Answer Sheet

Q1)

A Company has two batch processing applications that consume financial data about the day's stock transactions.

Each transaction needs to be stored durably and guarantee that a record of each application is delivered so the audit and billing batch processing applications can process the data.

However, the two applications run separately and several hours apart and need access to the same transaction information. After reviewing the transaction information for the day, the information no longer needs to be stored.

What is the best way to architect this application? Choose the correct answer from the options below

✅ Use Kinesis to store the transaction information. The billing application will consume data from the stream, the audit application can consume the same data several hours later.
⚪ Store the transaction information in a DynamoDB table. The billing application can read the rows while the audit application will read the rows them remove the data.
⚪ Use SQS for storing the transaction messages; when the billing batch process performs first and consumes the message, write the code in a way that does not remove the message after consumed, so it is available for the audit application several hours later. The audit application can consume the SQS message and remove it from the queue when completed.
⚪ Use SQS for storing the transaction messages. When the billing batch process consumes each message, have the application create an identical message and place it in a different SQS for the audit application to use several hours later.

Q2)

A research scientist is planning for the one-time launch of an Elastic MapReduce cluster and is encouraged by her manager to minimize the costs.

The cluster is designed to ingest 200TB of genomics data with a total of 100 Amazon EC2 instances and is expected to run for around four hours.

The resulting data set must be stored temporarily until archived into an Amazon RDS Oracle instance.

Which option will help save the most money while meeting requirements?

⚪ Deploy on-demand master, core and task nodes and store ingest and output files in Amazon S3 RRS
⚪ Store the ingest files in Amazon S3 RRS and store the output files in S3. Deploy Reserved Instances for the master and core nodes and on-demand for the task nodes.
⚪ Optimize by deploying a combination of on-demand, RI and spot-pricing models for the master, core and task nodes. Store ingest and output files in Amazon S3 with a lifecycle policy that archives them to Amazon Glacier.
✅ Store ingest and output files in Amazon S3. Deploy on-demand for the master and core nodes and spot for the task nodes.

Q3) TPT Limited stores millions of sensitive transactions across thousands of 100-GB files which should be encrypted during transit and at rest for compliance purposes. For business decision analysts often require subsets of files with up to 5TB of space for generating simulations. Which of the given AWS solution cost effectively accommodate the long-term storage and in-flight subsets of data?

⚪ Use HDFS on Amazon Elastic MapReduce (EMR), and run simulations on subsets in-memory on Amazon Elastic Compute Cloud (EC2).
⚪ Use HDFS on Amazon EMR, and run simulations on subsets in ephemeral drives on Amazon EC2.
⚪ Use Amazon S3 with server-side encryption, and run simulations on subsets in-memory on Amazon EC2.
✅ Use Amazon Simple Storage Service (S3) with server-side encryption, and run simulations on subsets in ephemeral drives on Amazon EC2.
⚪ None of these
⚪ Store the full data set in encrypted Amazon Elastic Block Store (EBS) volumes, and regularly capture snapshots that can be cloned to EC2 workstations

---

Q4)

A customer's nightly EMR job processes a single 2-TB data file stored on Amazon Simple Storage Service (S3).

The Amazon Elastic Map Reduce (EMR) job runs on two On-Demand core nodes and three On-Demand task nodes.

Which of the following may help reduce the EMbR job completion time? Choose 2 answers

✅ Adjust the number of simultaneous mapper tasks.
⚪ Launch the core nodes and task nodes within an Amazon Virtual Cloud.
⚪ Use a bootstrap action to present the S3 bucket as a local filesystem.
✅ Change the input split size in the MapReduce job configuration.
⚪ Use three Spot Instances rather than three On-Demand instances for the task nodes.
⚪ Enable termination protection for the job flow.

Q5)

Your company is in the process of developing a next generation pet collar that collects biometric information to assist families with promoting healthy lifestyles for their pets.

Each collar will push 30kb of biometric data In JSON format every 2 seconds to a collection platform that will process and analyze the data providing health trending information back to the pet owners and veterinarians via a web portal Management has tasked you to architect the collection platform ensuring the following requirements are met.

Provide the ability for real-time analytics of the inbound biometric data to ensure processing of the biometric data is highly durable, Elastic and parallel. The results of the analytic processing should be persisted for data mining.

Which architecture outlined below will meet the initial requirements for the collection platform?

- Utilize EMR to collect the inbound sensor data, analyze the data from EMR with Amazon Kinesis and save the results to DynamoDB.
- Utilize SQS to collect the inbound sensor data analyze the data from SQS with Amazon Kinesis and save the results to a Microsoft SQL Server RDS instance.
- ✅ Utilize Amazon Kinesis to collect the inbound sensor data, analyze the data with Kinesis clients and save the results to a Redshift cluster using EMR.
- Utilize S3 to collect the inbound sensor data analyze the data from S3 with a daily scheduled Data Pipeline and save the results to a Redshift Cluster.

Q6)

A company is developing a video application that will emit a log stream. Each record in the stream may contain up to 400 KB of data.

To improve the video-streaming experience, it is necessary to collect a subset of metrics from the stream to be analyzed for trends over time using complex SQL queries.

A Solutions Architect will create a solution that allows the application to scale without customer interaction.

Which solution should be implemented to meet these requirements?

- Send the log data to an Amazon Kinesis data stream. Subscribe an AWS Lambda function to the stream that transforms the data and sends it to a second data stream. Use Amazon Kinesis Data Analytics to query the data in the second stream.
- ✅ Send the log data to an Amazon CloudWatch Logs log group. Make the log group an event source for an AWS Lambda function that transforms the data and stores it in an Amazon S3 bucket. Query the data with Amazon Athena.
- Send the log data to an Amazon SQS standard queue. Make the queue an event source for an AWS Lambda function that transforms the data and stores it in Amazon Redshift. Query the data in Amazon Redshift.
- Send the log data to an Amazon Kinesis Data Firehose delivery stream. Use an AWS Lambda function to transform the data. Deliver the data to Amazon Redshift. Query the data in Amazon Redshift.

Q7) TPT Limited serves on-demand training videos in high resolution MP4 format, to its employees. All the employees are distributed globally and usually on the move with company-provided tablets needing the HTTP Live Streaming (HLS) protocol to watch a video. TPT Limited has no experience in transcoding. Which of the following will be the most cost-efficient architecture without compromising high availability and quality of video delivery?

- ✅ Elastic Transcoder to transcode original high-resolution MP4 videos to HLS. S3 to host videos with lifecycle Management to archive original flies to Glacier after a few days. CloudFront to serve HLS transcoded videos from S3
- A video transcoding pipeline running on EC2 using SQS to distribute tasks and Auto Scaling to adjust the number or nodes depending on the length of the queue S3 to host videos with Lifecycle Management to archive all files to Glacier after a few days CloudFront to serve HLS transcoding videos from Glacier
- Elastic Transcoder to transcode original high-resolution MP4 videos to HLS EBS volumes to host videos and EBS snapshots to incrementally backup original rues after a few days. CloudFront to serve HLS transcoded videos from EC2.
- A video transcoding pipeline running on EC2 using SQS to distribute tasks and Auto Scaling to adjust the number of nodes depending on the length of the queue. EBS volumes to host videos and EBS snapshots to incrementally backup original files after a few days. CloudFront to serve HLS transcoded videos from EC2
- None of these

Q8)

Your company releases new features with high frequency while demanding high application availability.

As part of the application's A/B testing, logs from each updated Amazon EC2 instance of the application need to be analyzed in near real-time, to ensure that the application is working flawlessly after each deployment.

If the logs show any anomalous behavior, then the application version of the instance is changed to a more stable one.

Which of the following methods should you use for shipping and analyzing the logs in a highly available manner?

- ✅ Ship the logs to an Amazon Kinesis stream and have the consumers analyze the logs in a live manner.
- Ship the logs to Amazon CloudWatch Logs and use Amazon EMR to analyze the logs in a batch manner each hour.
- Ship the logs to Amazon S3 for durability and use Amazon EMR to analyze the logs in a batch manner each hour.
- Ship the logs to a large Amazon EC2 instance and analyze the logs in a live manner.

Q9)

A company needs to deploy a data lake solution for their data scientists in which all company data is accessible and stored in a central S3 bucket.

The company segregates the data by business unit, using specific prefixes. Scientists can only access the data from their own business unit. The company needs a single sign-on identity and management solution based on Microsoft Active Directory (AD) to manage access to the data in Amazon S3.

Which method meets these requirements?

- Use Amazon S3 API integration with AD to impersonate the users on access in a transparent manner.
- Deploy the AD Synchronization service to create AWS IAM users and groups based on AD information.
- Create bucket policies that only allow access to the authorized prefixes based on the users' group name in Active Directory.
- ✅ Use AWS IAM Federation functions and specify the associated role based on the users' groups in AD.

Q10) TPT Limited deployed a web application that emits events to Amazon Kinesis Streams for operational reporting. Both critical and informational events are reported. James has been asked to suggest a solution that meets the following requirements -

1. Capturing critical events immediately before processing continues
2. Informational events should not delay processing
Which of the following solutions will help meet the companies goal?

- Log all events using the PutRecords API method.
- None of these
- ✅ Log critical events using the PutRecords API method, and log informational events using the Kinesis Producer Library.
- Log critical events using the Kinesis Producer Library, and log informational events using the PutRecords API method.
- Log all events using the Kinesis Producer Library.

Q11)

A data engineer needs to collect data from multiple Amazon Redshift clusters within a business and consolidate the data into a single central data warehouse. Data must be encrypted at all times while at rest or in flight.

What is the most scalable way to build this data collection process?

- Connect to the source cluster over an SSL client connection, and write data records to Amazon Kinesis Firehose to load into your target data warehouse.
- Run an UNLOAD command that stores the data in an S3 bucket encrypted with an AWS KMS data key; run a COPY command to move the data into the target cluster.
- ✅ Use AWS KMS data key to run an UNLOAD ENCRYPTED command that stores the data in an unencrypted S3 bucket; run a COPY command to move the data into the target cluster.
- Run an ETL process that connects to the source clusters using SSL to issue a SELECT query for new data, and then write to the target data warehouse using an INSERT command over another SSL secured connection.

Q12)

A data engineer needs to architect a data warehouse for an online retail company to store historic purchases.

The data engineer needs to use Amazon Redshift. To comply with PCI:DSS and meet corporate data protection standards, the data engineer must ensure that data is encrypted at rest and that the keys are managed by a corporate on-premises HSM.

Which approach meets these requirements in the most cost-effective manner?

- Use AWS Import/Export to import the corporate HSM device into the AWS Region where the Amazon Redshift cluster will launch, and configure Redshift to use the imported HSM.
- Configure the AWS Key Management Service to point to the corporate HSM device, and then launch the Amazon Redshift cluster with the KMS managing the encryption keys.
- Use the AWS CloudHSM service to establish a trust relationship between the CloudHSM and the corporate HSM over a Direct Connect connection. Configure Amazon Redshift to use the CloudHSM device.
- ✅ Create a VPC, and then establish a VPN connection between the VPC and the on-premises network. Launch the Amazon Redshift cluster in the VPC, and configure it to use your corporate HSM.

Q13) Robert is a data engineer working at TPT Limited. He has been asked to design a data processing platform which will receives a huge volume of unstructured data. To handle the data Robert has been asked to populate a well-structured star schema in Amazon Redshift. Which of the following is most efficient architecture strategy that will help meet the requirement?

- When the data is saved to Amazon S3, use S3 Event Notifications and AWS Lambda to transform the file contents. Insert the data into the analysis schema on Redshift.
- Normalize the data using an AWS Marketplace ETL tool, persist the results to Amazon S3, and use AWS Lambda to INSERT the data into Redshift.
- Load the unstructured data into Redshift, and use string parsing functions to extract structured data for inserting into the analysis schema.
- ✅ Transform the unstructured data using Amazon EMR and generate CSV datCOPY the CSV data into the analysis schema within Redshift.
- None of these

Q14)

A company has several teams of analysts. Each team of analysts has their own cluster.

The teams need to run SQL queries using Hive, Spark-SQL, and Presto with Amazon EMR.

The company needs to enable a centralized metadata layer to expose the Amazon S3 objects as tables to the analysts.

Which approach meets the requirement for a centralized metadata layer?

- s3distcp with the output Manifest option to generate RDS DDL
- ✅ Bootstrap action to change the Hive Metastore to an Amazon RDS database
- EMRFS consistent view with a common Amazon DynamoDB table
- Naming scheme support with automatic partition discovery from Amazon S3

Q15) TPT Limited has deployed an Amazon Redshift data warehouse for different user teams that queries the same table but with very different query types. Recently the user teams are facing issues of poor performance. Which of the given option will help in improving the performance for the user teams?

- Add support for workload management queue hopping.
- Maintain team-specific copies of the table.
- None of these
- ✅ Add interleaved sort keys per team.
- Create custom table views.

Q16)

An administrator needs to design the event log storage architecture for events from mobile devices.

The event data will be processed by an Amazon EMR cluster daily for aggregated reporting and analytics before being archived.

How should the administrator recommend storing the log data?

- Create an Amazon DynamoDB table partitioned on EventID, write log data to table. Execute the EMR job on the table.
- ✅ Create an Amazon S3 bucket and write data into folders by day. Execute the EMR job on the daily folder.
- Create an Amazon DynamoDB table partitioned on the device and sorted on date, write log data to table. Execute the EMR job on the Amazon DynamoDB table.
- Create an Amazon S3 bucket and write log data into folders by device. Execute the EMR job on the device folders.

Q17)

A company is building a new application in AWS. The architect needs to design a system to collect application log events.

The design should be a repeatable pattern that minimizes data loss if an application instance fails, and keeps a durable copy of a log data for at least 30 days.

What is the simplest architecture that will allow the architect to analyze the logs?

- Write them to CloudWatch Logs and use an AWS Lambda function to load them into HDFS on an Amazon Elastic MapReduce (EMR) cluster for analysis.
- Write them to the local disk and configure the Amazon CloudWatch Logs agent to load the data into CloudWatch Logs and subsequently into Amazon Elasticsearch Service.
- Write them to a file on Amazon Simple Storage Service (S3). Write an AWS Lambda function that runs in response to the S3 event to load the events into Amazon Elasticsearch Service for analysis.
- ✅ Write them directly to a Kinesis Firehose. Configure Kinesis Firehose to load the events into an Amazon Redshift cluster for analysis.

Q18)

A system needs to collect on-premises application spool files into a persistent storage layer in AWS. Each spool file is 2 KB.

The application generates 1 M files per hour. Each source file is automatically deleted from the local server after an hour.

What is the most cost-efficient option to meet these requirements?

- Copy files to Amazon S3 infrequent Access Storage.
- Write file contents to Amazon ElastiCache.
- Copy files to Amazon S3 Standard Storage.
- ✅ Write file contents to an Amazon DynamoDB table.

Q19)

A game company needs to properly scale its game application, which is backed by DynamoDB.

Amazon Redshift has the past two years of historical data. Game traffic varies throughout the year based on various factors such as season, movie release, and holiday season.

An administrator needs to calculate how much read and write throughput should be provisioned for DynamoDB table for each week in advance.

How should the administrator accomplish this task?

- Feed the data into Amazon Machine Learning and build a binary classification model.
- Feed the data into Apache Mahout and build a multi-classification model.
- Feed the data into Spark Mlib and build a random forest model.
- ✅ Feed the data into Amazon Machine Learning and build a regression model.

Q20)

A solutions architect for a logistics organization ships packages from thousands of suppliers to end customers.

The architect is building a platform where suppliers can view the status of one or more of their shipments.

Each supplier can have multiple roles that will only allow access to specific fields in the resulting information.

Which strategy allows the appropriate level of access control and requires the LEAST amount of management work?

- Send the tracking data to Amazon Kinesis Firehose. Store the data in an Amazon Redshift cluster. Create views for the suppliers, users and roles. Allow suppliers access to the Amazon Redshift cluster using a user limited to the applicable view.
- Send the tracking data to Amazon Kinesis Streams. Use Amazon EMR with Spark Streaming to store the data in HBase. Create one table per supplier. Use HBase Kerberos integration with the suppliers? users. Use HBase ACL-based security to limit access for the roles to their specific table and columns.
- Send the tracking data to Amazon Kinesis Firehose. Use Amazon S3 notifications and AWS Lambda to prepare files in Amazon S3 with appropriate data for each supplier?s roles. Generate temporary AWS credentials for the suppliers? users with AWS STS. Limit access to the appropriate files through security policies.
- ✅ Send the tracking data to Amazon Kinesis Streams. Use AWS Lambda to store the data in an Amazon DynamoDB Table. Generate temporary AWS credentials for the suppliers? users with AWS STS, specifying fine-grained security policies to limit access only to their applicable dat(Correct)

Q21)

A data engineer wants to use an Amazon Elastic Map Reduce for an application.

The data engineer needs to make sure it complies with regulatory requirements.

The auditor must be able to confirm at any point which servers are running and which network access controls are deployed.

Which action should the data engineer take to meet this requirement?

- Provide the auditor with CloudFormation templates.
- Provide the auditor with SSH keys for access to the Amazon EMR cluster.
- ✅ Provide the auditor IAM accounts with the SecurityAudit policy attached to their group.
- Provide the auditor with access to AWS DirectConnect to use their existing tools.

Q22)

An online photo album app has a key design feature to support multiple screens (e.g, desktop, mobile phone, and tablet) with high-quality displays. Multiple versions of the image must be saved in different resolutions and layouts.

The image-processing Java program takes an average of five seconds per upload, depending on the image size and format.

Each image upload captures the following image metadata: user, album, photo label, upload timestamp.

The app should support the following requirements:

- Hundreds of user image uploads per second

- Maximum image upload size of 10 MB

- Maximum image metadata size of 1 KB

Image displayed in optimized resolution in all supported screens no later than one minute after image upload

Which strategy should be used to meet these requirements?

- ✅ Upload image with metadata to Amazon S3, use Lambda function to run the image processing and save the images output to Amazon S3 and metadata to the app repository DB.
- Write image and metadata to Amazon Kinesis. Use Amazon Elastic MapReduce (EMR) with Spark Streaming to run image processing and save the images output to Amazon S3 and metadata to app repository DB.
- Write images and metadata to Amazon Kinesis. Use a Kinesis Client Library (KCL) application to run the image processing and save the image output to Amazon S3 and metadata to the app repository DB.
- Write image and metadata to RDS with BLOB data type. Use AWS Data Pipeline to run the image processing and save the image output to Amazon S3 and metadata to the app repository DB.

Q23)

An online gaming company uses DynamoDB to store user activity logs and is experiencing throttled writes on the company?s DynamoDB table.

The company is NOT consuming close to the provisioned capacity. The table contains a large number of items and is partitioned on user and sorted by date.

The table is 200 GB and is currently provisioned at 10 K WCU and 20 K RCU.

Which two additional pieces of information are required to determine the cause of the throttling? (Choose two.)

- ✅ The structure of any LSIs that have been defined on the table
- Application-level metrics showing the average item size and peak update rates for each attribute
- ✅ CloudWatch data showing consumed and provisioned write capacity when writes are being throttled
- The structure of any GSIs that have been defined on the table
- The maximum historical WCU and RCU for the table

Q24)

A city has been collecting data on its public bicycle share program for the past three years. The 5PB dataset currently resides on Amazon S3.

The data contains the following datapoints:

- Bicycle origination points

- Bicycle destination points

- Mileage between the points

- Number of bicycle slots available at the station (which is variable based on the station location)

- Number of slots available and taken at a given time

The program has received additional funds to increase the number of bicycle stations available. All data is regularly archived to Amazon Glacier.

The new bicycle stations must be located to provide the most riders access to bicycles. How should this task be performed?

- ✅ Keep the data on Amazon S3 and use an Amazon EMR-based Hadoop cluster with spot instances to run a Spark job that performs a stochastic gradient descent optimization over EMRFS.
- Persist the data on Amazon S3 and use a transient EMR cluster with spot instances to run a Spark streaming job that will move the data into Amazon Kinesis.
- Use the Amazon Redshift COPY command to move the data from Amazon S3 into Redshift and perform a SQL query that outputs the most popular bicycle stations.

Move the data from Amazon S3 into Amazon EBS-backed volumes and use an EC2 based Hadoop cluster with spot instances to run a Spark job that performs a stochastic gradient descent optimization.

Q25)

An administrator is deploying Spark on Amazon EMR for two distinct use cases: machine learning algorithms and ad-hoc querying.

All data will be stored in Amazon S3. Two separate clusters for each use case will be deployed. The data volumes on Amazon S3 are less than 10 GB.

How should the administrator align instance types with the cluster's purpose?

- Machine Learning on T instance types and ad-hoc queries on M instance types
- Machine Learning on R instance types and ad-hoc queries on G2 instance types
- ✓ Machine Learning on C instance types and ad-hoc queries on R instance types
- Machine Learning on D instance types and ad-hoc queries on I instance types

---

Q26)

A large grocery distributor receives daily depletion reports from the field in the form of gzip archives of CSV files uploaded to Amazon S3.

The files range from 500MB to 5GB. These files are processed daily by an EMR job. Recently it has been observed that the file sizes vary, and the EMR jobs take too long. The distributor needs to tune and optimize the data processing workflow with this limited information to improve the performance of the EMR job.

Which recommendation should an administrator provide?

- Decompress the gzip archives and store the data as CSV files.
- ✓ Use bzip2 or Snappy rather than gzip for the archives.
- Reduce the HDFS block size to increase the number of task processors.
- Use Avro rather than gzip for the archives.

Q27)

Your application generates a 1 KB JSON payload that needs to be queued and delivered to EC2 instances for applications.

At the end of the day, the application needs to replay the data for the past 24 hours. In the near future, you also need the ability for other multiple EC2 applications to consume the same stream concurrently.

What is the best solution for this?

- Kinesis Firehose
- ✓ Kinesis Data Streams
- SNS
- SQS

Q28)

An administrator needs to design a distribution strategy for a star schema in a Redshift cluster.

The administrator needs to determine the optimal distribution style for the tables in the Redshift schema.

In which three circumstances would choosing key-based distribution be most appropriate? (Select three.)

- ✓ When the administrator needs to take advantage of data locality on a local node for joins and aggregates.
- ✓ When the administrator needs to balance data distribution and collocation data.
- When the administrator needs to optimize the fact table for parity with the number of slices.
- ✓ When the administrator needs to reduce cross-node traffic.
- When the administrator needs to optimize a large, slowly changing dimension table.

Q29)

An administrator receives about 100 files per hour into Amazon S3 and will be loading the files into Amazon Redshift.

Customers who analyze the data within Redshift gain significant value when they receive data as quickly as possible. The customers have agreed to a maximum loading interval of 5 minutes.

Which loading approach should the administrator use to meet this objective?

- Load the cluster when the number of files is less than the Cluster Slice Count.
- ✓ Load the cluster when the administrator has the number of files as multiple of files relative to Cluster Slice Count, or 5 minutes, whichever comes first.
- Load the cluster as soon as the administrator has the same number of files as nodes in the cluster.
- Load each file as it arrives because getting data into the cluster as quickly as possibly is the priority.

Q30)

A clinical trial will rely on medical sensors to remotely assess patient health. Each physician who participates in the trial requires visual reports each morning. The reports are built from aggregations of all the sensor data taken each minute.

What is the most cost-effective solution for creating this visualization each day?

- Use an EMR cluster to aggregate the patient sensor data each night and provide Zeppelin notebooks that look at the new data residing on the cluster each morning for the physician to review.
- Use Spark streaming on EMR to aggregate the patient sensor data in every 15 minutes and generate a QuickSight visualization on the new data each morning for the physician to review.
- ✅ Use a transient EMR cluster that shuts down after use to aggregate the patient sensor data each night and generate a QuickSight visualization on the new data each morning for the physician to review.
- Use Kinesis Aggregators Library to generate reports for reviewing the patient sensor data and generate a QuickSight visualization on the new data each morning for the physician to review.

Q31)

A company generates a large number of files each month and needs to use AWS import/export to move these files into Amazon S3 storage.

To satisfy the auditors, the company needs to keep a record of which files were imported into Amazon S3.

What is a low-cost way to create a unique log for each import job?

- Use the log file checksum in the import/export manifest files to create a unique log file in Amazon S3 for each import.
- ✅ Use the log file prefix in the import/export manifest files to create a unique log file in Amazon S3 for each import.
- Use the same log file prefix in the import/export manifest files to create a versioned log file in Amazon S3 for all imports.
- Use a script to iterate over files in Amazon S3 to generate a log after each import/export job.

Q32)

A company hosts a portfolio of e-commerce websites across the Oregon, N. Virginia, Ireland, and Sydney AWS regions.

Each site keeps log files that capture user behavior. The company has built an application that generates batches of product recommendations with collaborative filtering in Oregon.

Oregon was selected because the flagship site is hosted there and provides the largest collection of data to train machine learning models against. The other regions do NOT have enough historic data to train accurate machine learning models.

Which set of data processing steps improves recommendations for each region?

- ✅ Use the CloudWatch Logs agent to consolidate logs into a single CloudWatch Logs group.
- Use Kinesis as a buffer for web logs and replicate logs to the Kinesis stream of a neighboring region.
- Use Amazon S3 bucket replication to consolidate log entries and build a single model in Oregon.
- Use the e-commerce application in Oregon to write replica log files in each other region.

Q33)

A company's social media manager requests more staff on the weekends to handle an increase in customer contacts from a particular region.

The company needs a report to visualize the trends on weekends over the past 6 months using QuickSight.

How should the data be represented?

- A map of regions with a heatmap overlay to show the volume of customer contacts
- A pie chart per region plotting customer contacts per day of week
- ✅ A line graph plotting customer contacts vs. time, with a line for each region
- A bar graph plotting region vs. volume of social media contacts

---

Q34)

Company A operates in Country X. Company A maintains a large dataset of historical purchase orders that contains personal data of their customers in the form of full names and telephone numbers. The dataset consists of 5 text files, 1TB each.

Currently the dataset resides on-premises due to legal requirements of storing personal data in-country. The research and development department needs to run a clustering algorithm on the dataset and wants to use Elastic Map Reduce service in the closest AWS region.

Due to geographic distance, the minimum latency between the on-premises system and the closet AWS region is 200 ms.

Which option allows Company A to do clustering in the AWS Cloud and meet the legal requirement of maintaining personal data in-country?

- Use AWS Import/Export Snowball device to securely transfer the data to the AWS region and copy the files onto an EBS volume. Have the EMR cluster read the dataset using EMRFS.
- Encrypt the data files according to encryption standards of Country X and store them on AWS region in Amazon S3. Have the EMR cluster read the dataset using EMRFS.
- Establish a Direct Connect link between the on-premises system and the AWS region to reduce latency. Have the EMR cluster read the data directly from the on-premises storage system over Direct Connect.
- ✅ Anonymize the personal data portions of the dataset and transfer the data files into Amazon S3 in the AWS region. Have the EMR cluster read the dataset using EMRFS.

Q35)

A web-hosting company is building a web analytics tool to capture clickstream data from all of the websites hosted within its platform and to provide near-real-time business intelligence. This entire system is built on AWS services. The web-hosting company is interested in using Amazon Kinesis to collect this data and perform sliding window analytics.

What is the most reliable and fault-tolerant technique to get each website to send data to Amazon Kinesis with every click?

✅ After receiving a request, each web server sends it to Amazon Kinesis using the Amazon Kinesis PutRecord API. Use the exponential back-off algorithm for retries until a successful response is receive

⚪ Each web server buffers the requests until the count reaches 500 and sends them to Amazon Kinesis using the Amazon Kinesis PutRecord API.

⚪ After receiving a request, each web server sends it to Amazon Kinesis using the Amazon Kinesis Producer Library addRecords method.

⚪ After receiving a request, each web server sends it to Amazon Kinesis using the Amazon Kinesis PutRecord API. Use the sessionID as a partition key and set up a loop to retry until a success response is received.

Q36)

You need to analyze clickstream data on your website from multiple applications.

You want to analyze the pattern of pages a consumer clicks on and in what order.

You need to be able to use the data in real time and want to manage as little infrastructure as possible.

Which option would meet this requirement?

✅ Use Amazon Kinesis with a worker to process the data received from the Kinesis stream.

⚪ Use Elastic MapReduce to ingest the data and analyze it.

⚪ Publish web clicks by session to an Amazon SQS.

⚪ Send click events directly to Amazon Redshift and then analyze them with SQL.

Q37)

Your client has a high-volume DynamoDB table that serves comment information to an internal API.

Currently, the table allows you to query with a composite primary key with postId as a partition key and commentId as a sort key. Application validation ensures that each item has other fields including timestamp, userId, and sentimentScore. The client has several long-running users, and they would like to provide more effective ways of surfacing posts from them from different time frames.

How might the client enable this sort of functionality?

⚪ Create a Local Secondary Index with a partition key of userId and a sort key of timestamp.

⚪ Create a Local Secondary Index with a partition key of timestamp and a sort key of userId.

✅ Create a Global Secondary Index with a partition key of userId and a sort key of timestamp.

⚪ Create a Global Secondary Index with a partition key of timestamp and a sort key of userId.

Q38)

You have a JSON data file in S3 that you are attempting to load into a JavaScript visualization you are writing locally. **This visualization makes an HTTP GET request to the S3 location that fails.**

However, when you attempt to visit the URL being requested by the JavaScript directly from inside your browser, it seems to be loading fine.

You are also using a private/incognito window and are not signed into the AWS console. What is the most likely issue?

⚪ The IAM role you used to create and upload the JSON data in the S3 bucket is preventing the JavaScript from loading the file.

⚪ The bucket policies are preventing the JavaScript from loading the file.

✅ The CORS settings are preventing the JavaScript from loading the file.

⚪ The ACLs on the bucket are preventing the JavaScript from loading the file.

Q39)

You have to design an EMR system where you will be processing highly confidential data.

What can you do to ensure encryption of data at rest?

⚪ TLS

⚪ VPN

✅ SSE-KMS

✅ LUKS
