Practice Exam Questions

amazon web services

DAS-C01

Certified Data Analytics Specialty

EXAMKILLER

Help Pass Your Exam At First Try

Amazon Web Services

Exam DAS-C01

AWS Certified Data Analytics - Specialty

Version: 6.0

[Total Questions: 157]

Question No : 1

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 heat maps as the visual type.
B. Use Amazon QuickSight with Amazon S3 as the data source. Use heat maps as the visual type.
C. Use Amazon QuickSight with Amazon Athena as the data source. Use pivot tables as the visual type.
D. Use Amazon QuickSight with Amazon S3 as the data source. Use pivot tables as the visual type.

Answer: A

Question No : 2

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

Explanation: 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

Question No : 3

A company uses Amazon kinesis Data Streams to ingest and process customer behavior information from application users each day. A data analytics specialist notices that its data stream is throttling. The specialist has turned on enhanced monitoring for the Kinesis data stream and has verified that the data stream did not exceed the data limits. The specialist discovers that there are hot shards

Which solution will resolve this issue?

A. Use a random partition key to ingest the records.
B. Increase the number of shards Split the size of the log records.
C. Limit the number of records that are sent each second by the producer to match the capacity of the stream.
D. Decrease the size of the records that are sent from the producer to match the capacity of the stream.

Answer: A

Question No : 4

A financial services company needs to aggregate daily stock trade data from the exchanges into a data store. The company requires that data be streamed directly into the data store, but also occasionally allows data to be modified using SQL. The solution should integrate complex, analytic queries running with minimal latency. The solution must provide a business intelligence dashboard that enables viewing of the top contributors to anomalies in stock prices.

Which solution meets the company’s requirements?

A. Use Amazon Kinesis Data Firehose to stream data to Amazon S3. Use Amazon Athena as a data source for Amazon QuickSight to create a business intelligence dashboard.
B. Use Amazon Kinesis Data Streams to stream data to Amazon 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 Streams to stream data to Amazon S3. Use Amazon Athena as a data source for Amazon QuickSight to create a business intelligence dashboard.

Answer: C

Question No : 5

A marketing company has data in Salesforce, MySQL, and Amazon S3. The company wants to use data from these three locations and create mobile dashboards for its users. The company is unsure how it should create the dashboards and needs a solution with the least possible customization and coding.

Which solution meets these requirements?

A. Use Amazon Athena federated queries to join the data sources. Use Amazon QuickSight to generate the mobile dashboards.
B. Use AWS Lake Formation to migrate the data sources into Amazon S3. Use Amazon QuickSight to generate the mobile dashboards.
C. Use Amazon Redshift federated queries to join the data sources. Use Amazon QuickSight to generate the mobile dashboards.
D. Use Amazon QuickSight to connect to the data sources and generate the mobile dashboards.

Answer: C
Reference: https://aws.amazon.com/blogs/big-data/accessing-and-visualizing-data-from-multiple-data- sources-with-amazon-athena-and-amazon-quicksight/

Question No : 6

A company has an encrypted Amazon Redshift cluster. The company recently enabled

Amazon Redshift audit logs and needs to ensure that the audit logs are also encrypted at rest. The logs are retained for 1 year. The auditor queries the logs once a month.

What is the MOST cost-effective way to meet these requirements?

A. Encrypt the Amazon S3 bucket where the logs are stored by using AWS Key Management Service (AWS KMS). Copy the data into the Amazon Redshift cluster from Amazon S3 on a daily basis. Query the data as required.
B. Disable encryption on the Amazon Redshift cluster, configure audit logging, and encrypt the Amazon Redshift cluster. Use Amazon Redshift Spectrum to query the data as required.
C. Enable default encryption on the Amazon S3 bucket where the logs are stored by using AES-256 encryption. Copy the data into the Amazon Redshift cluster from Amazon S3 on a daily basis. Query the data as required.
D. Enable default encryption on the Amazon S3 bucket where the logs are stored by using AES-256 encryption. Use Amazon Redshift Spectrum to query the data as required.

Answer: A

Question No : 7

A large retailer has successfully migrated to an Amazon S3 data lake architecture. The company’s marketing team is using Amazon Redshift and Amazon QuickSight to analyze data, and derive and visualize insights. To ensure the marketing team has the most up-to-date actionable information, a data analyst implements nightly refreshes of Amazon Redshift using terabytes of updates from the previous day.

After the first nightly refresh, users report that half of the most popular dashboards that had been running correctly before the refresh are now running much slower. Amazon CloudWatch does not show any alerts.

What is the MOST likely cause for the performance degradation?

A. The dashboards are suffering from inefficient SQL queries.
B. The cluster is undersized for the queries being run by the dashboards.
C. The nightly data refreshes are causing a lingering transaction that cannot be automatically closed by Amazon Redshift due to ongoing user workloads.
D. The nightly data refreshes left the dashboard tables in need of a vacuum operation that could not be automatically performed by Amazon Redshift due to ongoing user workloads.

Answer: D

Explanation: https://github.com/awsdocs/amazon-redshift-developer-guide/issues/21

Question No : 8

A financial company hosts a data lake in Amazon S3 and a data warehouse on an Amazon Redshift cluster. The company uses Amazon QuickSight to build dashboards and wants to secure access from its on-premises Active Directory to Amazon QuickSight.

How should the data be secured?

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

Answer: A

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

Question No : 9

A company's data analyst needs to ensure that queries executed in Amazon Athena cannot scan more than a prescribed amount of data for cost control purposes. Queries that exceed the prescribed threshold must be canceled immediately.

What should the data analyst do to achieve this?

A. Configure Athena to invoke an AWS Lambda function that terminates queries when the prescribed threshold is crossed.
B. For each workgroup, set the control limit for each query to the prescribed threshold.

C. Enforce the prescribed threshold on all Amazon S3 bucket policies
D. For each workgroup, set the workgroup-wide data usage control limit to the prescribed threshold.

Answer: B

Explanation:

https://docs.aws.amazon.com/athena/latest/ug/manage-queries-control-costs-with-workgroups.html

Question No : 10

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

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

Which storage service will provide the MOST performant solution?

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

Answer: B

Question No : 11

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 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.
B. 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.
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 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.

Answer: A
Reference: https://docs.aws.amazon.com/prescriptive-guidance/latest/patterns/build-an-etl-service-pipeline-to-load-data-incrementally-from-amazon-s3-to-amazon-redshift-using-aws-glue.html

Question No : 12

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: A

Explanation:

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

Question No : 13

A financial company uses Amazon S3 as its data lake and has set up a data warehouse using a multi-node Amazon Redshift cluster. The data files in the data lake are organized in folders based on the data source of each data file. All the data files are loaded to one table in the Amazon Redshift cluster using a separate COPY command for each data file location. With this approach, loading all the data files into Amazon Redshift takes a long time to complete. Users want a faster solution with little or no increase in cost while maintaining the segregation of the data files in the S3 data lake.

Which solution meets these requirements?

A. Use Amazon EMR to copy all the data files into one folder and issue a COPY command to load the data into Amazon Redshift.
B. Load all the data files in parallel to Amazon Aurora, and run an AWS Glue job to load the data into Amazon Redshift.
C. Use an AWS Glue job to copy all the data files into one folder and issue a COPY command to load the data into Amazon Redshift.
D. Create a manifest file that contains the data file locations and issue a COPY command to load the data into Amazon Redshift.

Answer: D

Explanation:

https://docs.aws.amazon.com/redshift/latest/dg/loading-data-files-using-manifest.html "You can use a manifest to ensure that the COPY command loads all of the required files, and only the required files, for a data load"

Question No : 14

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 process. 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 format. Use Amazon Athena to query the data.
B. Create an Amazon RDS for PostgreSQL database and use AWS Database Migration Service (AWS DMS) to migrate the data into Amazon RDS. 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.
C. Configure an AWS Glue crawler to use a JDBC connection to catalog the data in the on-premises database. Use an AWS Glue job to enrich the data and save the result to Amazon S3 in Apache Parquet format. Create an Amazon Redshift cluster and use Amazon Redshift Spectrum to query the data.
D. Configure an AWS Glue crawler to use a JDBC connection to catalog the data in the on-premises database. Use an AWS Glue job to enrich the data and save the result to Amazon S3 in Apache Parquet format. Use Amazon Athena to query the data.

Answer: B

Question No : 15

A company is building an analytical solution that includes Amazon S3 as data lake storage and Amazon Redshift for data warehousing. The company wants to use Amazon Redshift Spectrum to query the data that is stored in Amazon S3.

Which steps should the company take to improve performance when the company uses Amazon Redshift Spectrum to query the S3 data files? (Select THREE)

Use gzip compression with individual file sizes of 1-5 GB

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

Answer: B,C,D

Question No : 16

A software company hosts an application on AWS, and new features are released weekly. As part of the application testing process, a solution must be developed that analyzes logs from each Amazon EC2 instance to ensure that the application is working as expected after each deployment. The collection and analysis solution should be highly available with the ability to display new information with minimal delays.

Which method should the company use to collect and analyze the logs?

A. Enable detailed monitoring on Amazon EC2, use Amazon CloudWatch agent to store logs in Amazon S3, and use Amazon Athena for fast, interactive log analytics.
B. Use the Amazon Kinesis Producer Library (KPL) agent on Amazon EC2 to collect and send data to Kinesis Data Streams to further push the data to Amazon Elasticsearch Service and visualize using Amazon QuickSight.
C. Use the Amazon Kinesis Producer Library (KPL) agent on Amazon EC2 to collect and send data to Kinesis Data Firehose to further push the data to Amazon Elasticsearch Service and Kibana.
D. Use Amazon CloudWatch subscriptions to get access to a real-time feed of logs and have the logs delivered to Amazon Kinesis Data Streams to further push the data to Amazon Elasticsearch Service and Kibana.

Answer: D
Reference:
https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/Subscriptions.html

Question No : 17

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: B,D

Reference: https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-db-encryption.html

Question No : 18

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

Question No : 19

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 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.
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 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.
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 base IAM roles. Create a security configuration mapping for the additional IAM roles to Active Directory user groups for each team.

Answer: C

Question No : 20

A medical company has a system with sensor devices that read metrics and send them in real time to an Amazon Kinesis data stream. The Kinesis data stream has multiple shards.

The company needs to calculate the average value of a numeric metric every second and set an alarm for whenever the value is above one threshold or below another threshold. The alarm must be sent to Amazon Simple Notification Service (Amazon SNS) in less than 30 seconds.

Which architecture meets these requirements?

A. Use an Amazon Kinesis Data Firehose delivery stream to read the data from the Kinesis data stream with an AWS Lambda transformation function that calculates the average per second and sends the alarm to Amazon SNS.
B. Use an AWS Lambda function to read from the Kinesis data stream to calculate the average per second and sent the alarm to Amazon SNS.
C. Use an Amazon Kinesis Data Firehose deliver stream to read the data from the Kinesis data stream and store it on Amazon S3. Have Amazon S3 trigger an AWS Lambda function that calculates the average per second and sends the alarm to Amazon SNS.
D. Use an Amazon Kinesis Data Analytics application to read from the Kinesis data stream and calculate the average per second. Send the results to an AWS Lambda function that sends the alarm to Amazon SNS.

Answer: D
Reference: https://docs.aws.amazon.com/firehose/latest/dev/firehose-dg.pdf

Question No : 21

A data analytics specialist is building an automated ETL ingestion pipeline using AWS Glue to ingest compressed files that have been uploaded to an Amazon S3 bucket. The ingestion pipeline should support incremental data processing.

Which AWS Glue feature should the data analytics specialist use to meet this requirement?

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

Answer: C
Reference: https://docs.aws.amazon.com/prescriptive-guidance/latest/patterns/build-an-etl-service-pipeline-to- load-data-incrementally-from-amazon-s3-to-amazon-redshift-using-aws-glue.html

Question No : 22

A data engineer is using AWS Glue ETL jobs to process data at frequent intervals The processed data is then copied into Amazon S3 The ETL jobs run every 15 minutes. The AWS Glue Data Catalog partitions need to be updated automatically after the completion of each job

Which solution will meet these requirements MOST cost-effectively?

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

Question No : 23

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

Which solution will allow the company to accomplish this goal?

A. Create a subscription filter in the production accounts CloudWatch Logs to target the Kinesis data stream in the test account as its destination In the test account create an 1AM role that grants access to the Kinesis data stream and the CloudWatch Logs resources in the production account
B. In the test account create an 1AM role that grants access to the Kinesis data stream and the CloudWatch Logs resources in the production account Create a destination data stream in Kinesis Data Streams in the test account with an 1AM role and a trust policy that allow CloudWatch Logs in the production account to write to the test account
C. In the test account, create an 1AM role that grants access to the Kinesis data stream and the CloudWatch Logs resources in the production account Create a destination data

stream in Kinesis Data Streams in the test account with an 1AM role and a trust policy that allow CloudWatch Logs in the production account to write to the test account
D. Create a destination data stream in Kinesis Data Streams in the test account with an 1AM role and a trust policy that allow CloudWatch Logs in the production account to write to the test account Create a subscription filter in the production accounts CloudWatch Logs to target the Kinesis data stream in the test account as its destination

Answer: D

Question No : 24

A large energy company is using Amazon QuickSight to build dashboards and report the historical usage data of its customers This data is hosted in Amazon Redshift The reports need access to all the fact tables' billions ot records to create aggregation in real time grouping by multiple dimensions

A data analyst created the dataset in QuickSight by using a SQL query and not SPICE Business users have noted that the response time is not fast enough to meet their needs

Which action would speed up the response time for the reports with the LEAST implementation effort?

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

Answer: A

Question No : 25

A company is sending historical datasets to Amazon S3 for storage. A data engineer at the company wants to make these datasets available for analysis using Amazon Athena. The engineer also wants to encrypt the Athena query results in an S3 results location by using AWS solutions for encryption. The requirements for encrypting the query results are as follows:

Use custom keys for encryption of the primary dataset query results.

Use generic encryption for all other query results.

Provide an audit trail for the primary dataset queries that shows when the keys were used and by whom.

Which solution meets these requirements?

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

Answer: A
Reference: https://d1.awsstatic.com/product-marketing/S3/Amazon_S3_Security_eBook_2020.pdf

Question No : 26

A company uses an Amazon EMR cluster with 50 nodes to process operational data and make the data available for data analysts These jobs run nightly use Apache Hive with the Apache Jez framework as a processing model and write results to Hadoop Distributed File System (HDFS) In the last few weeks, jobs are failing and are producing the following error message

"File could only be replicated to 0 nodes instead of 1"

A data analytics specialist checks the DataNode logs the NameNode logs and network connectivity for potential issues that could have prevented HDFS from replicating data The data analytics specialist rules out these factors as causes for the issue

Which solution will prevent the jobs from failing'?

A. Monitor the HDFSUtilization metric. If the value crosses a user-defined threshold add task nodes to the EMR cluster
B. Monitor the HDFSUtilization metri.c If the value crosses a user-defined threshold add core nodes to the EMR cluster

C. Monitor the MemoryAllocatedMB metric. If the value crosses a user-defined threshold, add task nodes to the EMR cluster
D. Monitor the MemoryAllocatedMB metric. If the value crosses a user-defined threshold, add core nodes to the EMR cluster.

Answer: C

Question No : 27

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

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

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

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

Answer: A,D,E

Question No : 28

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

Which solution will improve the data loading performance?

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

Answer: B

Explanation:

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

Question No : 29

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: D

Explanation:

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

Question No : 30