1) A company ingests a large set of clickstream data in nested JSON format from different sources and stores it in Amazon S3. Data analysts need to analyze this data in combination with data stored in an Amazon Redshift cluster. Data analysts want to build a cost-effective and automated solution for this need.

Which solution meets these requirements?

A) Use Apache Spark SQL on Amazon EMR to convert the clickstream data to a tabular format. Use the Amazon Redshift COPY command to load the data into the Amazon Redshift cluster.
B) Use AWS Lambda to convert the data to a tabular format and write it to Amazon S3. Use the Amazon Redshift COPY command to load the data into the Amazon Redshift cluster.
C) Use the Relationalize class in an AWS Glue ETL job to transform the data and write the data back to Amazon S3. Use Amazon Redshift Spectrum to create external tables and join with the internal tables.
D) Use the Amazon Redshift COPY command to move the clickstream data directly into new tables in the Amazon Redshift cluster.

2) A publisher website captures user activity and sends clickstream data to Amazon Kinesis Data Streams. The publisher wants to design a cost-effective solution to process the data to create a timeline of user activity within a session. The solution must be able to scale depending on the number of active sessions.

Which solution meets these requirements?

A) Include a variable in the clickstream data from the publisher website to maintain a counter for the number of active user sessions. Use a timestamp for the partition key for the stream. Configure the consumer application to read the data from the stream and change the number of processor threads based upon the counter. Deploy the consumer application on Amazon EC2 instances in an EC2 Auto Scaling group.
B) Include a variable in the clickstream to maintain a counter for each user action during their session. Use the action type as the partition key for the stream. Use the Kinesis Client Library (KCL) in the consumer application to retrieve the data from the stream and perform the processing. Configure the consumer application to read the data from the stream and change the number of processor threads based upon the counter. Deploy the consumer application on AWS Lambda.
C) Include a session identifier in the clickstream data from the publisher website and use as the partition key for the stream. Use the Kinesis Client Library (KCL) in the consumer application to retrieve the data from the stream and perform the processing. Deploy the consumer application on Amazon EC2 instances in an EC2 Auto Scaling group. Use an AWS Lambda function to reshard the stream based upon Amazon CloudWatch alarms.
D) Include a variable in the clickstream data from the publisher website to maintain a counter for the number of active user sessions. Use a timestamp for the partition key for the stream. Configure the consumer application to read the data from the stream and change the number of processor threads based upon the counter. Deploy the consumer application on AWS Lambda.

3) A company is currently using Amazon DynamoDB as the database for a user support application. The company is developing a new version of the application that will store a PDF file for each support case ranging in size from 1–10 MB. The file should be retrievable whenever the case is accessed in the application.

How can the company store the file in the MOST cost-effective manner?

A) Store the file in Amazon DocumentDB and the document ID as an attribute in the DynamoDB table.
B) Store the file in Amazon S3 and the object key as an attribute in the DynamoDB table.
C) Split the file into smaller parts and store the parts as multiple items in a separate DynamoDB table.
D) Store the file as an attribute in the DynamoDB table using Base64 encoding.

4) A company needs to implement a near-real-time fraud prevention feature for its ecommerce site. User and order details need to be delivered to an Amazon SageMaker endpoint to flag suspected fraud. The amount of input data needed for the inference could be as much as 1.5 MB.

Which solution meets the requirements with the LOWEST overall latency?

A) Create an Amazon Managed Streaming for Kafka cluster and ingest the data for each order into a topic. Use a Kafka consumer running on Amazon EC2 instances to read these messages and invoke the Amazon SageMaker endpoint.
B) Create an Amazon Kinesis Data Streams stream and ingest the data for each order into the stream. Create an AWS Lambda function to read these messages and invoke the Amazon SageMaker endpoint.
C) Create an Amazon Kinesis Data Firehose delivery stream and ingest the data for each order into the stream. Configure Kinesis Data Firehose to deliver the data to an Amazon S3 bucket. Trigger an AWS Lambda function with an S3 event notification to read the data and invoke the Amazon SageMaker endpoint.
D) Create an Amazon SNS topic and publish the data for each order to the topic. Subscribe the Amazon SageMaker endpoint to the SNS topic.

5) A media company is migrating its on-premises legacy Hadoop cluster with its associated data processing scripts and workflow to an Amazon EMR environment running the latest Hadoop release. The developers want to reuse the Java code that was written for data processing jobs for the on-premises cluster.

Which approach meets these requirements?

A) Deploy the existing Oracle Java Archive as a custom bootstrap action and run the job on the EMR cluster.
B) Compile the Java program for the desired Hadoop version and run it using a CUSTOM_JAR step on the EMR cluster.
C) Submit the Java program as an Apache Hive or Apache Spark step for the EMR cluster.
D) Use SSH to connect the master node of the EMR cluster and submit the Java program using the AWS CLI.

6) An online retail company wants to perform analytics on data in large Amazon S3 objects using Amazon EMR. An Apache Spark job repeatedly queries the same data to populate an analytics dashboard. The analytics team wants to minimize the time to load the data and create the dashboard.

Which approaches could improve the performance? (Select TWO.)

A) Copy the source data into Amazon Redshift and rewrite the Apache Spark code to create analytical reports by querying Amazon Redshift.
B) Copy the source data from Amazon S3 into Hadoop Distributed File System (HDFS) using s3distcp.
C) Load the data into Spark DataFrames.
D) Stream the data into Amazon Kinesis and use the Kinesis Connector Library (KCL) in multiple Spark jobs to perform analytical jobs.
E) Use Amazon S3 Select to retrieve the data necessary for the dashboards from the S3 objects.

7) A data engineer needs to create a dashboard to display social media trends during the last hour of a large company event. The dashboard needs to display the associated metrics with a consistent latency of less than 2 minutes.

Which solution meets these requirements?

A) Publish the raw social media data to an Amazon Kinesis Data Firehose delivery stream. Use Kinesis Data Analytics for SQL Applications to perform a sliding window analysis to compute the metrics and output the results to a Kinesis Data Streams data stream. Configure an AWS Lambda function to save the stream data to an Amazon DynamoDB table. Deploy a real-time dashboard hosted in an Amazon S3 bucket to read and display the metrics data stored in the DynamoDB table.
B) Publish the raw social media data to an Amazon Kinesis Data Firehose delivery stream. Configure the stream to deliver the data to an Amazon Elasticsearch Service cluster with a buffer interval of 0 seconds. Use Kibana to perform the analysis and display the results.
C) Publish the raw social media data to an Amazon Kinesis Data Streams data stream. Configure an AWS Lambda function to compute the metrics on the stream data and save the results in an Amazon S3 bucket. Configure a dashboard in Amazon QuickSight to query the data using Amazon Athena and display the results.
D) Publish the raw social media data to an Amazon SNS topic. Subscribe an Amazon SQS queue to the topic. Configure Amazon EC2 instances as workers to poll the queue, compute the metrics, and save the results to an Amazon Aurora MySQL database. Configure a dashboard in Amazon QuickSight to query the data in Aurora and display the results.

8) A real estate company is receiving new property listing data from its agents through .csv files every day and storing these files in Amazon S3. The data analytics team created an Amazon QuickSight visualization report that uses a dataset imported from the S3 files. The data analytics team wants the visualization report to reflect the current data up to the previous day.

How can a data analyst meet these requirements?

A) Schedule an AWS Lambda function to drop and re-create the dataset daily.
B) Configure the visualization to query the data in Amazon S3 directly without loading the data into SPICE.
C) Schedule the dataset to refresh daily.
D) Close and open the Amazon QuickSight visualization.

9) A financial company uses Amazon EMR for its analytics workloads. During the company's annual security audit, the security team determined that none of the EMR clusters' root volumes are encrypted. The security team recommends the company encrypt its EMR clusters' root volume as soon as possible.

Which solution would meet these requirements?

A) Enable at-rest encryption for EMR File System (EMRFS) data in Amazon S3 in a security configuration. Re-create the cluster using the newly created security configuration.
B) Specify local disk encryption in a security configuration. Re-create the cluster using the newly created security configuration.
C) Detach the Amazon EBS volumes from the master node. Encrypt the EBS volume and attach it back to the master node.
D) Re-create the EMR cluster with LZO encryption enabled on all volumes.

10) A company is providing analytics services to its marketing and human resources (HR) departments. The departments can only access the data through their business intelligence (BI) tools, which run Presto queries on an Amazon EMR cluster that uses the EMR File System (EMRFS). The marketing data analyst must be granted access to the advertising table only. The HR data analyst must be granted access to the personnel table only.

Which approach will satisfy these requirements?

A) Create separate IAM roles for the marketing and HR users. Assign the roles with AWS Glue resource-based policies to access their corresponding tables in the AWS Glue Data Catalog. Configure Presto to use the AWS Glue Data Catalog as the Apache Hive metastore.
B) Create the marketing and HR users in Apache Ranger. Create separate policies that allow access to the user's corresponding table only. Configure Presto to use Apache Ranger and an external Apache Hive metastore running in Amazon RDS.
C) Create separate IAM roles for the marketing and HR users. Configure EMR to use IAM roles for EMRFS access. Create a separate bucket for the HR and marketing data. Assign appropriate permissions so the users will only see their corresponding datasets.
D) Create the marketing and HR users in Apache Ranger. Create separate policies that allows access to the user's corresponding table only. Configure Presto to use Apache Ranger and the AWS Glue Data Catalog as the Apache Hive metastore.

Answers

1) C – The Relationalize PySpark transform can be used to flatten the nested data into a structured format. Amazon Redshift Spectrum can join the external tables and query the transformed clickstream data in place rather than needing to scale the cluster to accommodate the large dataset.

2) C – Partitioning by the session ID will allow a single processor to process all the actions for a user session in order. An AWS Lambda function can call the UpdateShardCount API action to change the number of shards in the stream. The KCL will automatically manage the number of processors to match the number of shards. Amazon EC2 Auto Scaling will assure the correct number of instances are running to meet the processing load.

3) B – Use Amazon S3 to store large attribute values that cannot fit in an Amazon DynamoDB item. Store each file as an object in Amazon S3 and then store the object path in the DynamoDB item.

4) A – An Amazon Managed Streaming for Kafka cluster can be used to deliver the messages with very low latency. It has a configurable message size that can handle the 1.5 MB payload.

5) B – A CUSTOM JAR step can be configured to download a JAR file from an Amazon S3 bucket and execute it. Since the Hadoop versions are different, the Java application has to be recompiled.

6) C, E – One of the speed advantages of Apache Spark comes from loading data into immutable dataframes, which can be accessed repeatedly in memory. Spark DataFrames organizes distributed data into columns. This makes summaries and aggregates much quicker to calculate. Also, instead of loading an entire large Amazon S3 object, load only what is needed using Amazon S3 Select. Keeping the data in Amazon S3 avoids loading the large dataset into HDFS.

7) A – Amazon Kinesis Data Analytics can query data in a Kinesis Data Firehose delivery stream in near-real time using SQL. A sliding window analysis is appropriate for determining trends in the stream. Amazon S3 can host a static webpage that includes JavaScript that reads the data in Amazon DynamoDB and refreshes the dashboard.

8) C – Datasets created using Amazon S3 as the data source are automatically imported into SPICE. The Amazon QuickSight console allows for the refresh of SPICE data on a schedule.

9) B – Local disk encryption can be enabled as part of a security configuration to encrypt root and storage volumes.

10) A – AWS Glue resource policies can be used to control access to Data Catalog resources.