AWS Certified Data Analytics - Specialty.VCEplus.premium.exam.65q

Number: DAS-C01
Passing Score: 800
Time Limit: 120 min
File Version: 1.0

DAS-C01

AWS Certified Data Analytics - Specialty

Version 1.0

Exam A

QUESTION 1
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.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 2
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.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 3
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 sameAmazon 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 separateAvailability Zone. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Reference: https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hbase-s3.html

QUESTION 4
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.

Correct Answer: D
Section: (none)
Explanation

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

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 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.
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 maximum capacity job parameter.
C. 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.
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.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Reference: https://docs.aws.amazon.com/glue/latest/dg/monitor-debug-capacity.html

QUESTION 6
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.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Reference: https://towardsdatascience.com/update-and-insert-upsert-data-from-aws-glue-698ac582e562

QUESTION 7
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.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 8
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.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 9
A manufacturing company has been collecting IoT sensor data from devices on its factory floor for a year and is storing the data in Amazon Redshift for daily analysis. A data analyst has determined that, at an expected ingestion rate of about 2 TB per day, the cluster will be undersized in less than 4 months. A long-term solution is needed. The data analyst has indicated that most queries only reference the most recent 13 months of data, yet there are also quarterly reports that need to query all the data generated from the past 7 years. The chief technology officer (CTO) is concerned about the costs, administrative effort, and performance of a long-term solution.

Which solution should the data analyst use to meet these requirements?

A. Create a daily job in AWS Glue to UNLOAD records older than 13 months to Amazon S3 and delete those records from Amazon Redshift. Create an external table in Amazon Redshift to point to the S3 location. Use Amazon RedshiftSpectrum to join to data that is older than 13 months.
B. Take a snapshot of the Amazon Redshift cluster. Restore the cluster to a new cluster using dense storage nodes with additional storage capacity.
C. 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.
D. 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 clusterusing Auto Scaling for any daily analytics needs, and use Amazon Athena for the quarterly reports, with both using the same AWS Glue Data Catalog.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 10

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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 11
A company that produces network devices has millions of users. Data is collected from the devices on an hourly basis and stored in an Amazon S3 data lake.

The company runs analyses on the last 24 hours of data flow logs for abnormality detection and to troubleshoot and resolve user issues. The company also analyzes historical logs dating back 2 years to discover patterns and look for improvement opportunities.

The data flow logs contain many metrics, such as date, timestamp, source IP, and target IP. There are about 10 billion events every day.

How should this data be stored for optimal performance?

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

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 12
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.

Correct Answer: BD
Section: (none)
Explanation

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

QUESTION 13
A company is planning to do a proof of concept for a machine earning (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. 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.
B. Create custom ETL jobs on-premises to curate the data. Use AWS DMS to ingest data into Amazon S3 for ML processing.
C. Ingest data into Amazon S3 using AWS DMS. Use AWS Glue to perform data curation and store the data in Amazon 3 for ML processing.
D. 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.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 14
A US-based sneaker retail company launched its global website. All the transaction data is stored in Amazon RDS and curated historic transaction data is stored in Amazon Redshift in the us-east-1 Region. The business intelligence (BI) team wants to enhance the user experience by providing a dashboard for sneaker trends.

The BI team decides to use Amazon QuickSight to render the website dashboards. During development, a team in Japan provisioned Amazon QuickSight in ap-northeast-1. The team is having difficulty connecting Amazon QuickSight from ap-northeast-1 to Amazon Redshift in us-east-1.

Which solution will solve this issue and meet the requirements?

A. In the Amazon Redshift console, choose to configure cross-Region snapshots and set the destination Region as ap-northeast-1. Restore the Amazon Redshift Cluster from the snapshot and connect to Amazon QuickSight launched inap-northeast-1.
B. Create a VPC endpoint from the Amazon QuickSight VPC to the Amazon Redshift VPC so Amazon QuickSight can access data from Amazon Redshift.
C. Create an Amazon Redshift endpoint connection string with Region information in the string and use this connection string in Amazon QuickSight to connect to Amazon Redshift.
D. Create a new security group for Amazon Redshift in us-east-1 with an inbound rule authorizing access from the appropriate IP address range for the Amazon QuickSight servers in ap-northeast-1.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 15
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.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 16
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 Amazon QuickSight console.
B. Edit the permissions for the new S3 bucket from within the Amazon QuickSight console.
C. Edit the permissions for the AWS Glue Data Catalog from within the AWS Glue console.
D. Edit the permissions for the new S3 bucket from within the S3 console.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Reference: https://aws.amazon.com/blogs/big-data/harmonize-query-and-visualize-data-from-various-providers-using-aws-glue-amazon-athena-and-amazon-quicksight/

QUESTION 17
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 topersist data to an S3 bucket.
B. 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 Firehoseon the 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 DataFirehose on the second Kinesis data stream to persist data to an S3 bucket.
D. 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 thesecond Kinesis data stream to persist data to an S3 bucket.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 18
A company currently uses Amazon Athena to query its global datasets. The regional data is stored in Amazon S3 in the us-east-1 and us-west-2 Regions. The data is not encrypted. To simplify the query process and manage it centrally, the company wants to use Athena in us-west-2 to query data from Amazon S3 in both Regions. The solution should be as low-cost as possible.

What should the company do to achieve this goal?

A. Use AWS DMS to migrate the AWS Glue Data Catalog from us-east-1 to us-west-2. Run Athena queries in us-west-2.
B. Run the AWS Glue crawler in us-west-2 to catalog datasets in all Regions. Once the data is crawled, run Athena queries in us-west-2.
C. Enable cross-Region replication for the S3 buckets in us-east-1 to replicate data in us-west-2. Once the data is replicated in us-west-2, run the AWS Glue crawler there to update the AWS Glue Data Catalog in us-west-2 and run Athenaqueries.
D. Update AWS Glue resource policies to provide us-east-1 AWS Glue Data Catalog access to us-west-2. Once the catalog in us-west-2 has access to the catalog in us-east-1, run Athena queries in us-west-2.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 19
A large company receives files from external parties in Amazon EC2 throughout the day. At the end of the day, the files are combined into a single file, compressed into a gzip file, and uploaded to Amazon S3. The total size of all the files is close to 100 GB daily. Once the files are uploaded to Amazon S3, an AWS Batch program executes a COPY command to load the files into an Amazon Redshift cluster.

Which program modification will accelerate the COPY process?

A. Upload the individual files to Amazon S3 and run the COPY command as soon as the files become available.
B. Split the number of files so they are equal to a multiple of the number of slices in the Amazon Redshift cluster. Gzip and upload the files to Amazon S3. Run the COPY command on the files.
C. Split the number of files so they are equal to a multiple of the number of compute nodes in the Amazon Redshift cluster. Gzip and upload the files to Amazon S3. Run the COPY command on the files.
D. Apply sharding by breaking up the files so the distkey columns with the same values go to the same file. Gzip and upload the sharded files to Amazon S3. Run the COPY command on the files.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Reference: https://docs.aws.amazon.com/redshift/latest/dg/t_splitting-data-files.html

QUESTION 20
A large ride-sharing company has thousands of drivers globally serving millions of unique customers every day. The company has decided to migrate an existing data mart to Amazon Redshift. The existing schema includes the following tables.

- A trips fact table for information on completed rides.
- A drivers dimension table for driver profiles.
- A customers fact table holding customer profile information.

The company analyzes trip details by date and destination to examine profitability by region. The drivers data rarely changes. The customers data frequently changes.

What table design provides optimal query performance?

A. Use DISTSTYLE KEY (destination) for the trips table and sort by date. Use DISTSTYLE ALL for the drivers and customers tables.
B. Use DISTSTYLE EVEN for the trips table and sort by date. Use DISTSTYLE ALL for the drivers table. Use DISTSTYLE EVEN for the customers table.
C. Use DISTSTYLE KEY (destination) for the trips table and sort by date. Use DISTSTYLE ALL for the drivers table. Use DISTSTYLE EVEN for the customers table.
D. Use DISTSTYLE EVEN for the drivers table and sort by date. Use DISTSTYLE ALL for both fact tables.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 21
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'sEMR 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 clusterEC2 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 clusterEC2 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 clusterEC2 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.

Correct Answer: C
Section: (none)
Explanation
Explanation/Reference:

QUESTION 22
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

Correct Answer: BEF
Section: (none)
Explanation

Explanation/Reference:
Reference: https://d1.awsstatic.com/whitepapers/Storage/data-lake-on-aws.pdf

QUESTION 23
A company wants to optimize the cost of its data and analytics platform. The company is ingesting a number of .csv and JSON files in Amazon S3 from various data sources. Incoming data is expected to be 50 GB each day. The company is using Amazon Athena to query the raw data in Amazon S3 directly. Most queries aggregate data from the past 12 months, and data that is older than 5 years is infrequently queried. The typical query scans about 500 MB of data and is expected to return results in less than 1 minute. The raw data must be retained indefinitely for compliance requirements.

Which solution meets the company's requirements?

A. Use an AWS Glue ETL job to compress, partition, and convert the data into a columnar data format. Use Athena to query the processed dataset. Configure a lifecycle policy to move the processed data into the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class 5 years after object creation. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after object creation.
B. Use an AWS Glue ETL job to partition and convert the data into a row-based data format. Use Athena to query the processed dataset. Configure a lifecycle policy to move the data into the Amazon S3 Standard-Infrequent Access (S3Standard-IA) storage class 5 years after object creation. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after object creation.
C. Use an AWS Glue ETL job to compress, partition, and convert the data into a columnar data format. Use Athena to query the processed dataset. Configure a lifecycle policy to move the processed data into the Amazon S3 Standard-Infrequent Access (S3 Standard-IA) storage class 5 years after the object was last accessed. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after the last date the object was accessed.
D. Use an AWS Glue ETL job to partition and convert the data into a row-based data format. Use Athena to query the processed dataset. Configure a lifecycle policy to move the data into the Amazon S3 Standard-Infrequent Access (S3Standard-IA) storage class 5 years after the object was last accessed. Configure a second lifecycle policy to move the raw data into Amazon S3 Glacier for long-term archival 7 days after the last date the object was accessed.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 24
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 data. Use theSpark Streaming application to detect the known event sequence and send the SNS message.

B. 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 incomingevents in the RDS database and query the latest data to detect the known event sequence and send the SNS message.
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 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 theknown 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.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Reference: https://aws.amazon.com/kinesis/data-streams/faqs/

QUESTION 25
A media content company has a streaming playback application. The company wants to collect and analyze the data to provide near-real-time feedback on playback issues. The company needs to consume this data and return results within 30 seconds according to the service-level agreement (SLA). The company needs the consumer to identify playback issues, such as quality during a specified timeframe. The data will be emitted as JSON and may change schemas over time.

Which solution will allow the company to collect data for processing while meeting these requirements?

A. Send the data to Amazon Kinesis Data Firehose with delivery to Amazon S3. Configure an S3 event trigger an AWS Lambda function to process the data. The Lambda function will consume the data and process it to identify potentialplayback issues. Persist the raw data to Amazon S3.
B. Send the data to Amazon Managed Streaming for Kafka and configure an Amazon Kinesis Analytics for Java application as the consumer. The application will consume the data and process it to identify potential playback issues. Persistthe raw data to Amazon DynamoDB.
C. Send the data to Amazon Kinesis Data Firehose with delivery to Amazon S3. Configure Amazon S3 to trigger an event for AWS Lambda to process. The Lambda function will consume the data and process it to identify potentialplayback issues. Persist the raw data to Amazon DynamoDB.
D. Send the data to Amazon Kinesis Data Streams and configure an Amazon Kinesis Analytics for Java application as the consumer. The application will consume the data and process it to identify potential playback issues. Persist the rawdata to Amazon S3.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 26
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 queriesusing 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. Create an AWS Glue Data Catalog of the data in Amazon S3. Use Amazon Athena to query the data.
D. 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.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 27
A company leverages Amazon Athena for ad-hoc queries against data stored in Amazon S3. The company wants to implement additional controls to separate query execution and query history among users, teams, or applications running in the same AWS account to comply with internal security policies.

Which solution meets these requirements?

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

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Reference: https://aws.amazon.com/athena/faqs/

QUESTION 28
A company wants to use an automatic machine learning (ML) Random Cut Forest (RCF) algorithm to visualize complex real-word scenarios, such as detecting seasonality and trends, excluding outers, and imputing missing values.

The team working on this project is non-technical and is looking for an out-of-the-box solution that will require the LEAST amount of management overhead.

Which solution will meet these requirements?

A. Use an AWS Glue ML transform to create a forecast and then use Amazon QuickSight to visualize the data.
B. Use Amazon QuickSight to visualize the data and then use ML-powered forecasting to forecast the key business metrics.
C. Use a pre-build ML AMI from the AWS Marketplace to create forecasts and then use Amazon QuickSight to visualize the data.
D. Use calculated fields to create a new forecast and then use Amazon QuickSight to visualize the data.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Reference: https://aws.amazon.com/blogs/big-data/query-visualize-and-forecast-trufactor-web-session-intelligence-with-aws-data-exchange/

QUESTION 29
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.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 30
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 theLambda function to run daily.

B. 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 Oozieworkflow in the cluster to invoke the Hive script daily.
C. 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.
D. 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.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 31
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.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Reference: https://aws.amazon.com/blogs/big-data/using-amazon-redshift-spectrum-amazon-athena-and-aws-glue-with-node-js-in-production/

QUESTION 32
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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 33
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.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Reference: https://aws.amazon.com/redshift/pricing/

QUESTION 34
A company stores its sales and marketing data that includes personally identifiable information (PII) in Amazon S3. The company allows its analysts to launch their own Amazon EMR cluster and run analytics reports with the data. To meet compliance requirements, the company must ensure the data is not publicly accessible throughout this process. A data engineer has secured Amazon S3 but must ensure the individual EMR clusters created by the analysts are not exposed to the public internet.

Which solution should the data engineer to meet this compliance requirement with LEAST amount of effort?

A. Create an EMR security configuration and ensure the security configuration is associated with the EMR clusters when they are created.
B. Check the security group of the EMR clusters regularly to ensure it does not allow inbound traffic from IPv4 0.0.0.0/0 or IPv6 ::/0.
C. Enable the block public access setting for Amazon EMR at the account level before any EMR cluster is created.
D. Use AWS WAF to block public internet access to the EMR clusters across the board.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Reference: https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-security-groups.html

QUESTION 35
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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Reference: https://docs.aws.amazon.com/redshift/latest/dg/r_COPY.html

QUESTION 36
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

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 37
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 nearreal-time data.

Which visualization solution will meet these requirements?

A. 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.
B. 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.
C. 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.
D. 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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 38
A financial company uses Apache Hive on Amazon EMR for ad-hoc queries. Users are complaining of sluggish performance.

A data analyst notes the following:

▪ Approximately 90% of queries are submitted 1 hour after the market opens. ▪
Hadoop Distributed File System (HDFS) utilization never exceeds 10%.

Which solution would help address the performance issues?

A. Create instance fleet configurations for core and task nodes. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch CapacityRemainingGB metric. Create an automatic scaling policy toscale in the instance fleet based on the CloudWatch CapacityRemainingGB metric.
B. Create instance fleet configurations for core and task nodes. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch YARNMemoryAvailablePercentage metric. Create an automatic scalingpolicy to scale in the instance fleet based on the CloudWatch YARNMemoryAvailablePercentage metric.
C. Create instance group configurations for core and task nodes. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch CapacityRemainingGB metric. Create an automatic scaling policy toscale in the instance groups based on the CloudWatch CapacityRemainingGB metric.
D. Create instance group configurations for core and task nodes. Create an automatic scaling policy to scale out the instance groups based on the Amazon CloudWatch YARNMemoryAvailablePercentage metric. Create an automaticscaling policy to scale in the instance groups based on the CloudWatch YARNMemoryAvailablePercentage metric.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 39
A media company has been performing analytics on log data generated by its applications. There has been a recent increase in the number of concurrent analytics jobs running, and the overall performance of existing jobs is decreasing as the number of new jobs is increasing. The partitioned data is stored in Amazon S3 One Zone-Infrequent Access (S3 One Zone-IA) and the analytic processing is performed on Amazon EMR clusters using the EMR File System (EMRFS) with consistent view enabled. A data analyst has determined that it is taking longer for the EMR task nodes to list objects in Amazon S3.

Which action would MOST likely increase the performance of accessing log data in Amazon S3?

A. Use a hash function to create a random string and add that to the beginning of the object prefixes when storing the log data in Amazon S3.
B. Use a lifecycle policy to change the S3 storage class to S3 Standard for the log data.
C. Increase the read capacity units (RCUs) for the shared Amazon DynamoDB table.
D. Redeploy the EMR clusters that are running slowly to a different Availability Zone.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 40
A company has developed several AWS Glue jobs to validate and transform its data from Amazon S3 and load it into Amazon RDS for MySQL in batches once every day. The ETL jobs read the S3 data using a DynamicFrame. Currently, the ETL developers are experiencing challenges in processing only the incremental data on every run, as the AWS Glue job processes all the S3 input data on each run.

Which approach would allow the developers to solve the issue with minimal coding effort?

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

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 41
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 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 isaccessible to the finance team only. Use the COPY command to load the data from Amazon S3 to the finance table.
B. 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 COPYcommand with the IAM role that has access to the KMS key to load the data from S3 to the finance table.
C. 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 thedata from DynamoDB and then insert the output to the finance table in Amazon Redshift.
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 hasaccess to the KMS key to load the data from Amazon S3 to the finance table in Amazon Redshift.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 42
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.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 43
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 Amazon Redshift Enhanced VPC Routing. Enable VPC Flow Logs to monitor traffic.
B. Allow access to the Amazon Redshift database using AWS IAM only. Log access using AWS CloudTrail.
C. Enable audit logging for Amazon Redshift using the AWS Management Console or the AWS CLI.
D. Enable and download audit reports from AWS Artifact.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Reference: https://docs.aws.amazon.com/redshift/latest/mgmt/db-auditing.html

QUESTION 44
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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 45
Once a month, a company receives a 100 MB .csv file compressed with gzip. The file contains 50,000 property listing records and is stored in Amazon S3 Glacier. The company needs its data analyst to query a subset of the data for a specific vendor.

What is the most cost-effective solution?
A. Load the data into Amazon S3 and query it with Amazon S3 Select.
B. Query the data from Amazon S3 Glacier directly with Amazon Glacier Select.
C. Load the data to Amazon S3 and query it with Amazon Athena.
D. Load the data to Amazon S3 and query it with Amazon Redshift Spectrum.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Reference: https://aws.amazon.com/athena/faqs/

QUESTION 46
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.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 47
A data analyst is designing a solution to interactively query datasets with SQL using a JDBC connection. Users will join data stored in Amazon S3 in Apache ORC format with data stored in Amazon Elasticsearch Service (Amazon ES) and Amazon Aurora MySQL.

Which solution will provide the MOST up-to-date results?

A. Use AWS Glue jobs to ETL data from Amazon ES and Aurora MySQL to Amazon S3. Query the data with Amazon Athena.
B. Use Amazon DMS to stream data from Amazon ES and Aurora MySQL to Amazon Redshift. Query the data with Amazon Redshift.
C. Query all the datasets in place with Apache Spark SQL running on an AWS Glue developer endpoint.
D. Query all the datasets in place with Apache Presto running on Amazon EMR.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 48
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 logs. Use Athena to perform ad-hoc analyses and use Amazon QuickSight to develop data visualizations.
B. 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 datavisualizations.
C. Create an AWS Lambda function to convert the logs into .csv format. Then add the function to the Kinesis Data Firehose transformation configuration. Use Amazon Redshift to perform ad-hoc analyses of the logs using SQL queries anduse Amazon QuickSight to develop data visualizations.
D. Create an Amazon EMR cluster and use Amazon S3 as the data source. Create an Apache Spark job to perform ad-hoc analyses and use Amazon QuickSight to develop data visualizations.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 49
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 catalogand 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 toscan 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 Formationservice-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.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 50
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 theJSON 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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 51
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.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 52
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. 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.
B. 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.
C. 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.
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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 53
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.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 54
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 Redshift. Configure Amazon QuickSight with Amazon Redshift as the data source.
B. 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 datasource.
C. 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.
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 EMRas the data source.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 55
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.

Correct Answer: BCD
Section: (none)
Explanation

Explanation/Reference:

QUESTION 56
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.

Correct Answer: C
Section: (none)
Explanation

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

QUESTION 57
A company is migrating its existing on-premises ETL jobs to Amazon EMR. The code consists of a series of jobs written in Java. The company needs to reduce overhead for the system administrators without changing the underlying code. Due to the sensitivity of the data, compliance requires that the company use root device volume encryption on all nodes in the cluster. Corporate standards require that environments be provisioned though AWS CloudFormation when possible.

Which solution satisfies these requirements?

A. Install open-source Hadoop on Amazon EC2 instances with encrypted root device volumes. Configure the cluster in the CloudFormation template.
B. Use a CloudFormation template to launch an EMR cluster. In the configuration section of the cluster, define a bootstrap action to enable TLS.
C. Create a custom AMI with encrypted root device volumes. Configure Amazon EMR to use the custom AMI using the CustomAmild property in the CloudFormation template.
D. 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.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Reference: https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-custom-ami.html

QUESTION 58
A transportation company uses IoT sensors attached to trucks to collect vehicle data for its global delivery fleet. The company currently sends the sensor data in small .csv files to Amazon S3. The files are then loaded into a 10-node Amazon Redshift cluster with two slices per node and queried using both Amazon Athena and Amazon Redshift. The company wants to optimize the files to reduce the cost of querying and also improve the speed of data loading into the Amazon Redshift cluster.

Which solution meets these requirements?

A. Use AWS Glue to convert all the files from .csv to a single large Apache Parquet file. COPY the file into Amazon Redshift and query the file with Athena from Amazon S3.
B. Use Amazon EMR to convert each .csv file to Apache Avro. COPY the files into Amazon Redshift and query the file with Athena from Amazon S3.
C. Use AWS Glue to convert the files from .csv to a single large Apache ORC file. COPY the file into Amazon Redshift and query the file with Athena from Amazon S3.
D. Use AWS Glue to convert the files from .csv to Apache Parquet to create 20 Parquet files. COPY the files into Amazon Redshift and query the files with Athena from Amazon S3.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 59
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 Amazon Elasticsearch Service deployed on Amazon EC2 to aggregate, filter, and process the data. Refresh content performance dashboards in near-real time.
D. Use Kibana to aggregate, filter, and visualize the data stored in Amazon Elasticsearch Service. 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.

Correct Answer: CE
Section: (none)
Explanation

Explanation/Reference:

QUESTION 60
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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 61
A media analytics company consumes a stream of social media posts. The posts are sent to an Amazon Kinesis data stream partitioned on user_id. An AWS Lambda function retrieves the records and validates the content before loading the posts into an Amazon Elasticsearch cluster. The validation process needs to receive the posts for a given user in the order they were received. A data analyst has noticed that, during peak hours, the social media platform posts take more than an hour to appear in the Elasticsearch cluster.

What should the data analyst do reduce this latency?

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

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 62
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.

Correct Answer: AC
Section: (none)
Explanation

Explanation/Reference:

QUESTION 63
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 enabled. Use Amazon EMR running PySpark to process the data inAmazon S3.
B. Set up an AWS Lambda function with a Python runtime environment. Process individual Kinesis data stream messages from the connected devices and sensors using Lambda.
C. Launch an Amazon Redshift cluster. Copy the collected data from Amazon S3 to Amazon Redshift and move the data processing jobs from Amazon EMR to Amazon Redshift.
D. Set up AWS Glue Python jobs to merge the small data files in Amazon S3 into larger files and transform them to Apache Parquet format. Migrate the downstream PySpark jobs from Amazon EMR to AWS Glue.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 64
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 the COPY command with the manifest file to load data into Amazon Redshift.
B. Use S3DistCp to load files into Amazon Redshift.
C. Use temporary staging tables during the loading process.
D. Use the UNLOAD command to upload data into Amazon Redshift.
E. Use Amazon Redshift Spectrum to query files from Amazon S3.

Correct Answer: CE
Section: (none)
Explanation

Explanation/Reference:
Reference: https://aws.amazon.com/blogs/big-data/top-8-best-practices-for-high-performance-etl-processing-using-amazon-redshift/

QUESTION 65
A university intends to use Amazon Kinesis Data Firehose to collect JSON-formatted batches of water quality readings in Amazon S3. The readings are from 50 sensors scattered across a local lake. Students will query the stored data using Amazon Athena to observe changes in a captured metric over time, such as water temperature or acidity. Interest has grown in the study, prompting the university to reconsider how data will be stored. Which data format and partitioning choices will MOST significantly reduce costs? (Choose two.)

A. Store the data in Apache Avro format using Snappy compression.
B. Partition the data by year, month, and day.
C. Store the data in Apache ORC format using no compression.
D. Store the data in Apache Parquet format using Snappy compression.
E. Partition the data by sensor, year, month, and day.

Correct Answer: CD
Section: (none)
Explanation

Explanation/Reference:
Reference: https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html