Answer Sheet

Q1)

You have been asked to cost optimize a business critical and long-running EMR cluster.

The EMR cluster is currently on-demand for the master nodes, core nodes and task nodes.

The costs for running the cluster have been steadily increasing as nodes have been added and resized.

What would you suggest the business does to reduce the costs without requiring any long-term commitment?

- ✅ Leave the master and core nodes as on-demand and use spot instances for the task nodes
- Leave all nodes running on-demand instances, the cluster is already cost optimized.
- Leave the master node to use on-demand and change the core and task nodes to spot
- Recreate the cluster using spot instances for the master, core and task nodes.

Q2)

You have just joined a new company and have been put in charge of EC2 instances and any other services that use EC2 instances.

You notice that the company has been slow to take advantage of AWS per-second Billing, specifically in the area of EMR and Spot Instances.

What immediate steps can you take on EMR with spot instances to improve cost saving and performance?

- Run fewer instances for a longer amount of time.
- Run fewer instances for a shorter amount of time.
- Use on-demand instances instead.
- ✅ Run more instances for a shorter amount of time.

Q3)

You are designing a service that aggregates clickstream data in batch and delivers reports to subscribers via email only once per week.

Data is extremely spikey, geographically distributed, high-scale, and unpredictable.

How should you design this system?

- Use AWS Elasticsearch service and EC2 Auto Scaling groups. The Autoscaling groups scale based on click throughput and stream into the Elasticsearch domain, which is also scalable. Use Kibana to generate reports periodically.
- Use API Gateway invoking Lambdas which PutRecords into Kinesis, and EMR running Spark performing GetRecords on Kinesis to scale with spikes. Spark on EMR outputs the analysis to S3, which are sent out via email.
- ✅ Use a CloudFront distribution with access log delivery to S3. Clicks should be recorded as query string GETs to the distribution. Reports are built and sent by periodically running EMR jobs over the access logs in S3.
- Use a large Redshift cluster to perform the analysis, and a fleet of Lambdas to perform record inserts into the Redshift tables. Lambda will scale rapidly enough for the traffic spikes.

Q4)

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.

Q5)

A customer needs to load a 550-GB data file into an Amazon Redshift cluster from Amazon S3, using the COPY command.

The input file has both known and unknown issues that will probably cause the load process to fail.

The customer needs the most efficient way to detect load errors without performing any cleanup if the load process fails.

Which technique should the customer use?

- Write a script to delete the data from the tables in case of errors.
- ✅ Use COPY with NOLOAD parameter.
- Split the input file into 50-GB blocks and load them separately.
- Compress the input file before running COPY.

Q6)

An administrator decides to use the Amazon Machine Learning service to classify social media posts that mention your company into two categories: posts that require a response and posts that do not.

The training dataset of 10,000 posts contains the details of each post, including the timestamp, author, and full text of the post.

You are missing the target labels that are required for training. Which two options will create valid target label data?

- Using the a priori probability distribution of the two classes, use Monte-Carlo simulation to generate the labels.
- ✅ Use the Amazon Mechanical Turk web service to publish Human Intelligence Tasks that ask Turk workers to label the posts.
- Use the sentiment analysis NLP library to determine whether a post requires a response.
- ✅ Ask the social media handling team to review each post and provide the label.

Q7)

A company logs data from its application in large files and runs regular analytics of these logs to support internal reporting for three months after the logs are generated.

After three months, the logs are infrequently accessed for up to a year. The company also has a regulatory control requirement to store application logs for seven years.

Which course of action should the company take to achieve these requirements in the most cost-efficient way?

- Store the files in S3 Standard with a lifecycle policy to remove them after a year. Simultaneously store the files in Amazon S3 Glacier with a Deny Delete vault lock policy for archives less than seven years old.
- ✅ Store the files in S3 Standard with lifecycle policies to transition the storage class to Standard ? IA after three months and delete them after a year. Simultaneously store the files in Amazon Glacier with a Deny Delete vault lock policy for archives less than seven years old.
- Store the files in S3 Standard with a lifecycle policy to transition the storage class to Standard - IA after three months. After a year, transition the files to Glacier and add a Deny Delete vault lock policy for archives less than seven years old.
- Store the files in S3 Glacier with a Deny Delete vault lock policy for archives less than seven years old and a vault access policy that restricts read access to the analytics IAM group and write access to the log writer service role.

Q8)

A new algorithm has been written in Python to identify SPAM e-mails.

The algorithm analyzes the free text contained within a sample set of 1 million e-mails stored on Amazon S3.

The algorithm must be scaled across a production dataset of 5 PB, which also resides in Amazon S3 storage.

Which AWS service strategy is best for this use case?

- Initiate a Python job from AWS Data Pipeline to run directly against the Amazon S3 text files.
- Use Amazon Elasticsearch Service to store the text and then use the Python Elasticsearch Client to run analysis against the text index.
- ✅ Use Amazon EMR to parallelize the text analysis tasks across the cluster using a streaming program step.
- Copy the data into Amazon ElastiCache to perform text analysis on the in-memory data and export the results of the model into Amazon Machine Learning.

Q9)

An administrator needs to manage a large catalog of items from various external sellers.

The administrator needs to determine if the items should be identified as minimally dangerous, dangerous, or highly dangerous based on their textual descriptions.

The administrator already has some items with the danger attribute, but receives hundreds of new item descriptions every day without such classification.

The administrator has a system that captures dangerous goods reports from customer support team or from user feedback.

What is a cost-effective architecture to solve this issue?

- Build a machine learning model with binary classification for dangerous goods and run it on the DynamoDB Streams as every new item description is added to the system.
- ✅ Build a machine learning model to properly classify dangerous goods and run it on the DynamoDB Streams as every new item description is added to the system.
- Build a Kinesis Streams process that captures and marks the relevant items in the dangerous goods reports using a Lambda function, once more than two reports have been filed.
- Build a set of regular expression rules that are based on the existing examples, and run them on the DynamoDB Streams as every new item description is added to the system.

Q10)

A customer is collecting clickstream data using Amazon Kinesis and is grouping the events by IP address into 5-minute chunks stored in Amazon S3.

Many analysts in the company use Hive on Amazon EMR to analyze this data. Their queries always reference a single IP address. Data must be optimized for querying based on IP address using Hive running on Amazon EMR.

What is the most efficient method to query the data with Hive?

- Store the events for an IP address as a single file in Amazon S3 and add metadata with keys:Hive_Partitioned_IPAddress.
- Store the data in an HBase table with the IP address as the row key.
- ✅ Store the Amazon S3 objects with the following naming scheme bucket_name/source=ip_address/year=yy/month=mm/day=dd/hour=hh/filename.
- Store an index of the files by IP address in the Amazon DynamoDB metadata store for EMRFS.

Q11)

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.

Q12)

A company is centralizing a large number of unencrypted small files from multiple Amazon S3 buckets.

The company needs to verify that the files contain the same data after centralization.

Which method meets the requirements?

- Compare the size of the source and destination objects.
- Place a HEAD request against the source and destination objects comparing SIG v4.
- Call the S3 CompareObjects API for the source and destination objects.
- ✅ Compare the S3 Etags from the source and destination objects.

Q13)

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.

Q14) TPT Limited generates a large number of files every month which are to be imported into Amazon S3 storage using AWS import/export. Which of the following is a low-cost way to create a unique log for each import job at TPT Limited for compliance purposes?

- Using a script to iterate over files in Amazon S3 to generate a log after each import/export job.
- Using the log file checksum in the import/export manifest files to create a unique log file in Amazon S3 for each import.
- ✅ Using the log file prefix in the import/export manifest files to create a unique log file in Amazon S3 for each import.
- None of these
- Using the same log file prefix in the import/export manifest files to create a versioned log file in Amazon S3 for all imports.

Q15)

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.

Q16) TPT Limited is experiencing increase in customer contacts from a specific region. The social media manager at TPT Limited, requests additional staff on the weekends to address the increased customer contacts. Which is the correct manner of data representation using QuickSight to visualize the trends on weekends over the past 6 months?

- A map of regions with a heatmap overlay to show the volume of customer contacts
- A bar graph plotting region vs. volume of social media contacts
- A pie chart per region plotting customer contacts per day of week
- None of these
- ✅ A line graph plotting customer contacts vs. time, with a line for each region

---

Q17)

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.

Q18)

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.

Q19) Robert, head of digital marketing wants to analyze clickstream data from multiple applications on the website of TPT Limited, for pattern of pages a consumer clicks on and in what order by using data real time with minimum infrastructure. Which of the following options should Robert use to meet the requirements?

- Send click events directly to Amazon Redshift and then analyze them with SQL.
- Use Elastic MapReduce to ingest the data and analyze it.
- ✅ Use Amazon Kinesis with a worker to process the data received from the Kinesis stream.
- None of these
- Publish web clicks by session to an Amazon SQS.

Q20)

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.

Q21)

A retailer exports data daily from its transactional databases into an S3 bucket in the Sydney region.

The retailer's Data Warehousing team wants to import this data into an existing Amazon Redshift cluster in their VPC at Sydney.

Corporate security policy mandates that data can only be transported within a VPC.

What combination of the following steps will satisfy the security policy? Choose 2 answers

- Create a NAT gateway in a public subnet to allow the Amazon Redshift cluster to access Amazon S3.
- Create a Cluster Security Group to allow the Amazon Redshift cluster to access Amazon S3.
- ✅ Enable Amazon Redshift Enhanced VPC Routing.
- ✅ Create and configure an Amazon S3 VPC endpoint.

Q22)

Your company needs to design a data warehouse for a client in the retail industry. The data warehouse will store historic purchases in Amazon Redshift.

To comply with PCI:DSS requirements and meet data protection standards, the data must be encrypted at rest and have keys

managed by a corporate on-premises HSM.

How can you meet these requirements in a cost-effective manner?

- 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 VPN connection between a VPC you create in AWS and an on-premises network. Then launch the Redshift cluster in the VPC, and configure it to use your corporate HSM.
- Use AWS Import/Export to import a company HSM device into AWS alongside the Amazon Redshift cluster, and configure Redshift to use the imported HSM.

Q23)

Your client needs to load a 600 GB file into a Redshift cluster from S3, using the Redshift COPY command.

The file has several known (and potentially some unknown) issues that will probably cause the load process to fail.

How should the client most efficiently detect load errors without needing to perform cleanup if the load process fails?

- Write a script to delete the data from the tables in case of errors.
- Compress the input file before running COPY.
- Split the 600 GB file into smaller 25 GB chunks and load each separately.
- ✅ Use the COPY command with the NOLOAD parameter.

Q24)

A company has lot of web applications, databases and data warehouse built on Teradata, NoSQL databases, and other types of data stores.

They have lot of data assets in terms of logs, documents; excel files, CSV files, PDF documents and others. Web Application has different user workloads at different parts of the day. They are running one of their web application Node.js supported by MongoDB Database.

The schema designed is document based. The team wants to migrate the platform on to AWS.

Which NoSQL Managed service provides the document management capability?

- Amazon Neptune Database, being a graph database support document models and NoSQL requirements
- ✅ Amazon DynamoDB Database, being a document database support document models and NoSQL requirements
- Amazon RDS Database, being a multi-modal database support document models and NoSQL requirements
- Amazon Aurora Database, being a multi-modal database support document models and NoSQL requirements

---

Q25)

A company launched EMR cluster to support their big data analytics requirements. They have multiple data sources built out of S3, SQL databases,MongoDB, Redis, RDS, other file systems.

They are looking for distributed processing framework and programming model that helps you do machine learning, stream processing, or graph analytics using Amazon EMR clusters.

Which EMR Hadoop ecosystem fulfils the requirements?

- ✅ Apache Spark
- Apache HCatalog
- Apache HBase
- Apache Hive

Q26)

A media advertising company handles a large number of real-time messages sourced from over 200 websites.

The company's data engineer needs to collect and process records in real time for analysis using Spark Streaming on Amazon Elastic MapReduce (EMR).

The data engineer needs to fulfill a corporate mandate to keep ALL raw messages as they are received as a top priority.

Which Amazon Kinesis configuration meets these requirements?

- Publish messages to Amazon Kinesis Streams, pull messages off with Spark Streaming, and write raw data to Amazon Simple Storage Service (S3) before and after processing.
- Publish messages to Amazon Kinesis Firehose backed by Amazon Simple Storage Service (S3). Use AWS Lambda to pull messages from Firehose to Streams for processing with Spark Streaming.
- ✅ Publish messages to Amazon Kinesis Streams. Pull messages off Streams with Spark Streaming in parallel to AWS Lambda pushing messages from Streams to Firehose backed by Amazon Simple Storage Service (S3).
- Publish messages to Amazon Kinesis Firehose backed by Amazon Simple Storage Service (S3). Pull messages off Firehose with Spark Streaming in parallel to persistence to Amazon S3.

Q27)

A customer has an Amazon S3 bucket. Objects are uploaded simultaneously by a cluster of servers from multiple streams of data.

The customer maintains a catalog of objects uploaded in Amazon S3 using an Amazon DynamoDB table. This catalog has the following fields: StreamName, TimeStamp, and ServerName, from which ObjectName can be obtained.

The customer needs to define the catalog to support querying for a given stream or server within a defined time range.

Which DynamoDB table scheme is most efficient to support these queries?

- Define a Primary Key with ServerName as Partition Key. Define a Local Secondary Index with TimeStamp as Partition Key. Define a Global Secondary Index with StreamName as Partition Key and TimeStamp as Sort Key.
- Define a Primary Key with ServerName as Partition Key. Define a Local Secondary Index with StreamName as Partition Key. Define a Global Secondary Index with TimeStamp as Partition Key.
- Define a Primary Key with ServerName as Partition Key and TimeStamp as Sort Key. Do NOT define a Local Secondary Index or Global Secondary Index.
- ✅ Define a Primary Key with StreamName as Partition Key and TimeStamp followed by ServerName as Sort Key. Define a Global Secondary Index with ServerName as partition key and TimeStamp followed by StreamName.

Q28)

An Amazon EMR cluster using EMRFS has access to petabytes of data on Amazon S3, originating from multiple unique data sources.

The customer needs to query common fields across some of the data sets to be able to perform interactive joins and then display results quickly.

Which technology is most appropriate to enable this capability?

- Pig
- MicroStrategy
- ✅ Presto
- R Studio

Q29)

A data engineer is running a DWH on a 25-node Redshift cluster of a SaaS service. The data engineer needs to build a dashboard that will be used by customers.

Five big customers represent 80% of usage, and there is a long tail of dozens of smaller customers. The data engineer has selected the dashboarding tool.

How should the data engineer make sure that the larger customer workloads do NOT interfere with the smaller customer workloads?

- Route the largest customers to a dedicated Redshift cluster. Raise the concurrency of the multi-tenant Redshift cluster to accommodate the remaining customers.
- Push aggregations into an RDS for Aurora instance. Connect the dashboard application to Aurora rather than Redshift for faster queries.
- ✅ Place the largest customers into a single user group with a dedicated query queue and place the rest of the customers into a different query queue.
- Apply query filters based on customer-id that can NOT be changed by the user and apply distribution keys on customer-id.

Q30)

A customer needs to determine the optimal distribution strategy for the ORDERS fact table in its Redshift schema.

The ORDERS table has foreign key relationships with multiple dimension tables in this schema.

How should the company determine the most appropriate distribution key for the ORDERS table?

- ✅ Identify the largest and the most frequently joined dimension table and designate the key of this dimension table as the distribution key of the ORDERS table.
- Identify the smallest dimension table and designate the key of this dimension table as the distribution key of the ORDERS table.
- Identify the largest dimension table and designate the key of this dimension table as the distribution key of the ORDERS table.
- Identify the largest and most frequently joined dimension table and ensure that it and the ORDERS table both have EVEN distribution.

Q31)

Your customer is willing to consolidate their log streams (access logs, application logs, security logs etc.) in one single system.

Once consolidated, the customer wants to analyze these logs in real time based on heuristics.

From time to time, the customer needs to validate heuristics, which requires going back to data samples extracted from the last 12 hours.

What is the best approach to meet your customer's requirements?

- Setup an Auto Scaling group of EC2 syslogd servers, store the logs on S3 use EMR to apply heuristics on the logs
- Configure Amazon CloudTrail to receive custom logs, use EMR to apply heuristics the logs
- ✅ Send all the log events to Amazon Kinesis develop a client process to apply heuristics on the logs
- Send all the log events to Amazon SQS. Setup an Auto Scaling group of EC2 servers to consume the logs and apply the heuristics.

Q32)

Your team is building up a smart home iOS APP. The end users will use your company's camera-equipped home devices such as baby monitors, webcams, and home surveillance systems. Then the videos would be uploaded to AWS.

The users can then play the on-demand or live videos using the format of HTTP Live Streaming (HLS) through the Mobile application.

Which combinations of steps should you use to design the solution? (Select TWO)

- [x] In the mobile application, use HLS to display the video stream by using the HLS streaming session URL.
- [x] Create a Kinesis video stream to capture, store, and index the videos from the camera-equipped home devices.
- [ ] Transform the stream data to HLS compatible data by using Kinesis Data Analytics or customer code in EC2/Lambda. Then in the mobile application, use HLS protocol to display the video stream by using the converted HLS streaming data.
- [ ] Create a Kinesis Data Firehose to ingest, durably store and encrypt the live videos from the users' home devices.

Q33)

You require the ability to analyze a customer?s clickstream data on a website so they can do behavioral analysis.

Your customer needs to know what sequence of pages and ads their customer clicked on. This data will be used in real time to modify the page layouts as customers click through the site to increase stickiness and advertising click-through.

Which option meets the requirements for capturing and analyzing this data?

- [ ] Publish web clicks by session to an Amazon SQS queue and periodically drain these events to Amazon RDS and analyze with SQL
- [ ] Write click events directly to Amazon Redshift and then analyze with SQL
- [x] Push web clicks by session to Amazon Kinesis and analyze behavior using Kinesis workers
- [ ] Log clicks in weblogs by URL store to Amazon S3, and then analyze with Elastic MapReduce

---

Q34)

A company is performing a full migration of its systems from an on-premises data center to AWS. The company needs to move all the data stored on-premises to Amazon S3 within the next 4 weeks.

Currently, the on-premises storage holds 900 TB of data and is connected to the Internet over a 100 Mbps link. Up to 20% of the link's throughput is regularly used in real time by existing systems.

What is the MOST cost-effective way to perform the data migration in the given time frame?

- [ ] Set up an AWS Direct Connect link to upload the data.
- [ ] Use a multipart upload to transfer the data over the existing link.
- [x] Order multiple AWS Snowball devices to ship the data.
- [ ] Configure a VPN tunnel for the AWS environment to upload the data.

Q35)

You have been asked to cost optimize a business critical and long-running EMR cluster.

The EMR cluster is currently on-demand for the master nodes, core nodes and task nodes.

The costs for running the cluster have been steadily increasing as nodes have been added and resized.

What would you suggest the business does to reduce the costs without requiring any long-term commitment?

- [x] Leave the master and core nodes as on-demand and use spot instances for the task nodes
- [ ] Leave all nodes running on-demand instances, the cluster is already cost optimized.
- [ ] Leave the master node to use on-demand and change the core and task nodes to spot
- [ ] Recreate the cluster using spot instances for the master, core and task nodes.

Q36)

You have just joined a new company and have been put in charge of EC2 instances and any other services that use EC2 instances.

You notice that the company has been slow to take advantage of AWS per-second Billing, specifically in the area of EMR and Spot Instances.

What immediate steps can you take on EMR with spot instances to improve cost saving and performance?

- [x] Run more instances for a shorter amount of time.
- [ ] Run fewer instances for a longer amount of time.
- [ ] Run fewer instances for a shorter amount of time.
- [ ] Use on-demand instances instead.

Q37)

You are designing a service that aggregates clickstream data in batch and delivers reports to subscribers via email only once per week.

Data is extremely spikey, geographically distributed, high-scale, and unpredictable.

How should you design this system?

- [ ] Use AWS Elasticsearch service and EC2 Auto Scaling groups. The Autoscaling groups scale based on click throughput and stream into the Elasticsearch domain, which is also scalable. Use Kibana to generate reports periodically.
- [ ] Use API Gateway invoking Lambdas which PutRecords into Kinesis, and EMR running Spark performing GetRecords on Kinesis to scale with spikes. Spark on EMR outputs the analysis to S3, which are sent out via email.

✅ Use a CloudFront distribution with access log delivery to S3. Clicks should be recorded as query string GETs to the distribution. Reports are built and sent by periodically running EMR jobs over the access logs in S3.
⚪ Use a large Redshift cluster to perform the analysis, and a fleet of Lambdas to perform record inserts into the Redshift tables. Lambda will scale rapidly enough for the traffic spikes.

Q38)

A manufacturing company stores all telemetry data inside an Amazon DynamoDB table.

The company is satisfied with the performance but concerned that the storage cost might increase over time.

Which combination of steps should be taken to move old data into archival storage? (Select THREE.)

✅ Create an Amazon Kinesis Data Firehose delivery stream to load the data into Amazon S3 and set lifecycle policies to archive it to Amazon Glacier.
⚪ Create Amazon CloudWatch Events when a new item is added to the DynamoDB tablInvoke an AWS Lambda function to capture the changes and write to Amazon Kinesis Data Streams.
✅ Create a custom AWS Lambda function to regularly poll the DynamoDB Stream and deliver the batch records to an Amazon Kinesis Data Firehose.
⚪ Create rolling tables on DynamoDB to store data in a particular order and create custom application logic to handle the creation and deletion of tables.
✅ Enable DynamoDB Streams on the table and TTL.
⚪ Enable DynamoDB Streams and use the KCL with the DynamoDB Streams Kinesis Adapter to capture changes on DynamoDB tables.

Q39) TPT Limited has well-structured data coming with high velocity data, to be used for ad-hoc business analytics queries by the BI (business intelligence) team having good knowledge of SQL. Which of the following AWS service should be used in this case?

⚪ EMR running Apache Spark
⚪ EMR using Hive
⚪ None of these
✅ Kinesis Firehose + Redshift
⚪ Kinesis Firehose + RDS

Q40) Robert is an AWS administrator at TPT Limited. He has been asked to run a nightly COPY command on a 500-GB file in Amazon S3, into a 10-node Amazon Redshift cluster. How should Robert prepare the data, for optimizing the performance of the COPY command?

⚪ Split the file into 10 files of equal size.
⚪ Convert the file format to AVRO.
⚪ None of these
✅ Split the file into 500 smaller files.
⚪ Compress the file using gz compression.

Q41)

A customer needs to load a 550-GB data file into an Amazon Redshift cluster from Amazon S3, using the COPY command.

The input file has both known and unknown issues that will probably cause the load process to fail.

The customer needs the most efficient way to detect load errors without performing any cleanup if the load process fails.

Which technique should the customer use?

⚪ Write a script to delete the data from the tables in case of errors.
✅ Use COPY with NOLOAD parameter.
⚪ Split the input file into 50-GB blocks and load them separately.
⚪ Compress the input file before running COPY.

Q42)

An administrator decides to use the Amazon Machine Learning service to classify social media posts that mention your company into two categories: posts that require a response and posts that do not.

The training dataset of 10,000 posts contains the details of each post, including the timestamp, author, and full text of the post. You are missing the target labels that are required for training.

Which two options will create valid target label data?

⚪ Using the a priori probability distribution of the two classes, use Monte-Carlo simulation to generate the labels.
✅ Use the Amazon Mechanical Turk web service to publish Human Intelligence Tasks that ask Turk workers to label the posts.
⚪ Use the sentiment analysis NLP library to determine whether a post requires a response.
✅ Ask the social media handling team to review each post and provide the label.

Q43)

A company logs data from its application in large files and runs regular analytics of these logs to support internal reporting for three months after the logs are generated.

After three months, the logs are infrequently accessed for up to a year. The company also has a regulatory control requirement to store application logs for seven years.

Which course of action should the company take to achieve these requirements in the most cost-efficient way?

○ Store the files in S3 Standard with a lifecycle policy to remove them after a year. Simultaneously store the files in Amazon S3 Glacier with a Deny Delete vault lock policy for archives less than seven years old.
✓ Store the files in S3 Standard with lifecycle policies to transition the storage class to Standard ? IA after three months and delete them after a year. Simultaneously store the files in Amazon Glacier with a Deny Delete vault lock policy for archives less than seven years old.
○ Store the files in S3 Standard with a lifecycle policy to transition the storage class to Standard - IA after three months. After a year, transition the files to Glacier and add a Deny Delete vault lock policy for archives less than seven years old.
○ Store the files in S3 Glacier with a Deny Delete vault lock policy for archives less than seven years old and a vault access policy that restricts read access to the analytics IAM group and write access to the log writer service role.

Q44)

A new algorithm has been written in Python to identify SPAM e-mails.

The algorithm analyzes the free text contained within a sample set of 1 million e-mails stored on Amazon S3.

The algorithm must be scaled across a production dataset of 5 PB, which also resides in Amazon S3 storage.

Which AWS service strategy is best for this use case?

○ Use Amazon Elasticsearch Service to store the text and then use the Python Elasticsearch Client to run analysis against the text index.
✓ Use Amazon EMR to parallelize the text analysis tasks across the cluster using a streaming program step.
○ Copy the data into Amazon ElastiCache to perform text analysis on the in-memory data and export the results of the model into Amazon Machine Learning.
○ Initiate a Python job from AWS Data Pipeline to run directly against the Amazon S3 text files.

Q45)

An administrator needs to manage a large catalog of items from various external sellers.

The administrator needs to determine if the items should be identified as minimally dangerous, dangerous, or highly dangerous based on their textual descriptions. The administrator already has some items with the danger attribute, but receives hundreds of new item descriptions every day without such classification. The administrator has a system that captures dangerous goods reports from customer support team or from user feedback.

What is a cost-effective architecture to solve this issue?

○ Build a machine learning model with binary classification for dangerous goods and run it on the DynamoDB Streams as every new item description is added to the system.
✓ Build a machine learning model to properly classify dangerous goods and run it on the DynamoDB Streams as every new item description is added to the system.
○ Build a Kinesis Streams process that captures and marks the relevant items in the dangerous goods reports using a Lambda function, once more than two reports have been filed.
○ Build a set of regular expression rules that are based on the existing examples, and run them on the DynamoDB Streams as every new item description is added to the system.

Q46)

A customer is collecting clickstream data using Amazon Kinesis and is grouping the events by IP address into 5-minute chunks stored in Amazon S3.

Many analysts in the company use Hive on Amazon EMR to analyze this data. Their queries always reference a single IP address.

Data must be optimized for querying based on IP address using Hive running on Amazon EMR.

What is the most efficient method to query the data with Hive?

○ Store the events for an IP address as a single file in Amazon S3 and add metadata with keys:Hive_Partitioned_IPAddress.
○ Store the data in an HBase table with the IP address as the row key.
✓ Store the Amazon S3 objects with the following naming scheme bucket_name/source=ip_address/year=yy/month=mm/day=dd/hour=hh/filename.
○ Store an index of the files by IP address in the Amazon DynamoDB metadata store for EMRFS.

Q47)

An organization uses a custom map reduce application to build monthly reports based on many small data files in an Amazon S3 bucket.

The data is submitted from various business units on a frequent but unpredictable schedule. As the dataset continues to grow, it becomes increasingly difficult to process all of the data in one day. The organization has scaled up its Amazon EMR cluster, but other optimizations could improve performance. The organization needs to improve performance with minimal changes to existing processes and applications.

What action should the organization take?

○ Have business units submit data via Amazon Kinesis Firehose to aggregate data hourly into Amazon S3.
✓ Schedule a daily AWS Data Pipeline process that aggregates content into larger files using S3DistCp.
○ Use Amazon S3 Event Notifications and AWS Lambda to index each file into an Amazon Elasticsearch Service cluster.
○ Add Spark to the Amazon EMR cluster and utilize Resilient Distributed Datasets in-memory.
○ Use Amazon S3 Event Notifications and AWS Lambda to create a quick search file index in DynamoDB.

Q48)

An administrator tries to use the Amazon Machine Learning service to classify social media posts that mention the administrator's company into posts that require a response and posts that do not.

The training dataset of 10,000 posts contains the details of each post including the timestamp, author, and full text of the post.

The administrator is missing the target labels that are required for training.

Which Amazon Machine Learning model is the most appropriate for the task?

- Regression model where the predicted value is the probability that the post requires a response
- Multi-class prediction model, with two classes: require-response post and does-not-require-response
- ✅ Binary classification model, where the two classes are the require-response post and does-not-require-response
- Unary classification model, where the target class is the require-response post

---

Q49)

A company develops a tool whose coverage includes blogs, news sites, forums, videos, reviews, images and social networks such as Twitter and Facebook.

Users can search data by using Text and Image Search, and use charting, categorization, sentiment analysis and other features to provide further information and analysis. They have access to over 80 million sources. They want to provide Image and text analysis capabilities to the applications which includes identify objects, people, text, scenes, and activities and also provides highly accurate facial analysis and facial recognition.

What service can provide this capability?

- Amazon Polly
- ✅ Amazon Rekognition
- Amazon Comprehend
- Amazon SageMaker

Q50)

You manage a web advertising platform on a single AWS account. This platform produces real-time ad-click data that you store as objects in an Amazon S3 bucket called "click-data".

Your advertising partners want to use Amazon Elastic MapReduce in their own AWS accounts to do analytics on the ad-click data. They've asked for immediate access to the ad-dick data so that they can run analytics.

Which two choices are required to facilitate secure access to this data? Choose 2 answers.

- Configure AWS Data Pipeline in the partner AWS accounts to use the web Identity Federation API to access data in the "click-data" bucket.
- Configure AWS Data Pipeline to transfer the data from the "click-data" bucket to the partner's Amazon Elastic MapReduce cluster.
- Configure the Amazon S3 bucket access control list to allow access to the partners Amazon Elastic MapReduce cluster.
- ✅ Configure an Amazon S3 bucket policy for the "click-data" bucket that allows Read-Only access to the objects and associate this policy with an IAM role.
- Create a new IAM group for AWS Data Pipeline users with a trust policy that contains partner AWS account IDs.
- ✅ Create a cross-account IAM role with a trust policy that contains partner AWS account IDs and a unique external ID

Q51)

A Solutions Architect is designing a weather forecast application. Every hour, the application will receive a new set of raw data from weather stations. The application will analyze this data and produce a set of local weather forecasts available for users to download. The analysis takes 50 minutes to run on 2,000 vCPUs. The analysis must complete before the next set of data is available. Each local weather forecast is typically 10 GB in size. The forecasts are accessed heavily during the first hour they are available, with usage dropping rapidly as newer forecasts become available.

Which combination of steps is the MOST cost-effective architecture? (Select TWO.)

- Store weather forecast data in Amazon S3 Standard-Infrequent Access (S3 Standard-IA). Configure a lifecycle policy to transition the data to Amazon Glacier after 90 days.
- ✅ Store weather forecast data in Amazon S3 One Zone-Infrequent Access (S3 One Zone-IA). Configure a lifecycle policy to transition the data to Amazon Glacier after 90 days.
- Store weather forecast data in Amazon S3 Standard. Configure a lifecycle policy to transition the data to Amazon S3 Standard-Infrequent Access (S3 Standard-IA) after 30 days.
- Conduct the analysis on an Amazon EC2-based cluster using 1-hour Spot blocks in multiple AWS Regions.
- ✅ Conduct the analysis on a cluster of Amazon EC2 instances using Reserved Instances in a single AWS Region.

Q52)

A company has two different types of reporting needs on their 200-GB data warehouse;

Data scientists run a small number of concurrent adhoc SQL queries that can take several minutes each to run.

Display screens throughout the company run many fast SQL queries to populate dashboards,

Which design would meet these requirements with the LEAST cost?

- Use Amazon Redshift for Data Scientists; Run automated dashboard queries against Redshift and store the results in Amazon ElastiCache, Dashboards query ElastiCache.
- ✅ Use Amazon Redshift for both requirements, with separate query queues configured in workload management.
- Configure auto-replication between Amazon Redshift and Amazon RDS. Data scientists use Redshift and Dashboards use RDS
- Replicate relevant data between Amazon Redshift and Amazon DynamoDB. Data scientists use Redshift. Dashboards use DynamoDB

Q53)

Your company sells consumer devices and needs to record the first activation of all sold devices. Devices are not activated until the information is written on a persistent database.

Activation data is very important for your company and must be analyzed daily with a MapReduce job. The execution time of the data analysis process must be less than three hours per day. Devices are usually sold evenly during the year, but when a new device model is out, there is a predictable peak in activation's, that is, for a few days there are 10 times or even 100 times more activation's than in average day.

Which of the following databases and analysis framework would you implement to better optimize costs and performance for this workload?

- Amazon DynamoDB and Amazon Elastic MapReduce with Reserved instances
- Amazon RDS and Amazon Elastic MapReduce with Reserved instances.
- ✅ Amazon DynamoDB and Amazon Elastic MapReduce with Spot instances.
- Amazon RDS and Amazon Elastic MapReduce with Spot instances.

Q54)

A medical record filing system for a government medical fund is using an Amazon S3 bucket to archive documents related to patients.

Every patient visit to a physician creates a new file, which can add up millions of files each month. Collection of these files from each physician is handled via a batch process that runs ever? night using AWS Data Pipeline. This is sensitive data, so the data and any associated metadata must be encrypted at rest. Auditors review some files on a quarterly basis to see whether the records are maintained according to regulations.

Auditors must be able to locate any physical file in the S3 bucket for a given date, patient, or physician. Auditors spend a significant amount of time location such files.

What is the most cost and time efficient collection methodology in this situation?

- Use Amazon S3 event notification to populate an Amazon Redshift table with metadata about every file loaded to Amazon S3, and partition them based on the month and year of the file.
- ✅ Use Amazon S3 event notification to populate an Amazon DynamoDB table with metadata about every file loaded to Amazon S3, and partition them based on the month and year of the file.
- Use Amazon API Gateway to get the data feeds directly from physicians, batch them using a Spark application on Amazon Elastic MapReduce (EMR), and then store them in Amazon S3 with folders separated per physician.
- Use Amazon Kinesis to get the data feeds directly from physicians, batch them using a Spark application on Amazon Elastic MapReduce (EMR), and then store them in Amazon S3 with folders separated per physician.

Q55)

A company uses Amazon Redshift for its enterprise data warehouse. A new on-premises PostgreSQL OLTP DB must be integrated into the data warehouse.

Each table in the PostgreSQL DB has an indexed last_modified timestamp column. The data warehouse has a staging layer to load source data into the data warehouse environment for further processing. The data lag between the source PostgreSQL DB and the Amazon Redshift staging layer should NOT exceed four hours.

What is the most efficient technique to meet these requirements?

- Extract the incremental changes periodically using a SQL query. Upload the changes to a single Amazon Simple Storage Service (S3) object, and run the COPY command to load to the Amazon Redshift staging layer.
- ✅ Extract the incremental changes periodically using a SQL query. Upload the changes to multiple Amazon Simple Storage Service (S3) objects, and run the COPY command to load to the Amazon Redshift staging layer.
- Use a PostgreSQL trigger on the source table to capture the new insert/update/delete event and write it to Amazon Kinesis Streams. Use a KCL application to execute the event on the Amazon Redshift staging table.
- Create a DBLINK on the source DB to connect to Amazon Redshift. Use a PostgreSQL trigger on the source table to capture the new insert/update/delete event and execute the event on the Amazon Redshift staging table.

Q56)

An organization needs a data store to handle the following data types and access patterns:

- Faceting

- Search

- Flexible schema (JSON) and fixed schema

- Noise word elimination

Which data store should the organization choose?

- ✅ Amazon Elasticsearch Service
- Amazon Relational Database Service (RDS)
- Amazon DynamoDB
- Amazon Redshift

Q57)

You are provisioning an application using EMR. You have requested 100 instances. You are charged $0.015 per hour, per instance.

In the first 10 minutes after your launch request, Amazon EMR starts your cluster. 90 of your instances are available. **It takes your**

cluster one hour to complete.

How much will you be charged for this EMR usage for the first hour?

- TRUE
- ✓ $1.35 per hour
- $1.50 per hour
- 0.015

Q58)

You have been asked to ensure that all AWS API calls are collected across your company's AWS account and that they are kept around for 90 days for analysis. After that, they must be able to be restored for 3 years.

How can you meet these needs in a scalable, cost-effective way?

- Enable CloudTrail logging to Glacier, and set a lifecycle policy to expire the data after 3 years.
- Enable CloudTrail logging in all accounts into S3 buckets, and set a lifecycle policy to expire the data in each bucket after 3 years.
- ✓ Enable CloudTrail logging to a centralized S3 bucket, set a lifecycle policy to move the data to Glacier after 90 days, and expire the data after 3 years.
- Enable AWS CloudTrail logging across all accounts to a centralized Amazon S3 bucket with versioning enabled. Set a lifecycle policy to move the data to Amazon Glacier daily, and expire the data after 90 days.

Q59)

You company has launched an EMR cluster to support their big data analytics requirements. They are planning to build an application running on EMR which supports both OLTP and operational analytics allowing you to use standard SQL queries and JDBC APIs to work with an Apache HBase backing store. Also data transfer tool between Amazon S3, Hadoop, HDFS, and RDBMS databases.

Which EMR Hadoop ecosystem fulfils the requirements? (Select TWO)

- Apache Ganglia
- ✓ Apache Sqoop
- ✓ Apache Phoenix
- Apache Flink
- Apache Hue

Q60)

A company is looking to collect and process the log files in near real time that are generated from thousands of applications in their AWS cloud.

They are also collecting stock pricing information from stock price publishing data providers and using the information to recommend stocks to customers.

They are looking at querying streams and using Kinesis Analytics application to process all the stocks for recommendation if price changes greater than 10 percent.

What kind of Queries will help fulfill the requirement?

- Sliding windows queries
- Tumbling Windows queries
- Stagger Windows queries
- ✓ Continuous queries
