location
J. On each individual S3 bucket, modify the bucket policy to grant S3 permissions to the Lake Formation service-linked rol
K. Use Lake Formation permissions to addfine-grained access controls to allow senior analysts to view specific tables and columns.
L. Set up an individual AWS account for the central data lake and configure a central S3 bucke
M. Use an AWS Lake Formation blueprint to move the data from the various buckets into the central S3 bucke
N. On each individual bucket, modify the bucket policy to grant S3 permissions to the Lake Formation service-linked rol
O. Use Lake Formation permissions to add fine-grained access controls for both associate and senior analysts to view specific tables and columns.

Answer: C

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 33
A data analyst is designing an Amazon QuickSight dashboard using centralized sales data that resides in Amazon Redshift. The dashboard must be restricted so that a salesperson in Sydney, Australia, can see only the Australia view and that a salesperson in New York can see only United States (US) data.
What should the data analyst do to ensure the appropriate data security is in place?

A. Place the data sources for Australia and the US into separate SPICE capacity pools.
B. Set up an Amazon Redshift VPC security group for Australia and the US.
C. Deploy QuickSight Enterprise edition to implement row-level security (RLS) to the sales table.
D. Deploy QuickSight Enterprise edition and set up different VPC security groups for Australia and the US.

Answer: D

NEW QUESTION 36
A company hosts an on-premises PostgreSQL database that contains historical data. An internal legacy application uses the database for read-only activities. The company's business team wants to move the data to a data lake in Amazon S3 as soon as possible and enrich the data for analytics.
The company has set up an AWS Direct Connect connection between its VPC and its on-premises network. A data analytics specialist must design a solution that achieves the business team's goals with the least operational overhead.
Which solution meets these requirements?

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

Answer: B

NEW QUESTION 37
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. Change the partition key from facility ID to a randomly generated key
B. Increase the number of shards
C. Archive the data on the producers' side
D. Change the partition key from facility ID to capture date

Answer: B

NEW QUESTION 40
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 shar
B. The data analyst needs to make sure there is only a single shard in the stream and no stream resize runs.
C. The hash key generation process for the records is not working correctl
D. The data analyst should generate an explicit hash key on the producer side so the records are directed to the appropriate shard accurately.
E. The records are not being received by Kinesis Data Streams in orde
F. The producer should use the PutRecords API call instead of the PutRecord API call with the SequenceNumberForOrdering parameter.
G. The consumer is not processing the parent shard completely before processing the child shards after a stream resiz
H. The data analyst should process the parent shard completely first before processing the child shards.

Answer: D

Explanation:
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 42
A company developed a new elections reporting website that uses Amazon Kinesis Data Firehose to deliver full logs from AWS WAF to an Amazon S3 bucket.
The company is now seeking a low-cost option to perform this infrequent data analysis with visualizations of logs in a way that requires minimal development effort.
Which solution meets these requirements?

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

Answer: A

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

NEW QUESTION 44
A smart home automation company must efficiently ingest and process messages from various connected devices and sensors. The majority of these messages are comprised of a large number of small files. These messages are ingested using Amazon Kinesis Data Streams and sent to Amazon S3 using a Kinesis data stream consumer application. The Amazon S3 message data is then passed through a processing pipeline built on Amazon EMR running scheduled PySpark jobs.
The data platform team manages data processing and is concerned about the efficiency and cost of downstream data processing. They want to continue to use PySpark.
Which solution improves the efficiency of the data processing jobs and is well architected?

A. Send the sensor and devices data directly to a Kinesis Data Firehose delivery stream to send the data to Amazon S3 with Apache Parquet record format conversion enable
B. Use Amazon EMR running PySpark to process the data in Amazon S3.
C. Set up an AWS Lambda function with a Python runtime environmen
D. Process individual Kinesis data stream messages from the connected devices and sensors using Lambda.
E. Launch an Amazon Redshift cluste
F. Copy the collected data from Amazon S3 to Amazon Redshift and move the data processing jobs from Amazon EMR to Amazon Redshift.
G. Set up AWS Glue Python jobs to merge the small data files in Amazon S3 into larger files and transform them to Apache Parquet forma
H. Migrate the downstream PySpark jobs from Amazon EMR to AWS Glue.

Answer: D

Explanation:
https://aws.amazon.com/it/about-aws/whats-new/2020/04/aws-glue-now-supports-serverless-streaming-etl/

NEW QUESTION 47
A banking company is currently using an Amazon Redshift cluster with dense storage (DS) nodes to store sensitive data. An audit found that the cluster is unencrypted. Compliance requirements state that a database with sensitive data must be encrypted through a hardware security module (HSM) with automated key rotation.
Which combination of steps is required to achieve compliance? (Choose two.)

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

Answer: BD

NEW QUESTION 48

A streaming application is reading data from Amazon Kinesis Data Streams and immediately writing the data to an Amazon S3 bucket every 10 seconds. The application is reading data from hundreds of shards. The batch interval cannot be changed due to a separate requirement. The data is being accessed by Amazon Athena. Users are seeing degradation in query performance as time progresses.
Which action can help improve query performance?

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

Answer: A

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

NEW QUESTION 50
An insurance company has raw data in JSON format that is sent without a predefined schedule through an Amazon Kinesis Data Firehose delivery stream to an Amazon S3 bucket. An AWS Glue crawler is scheduled to run every 8 hours to update the schema in the data catalog of the tables stored in the S3 bucket. Data analysts analyze the data using Apache Spark SQL on Amazon EMR set up with AWS Glue Data Catalog as the metastore. Data analysts say that, occasionally, the data they receive is stale. A data engineer needs to provide access to the most up-to-date data.
Which solution meets these requirements?

A. Create an external schema based on the AWS Glue Data Catalog on the existing Amazon Redshift cluster to query new data in Amazon S3 with Amazon Redshift Spectrum.
B. Use Amazon CloudWatch Events with the rate (1 hour) expression to execute the AWS Glue crawler every hour.
C. Using the AWS CLI, modify the execution schedule of the AWS Glue crawler from 8 hours to 1 minute.
D. Run the AWS Glue crawler from an AWS Lambda function triggered by an S3:ObjectCreated:* eventnotification on the S3 bucket.

Answer: D

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

NEW QUESTION 51
A marketing company is using Amazon EMR clusters for its workloads. The company manually installs third- party libraries on the clusters by logging in to the master nodes. A data analyst needs to create an automated solution to replace the manual process.
Which options can fulfill these requirements? (Choose two.)

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

Answer: AE

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

NEW QUESTION 56
A marketing company wants to improve its reporting and business intelligence capabilities. During the planning phase, the company interviewed the relevant stakeholders and discovered that:

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

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

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

Answer: B

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

Answer: D

NEW QUESTION 59
A company is planning to create a data lake in Amazon S3. The company wants to create tiered storage based on access patterns and cost objectives. The solution must include support for JDBC connections from legacy clients, metadata management that allows federation for access control, and batch-based ETL using PySpark and Scala. Operational management should be limited.
Which combination of components can meet these requirements? (Choose three.)

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

Answer: BEF

NEW QUESTION 61
An airline has been collecting metrics on flight activities for analytics. A recently completed proof of concept demonstrates how the company provides insights to data analysts to improve on-time departures. The proof of concept used objects in Amazon S3, which contained the metrics in .csv format, and used Amazon Athena for querying the data. As the amount of data increases, the data analyst wants to optimize the storage solution to improve query performance.
Which options should the data analyst use to improve performance as the data lake grows? (Choose three.)

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

Answer: CDF

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

NEW QUESTION 64
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 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 dat
B. Use the Spark Streaming application to detect the known event sequence and send the SNS message.
C. 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.
D. Create an Amazon Kinesis Data Firehose delivery stream to capture the incoming sensor dat
E. Use an AWS Lambda transformation function to detect the known event sequence and send the SNS message.
F. Create an Amazon Kinesis data stream to capture the incoming sensor data and create another stream for alert message
G. Set up AWS Application Auto Scaling on bot
H. Create a Kinesis Data Analytics for Java application to detect the known event sequence, and add a message to the message strea
I. Configure an AWS Lambda function to poll the message stream and publish to the SNS topic.

Answer: D

NEW QUESTION 65

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

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

Answer: D

NEW QUESTION 67
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 Formatio
B. Once Lake Formation has the data, apply permissions on Lake Formation.
C. To create the data catalog, run an AWS Glue crawler on the existing Parquet dat
D. Register the Amazon S3 path and then apply permissions through Lake Formation to provide granular-level security.
E. Install Apache Ranger on an Amazon EC2 instance and integrate with Amazon EM
F. Using Ranger policies, create role-based access control for the existing data assets in Amazon S3.
G. Create multiple IAM roles for different users and group
H. Assign IAM roles to different data assets in Amazon S3 to create table-based and column-based access controls.

Answer: A

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

NEW QUESTION 70
A company analyzes its data in an Amazon Redshift data warehouse, which currently has a cluster of three dense storage nodes. Due to a recent business acquisition, the company needs to load an additional 4 TB of user data into Amazon Redshift. The engineering team will combine all the user data and apply complex calculations that require I/O intensive resources. The company needs to adjust the cluster's capacity to support the change in analytical and storage requirements.
Which solution meets these requirements?

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

Answer: C

NEW QUESTION 74
A retail company leverages Amazon Athena for ad-hoc queries against an AWS Glue Data Catalog. The data analytics team manages the data catalog and data access for the company. The data analytics team wants to separate queries and manage the cost of running those queries by different workloads and teams. Ideally, the data analysts want to group the queries run by different users within a team, store the query results in individual Amazon S3 buckets specific to each team, and enforce cost constraints on the queries run against the Data Catalog.
Which solution meets these requirements?

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

Answer: C

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

NEW QUESTION 76
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 tabl
B. 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 exist
C. Change the SQL query in the application to include the new field in the SELECT statement.
D. 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 applicatio
E. 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.
F. Store the mapping file in an Amazon S3 bucket and configure it as a reference data source for the Kinesis Data Analytics applicatio
G. 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.
H. Store the contents of the mapping file in an Amazon DynamoDB tabl
I. 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 exist
J. Forward the record from the Lambda function to the original application destination.

Answer: C

NEW QUESTION 79
.....

Thank You for Trying Our Product

We offer two products:

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

2nd - Questons and Answers in PDF Format

DAS-C01 Practice Exam Features:

* DAS-C01 Questions and Answers Updated Frequently

* DAS-C01 Practice Questions Verified by Expert Senior Certified Staff

* DAS-C01 Most Realistic Questions that Guarantee you a Pass on Your FirstTry

* DAS-C01 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year