Answer Sheet

Q1)

You operate a 32TB Redshift cluster made up of 16 ds2.xlarge nodes (with a 2TB HDD per node) in the Asia Pacific (Singapore) region that tracks time series data. A new business need requires you to triple the size of this cluster to 96TB by resizing your Redshift cluster.

What is the correct cluster resizing method?

✅ CLASSIC resize to 6 ds2.8xlarge nodes (with a 16TB HDD per node)
Explanation:-This change requires a CLASSIC resize operation because the current node type (ds2.xlarge) has a maximum storage per node of 2TB and a maximum node range of 1-32 (see https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html), which would only deliver 64TB of storage, not the 96TB of storage required. Option A (CLASSIC resize to 48 ds2.xlarge nodes [with a 2TB HDD per node]) is incorrect because the 48 nodes stated in the answer is greater than the 32 maximum nodes possible for that node type. Option B (ELASTIC resize to 48 ds2.xlarge nodes [with a 2TB HDD per node]) is also incorrect because it requires a number of nodes greater than the maximum number allowed for the stated node type. Option D (ELASTIC resize to 6 ds2.8xlarge nodes [with a 16TB HDD per node]) is incorrect because it applies ELASTIC resizing to a different node type, which is not possible.
- ELASTIC resize to 48 ds2.xlarge nodes (with a 2TB HDD per node)
- CLASSIC resize to 48 ds2.xlarge nodes (with a 2TB HDD per node)
- ELASTIC resize to 6 ds2.8xlarge nodes (with a 16TB HDD per node)

Q2)

You receive an emergency call because a critical query in your Redshift cluster is not finished and has been running for three times the typical duration.

What groups of tables will you investigate first to understand what is currently running in the cluster?

✅ STV
Explanation:-STV (System Table - Virtual) snapshots current system data, including currently running queries, current active users, current load states, and so on.
- System catalog tables
- STL
- PG metadata tables

Q3) What are three qualities of a Machine Learning program? (Select three.)

✅ Machine Learning programs improve accuracy with more data.
Explanation:-(Machine Learning programs improve accuracy with unsupervised learning) can be eliminated because unsupervised learning is not a type of Machine Learning.
✅ Machine Learning programs are able to alter themselves.
Explanation:-(Machine Learning programs improve accuracy with unsupervised learning) can be eliminated because unsupervised learning is not a type of Machine Learning.
✅ Machine Learning programs are created with data rather than rules.
Explanation:-(Machine Learning programs improve accuracy with unsupervised learning) can be eliminated because unsupervised learning is not a type of Machine Learning.
- Machine Learning programs improve accuracy with unsupervised learning.

Q4)

Your company is about to make a decision about a Big Data system that must include both interactive queries and visualizations, plus a Machine Learning pipeline powered by AWS Comprehend and AWS Lex. Your leadership is not sure about the use of AWS Athena and how it may fit into the overall EMR pipeline.

What two statements correctly describe AWS Athena and its technical architecture? (Select two.)

- Athena is powered by HIVE and SPARK and uses tables to contain metadata about underlying source data.
✅ Athena is powered by HIVE and PRESTO and uses tables to contain metadata about underlying source data.
Explanation:-Athena is a query service, while QuickSight is a visualization service; Athena is powered by HIVE and PRESTO.
- Athena is an interactive visualization service that uses a schema-on-read and does not require ETL in advance.
✅ Athena is an interactive query service that uses a schema-on-read and does not require ETL in advance.
Explanation:-Athena is a query service, while QuickSight is a visualization service; Athena is powered by HIVE and PRESTO.

Q5)

Your team operates a high-profile website that enables users to upload a photo and determine whether the photo is either a hotdog, not a hotdog, or a T-rex.

What kind of Machine Learning algorithm has your team implemented to produce this result on the website?

- Clustering
- Association
✅ Classification
Explanation:-Classification algorithms output discrete class categories.
- Regression

Q6)

You have been asked to use your department's existing continuous integration (CI) tool to test a three- tier web architecture defined in an AWS CloudFormation template. The tool already supports AWS APIs and can launch new AWS CloudFormation stacks after polling version control. The CI tool reports on the success of the AWS CloudFormation stack creation by using the DescribeStacks API to look for the CREATE_COMPLETE status.

The architecture tiers defined in the template consist of -

*** One load balancer**

*** Five Amazon EC2 instances running the web application**

*** One multi-AZ Amazon RDS instance How would you implement this?**

Select the most appropriate option for the following.

1. Define a WaitCondition and a WaitConditionhandle for the output of a output of a UserData command that does sanity checking of the application's post-install state.

2. Define a UserDataHandle for the output of a CustomResource that does sanity checking of the application's post-install state

3. Define a UserDataHandle for the output of a UserData command that does sanity checking of the application's post-install state and runs integration tests on the state of multiple tiers through load balancer to the application

4. Define a WaitCondition and use a WaitConditionHandle that leverages the AWS SDK to run the Describe Stacks API call until the CREATE_COMPLETE status is returned

- Only 1 and 4
- Option 2 and 4
- Option 2 and 3
- ✓ Option 1 and 2

Q7)

A customer wants to track access to their Amazon Simple Storage Service (S3) buckets and also use this information for their internal security and access audits.

Which of the following will meet the Customer requirement?

- Enable AWS CloudTrail to audit all Amazon S3 bucket access.
- ✓ Enable server access logging for all required Amazon S3 buckets.
- Enable the Requester Pays option to track access via AWS Billing
- Enable Amazon S3 event notifications for Put and Post.

Q8)

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 minimal changes to existing processes and applications.

What action should the organization take?

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

Q9)

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

Q10)

A user has setup an RDS DB with Oracle. The user wants to get notifications when someone modifies the security group of that DB.

How can the user configure that?

- Configure the CloudWatch alarm on the DB for a change in the security group
- It is not possible to get the notifications on a change in the security group
- ✓ Configure event notification on the DB security group
- Configure SNS to monitor security group changes

Q11)

A data engineer in a manufacturing company is designing a data processing platform that receives a large volume of unstructured data.

The data engineer must populate a well-structured star schema in Amazon Redshift.

What is the most efficient architecture strategy for this purpose?

- ✓ Transform the unstructured data using Amazon EMR and generate CSV data. COPY the CSV data into the analysis schema within Redshift.
- Load the unstructured data into Redshift, and use string parsing functions to extract structured data for inserting into the analysis schema
- 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.
- None of these

Q12)

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 1millione-mails stored on Amazon S3. The algorithm must be scaled across a production dataset of 5PB, which also resides in Amazon S3 storage.

Which AWS service strategy is best for this use case?

- Copy the data inro Amazon ElasriCache to perform text analysis on the in-memory data and export the results of the model inro amazon Machine Learning
- Use Amazon EMR to parallelize the text analysis task across the cluster using a streaming program step.
- 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 ot run analysis against the text index.

Q13)

A data engineer chooses Amazon DynamoDB as a data store for a regulated application. This application must be submitted to regulators for review.

The data engineer needs to provide a control framework that lists the security controls from the process to follow to add new users down to the physical controls of the data center, including items like security guards and cameras.

How should this control mapping be achieved using AWS?

- ✓ Request AWS third party audit reports and/or the AWS quality addendum and map the AWS responsibilities to the controls that must be provided.
- Request data center Temporary Auditor access to an AWS data center to verify the control mapping.
- Request relevant SLAs and security guidelines for Amazon DynamoDB and define these guidelines within the application's arcitecture to map to the control framework.
- Request Amazon DynamoDB system architecture designs to determine how to map the AWS responsibilities to the control that must be provided.

Q14)

An administrator needs to design a distribution strategy for a star schema in a Redshift cluster. The administrator needs to determine the optimal distribution styles for the tables in the Redshift schema.

In which three circumstances would choosing key-based distribution be most appropriate.

1. When the administrator needs to reduce cross-note traffic

2. When the administrator needs to optimize a large, slowly changing dimension table.

3. When the administrator needs to optimize the fact table for parity with the number of slices.

4. When the administrator needs to take advantage of data locality on a local node for joins and aggregates.

- Only 1, 2 and 4
- ✓ Only 2, 3 and 4
- Only 1, 2, and 3
- Only 1, 3 and 4

Q15) 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 200ms.

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

- Anontmize 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.
- ✓ 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.

Encrypt the data files according to encryption standards of country X and srore them on AWS region in Amazon S3. Hve the EMR cluster read the dataset using EMRFS.

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.

Q16)

An administrator needs to design a strategy for the schema in a Redshift cluster. The administrator needs to determine the optimal distribution style for the tables in the Redshift schema.

In which two circumstances would choosing EVEN distribution be most appropriate?

1. When the table are highly denormalized and do NOT participate in frequent joins.

2. When data must be grouped based on a specific key on a defined slice.

3. When a new table has been loaded and it is unclear how it will be joined to dimension.

4. When data transfer between nodes must be eliminated.

Only 1 and 4

Only 1 and 3

✅ Only 2 and 3

Only 1 and 2

Q17)

Which statements are true of sequence numbers in Amazon Kinesis?

1. Sequence numbers are assigned by Amazon Kinesis when a data producer calls PutRecords operation to add data to an Amazon Kinesis stream

2. A data pipeline is a group of data records in a stream.

3. The longer the time period between PutRecord or PutRecords requests, the larger the sequence number becomes.

4. Sequence numbers are assigned by Amazon Kinesis when a data producer calls PutRecord operation to add data to an Amazon Kinesis stream

Only 1, 3 and 4

✅ Only 1, 2, and 3

Explanation:-Sequence numbers in amazon Kinesis are assigned by Amazon Kinesis when a data producer calls PutRecord operation to add data to an Amazon Kinesis stream. Sequence numbers are assigned by Amazon Kinesis when a data producer calls PutRecords operation to add data to an Amazon Kinesis stream. Sequence numbers for the same partition key generally increase over time.The longer the time period between PutRecord or PutRecords requests, the larger the sequence number becomes.

Reference: http://docs.aws

Only 2, 3 and 4

Only 1, 2 and 4

Q18) How are Snowball logs stored?

In a SQLite table

✅ In a plaintext file

Explanation:-When you transfer data between your data center and a Snowball, the Snowball client generates a plaintext log and saves it to your workstation.

Reference: http://docs.aws.amazon.com/snowball/latest/ug/using-client.html

In a JSON file

In an XML file

Q19) How do you put your data into a Snowball?

Connect your data source to the Snowball and then press the "import" button.

Mount your data source onto the Snowball and ship it back together with the appliance.

Connect the Snowball to your datacenter and then copy the data from your data sources to the appliance via FTP

✅ Mount your data source onto a workstation in your datacenter and then use this workstation to transfer data to the Snowball.

Explanation:-To put your data into a Snowball, you mount your data source onto a workstation in your datacenter and then use this workstation to transfer data to the Snowball.

Reference: http://docs.aws.amazon.com/snowball/latest/ug/receive-appliance.html

---

Q20) Kinesis Partition keys are unicoded strings with a maximum length of _____.

128 bytes

✅ 256 bytes

Explanation:-Kinesis Partition keys are unicoded strings with a maximum length of 256 bytes

Reference: http://docs.aws.amazon.com/streams/latest/dev/working-with-kinesis.html

512 bytes

1024 bytes

Q21) Identify a factor that affects the speed of data transfer in AWS Snowball.

- The speed of the AGP card
- ✅ Local network speed

Explanation:-The Snowball client can be used to estimate the time taken to transfer data. Data transfer speed is affected by a number of factors including local network speed, file size, and the speed at which data can be read from local servers.
Reference: https://aws.amazon.com/importexport/faqs/

- Transcoder speed
- The speed of the L3 cache

Q22) How can AWS Snowball handle petabyte-scale data migration?

- Data is sent encoded (forward error correction) via a high speed network connection
- Data is sent compressed via a high speed network connection.
- ✅ Data is sent via a physical appliance sent to you by AWS.

Explanation:-Snowball uses secure appliances to transfer large amounts of data into and out of the AWS cloud; this is fast and cheaper than high-speed Internet.
Reference: https://aws.amazon.com/snowball/

- Data is sent via a shipping container, pulled by a semi-trailer truck.

Q23) The maximum size of a Kinesis data blob, the data payload before Base64 encoding is _____.

- Five megabytes
- Two megabytes
- One kilobyte
- ✅ One megabyte

Explanation:-The maximum size of a Kinesis data blob, the data payload before Base64 encoding is one megabtye
Reference: http://docs.aws.amazon.com/streams/latest/dev/working-with-kinesis.html

---

Q24) The Snowball client uses a(n) ____ to define what kind of data is transferred between the client's data center and a Snowball.

- JSON configuration file
- ✅ schema

Explanation:-The Snowball client uses schemas to define what kind of data is transferred between the client's data center and a Snowball. The schemas are declared when a command is issued.
Reference: http://docs.aws.amazon.com/snowball/latest/ug/using-client.html

- interface
- XML configuration file

Q25) An AWS Snowball appliance includes a(n) ____ network connection to minimize data transfer times.

- 1000BaseT
- ✅ 10GBaseT

Explanation:-An AWS Snowball appliance has a 10GBaseT network connection (both RJ45 as well as SFP+ with either a fiber or copper interface) to minimize data transfer times. This allows the Snowball appliance to transfer up to 80 terabytes of data from a data source to the appliance in about a day, plus shipping time.
Reference: https://aws.amazon.com/snowball/details/

- 40GBaseT
- Infiniband

Q26) The job management API for AWS Snowball is a network protocol based on HTTP that uses a(n) ____ model.

- ✅ RPC

Explanation:-The job management API for AWS Snowball is a network protocol based on HTTP. It uses JSON (RFC 4627) documents for HTTP request/response bodies and is an RPC model, in which there is a fixed set of operations, and the syntax for each operation is known to clients without any prior interaction.
Reference: http://docs.aws.amazon.com/snowball/latest/api-reference/api-reference.html

- MPI
- publish/subscribe
- RMI

Q27)

Which statements are true about re-sharding in Amazon Kinesis?

1. The shard or pair of shards that result from the re-sharding operation are referred to as child shards.

2. When you re-shard, data records that were flowing to the parent shards are rerouted to flow to the child shards based on the hash key values that the data record partition keys map to.

3. The shard or pair of shards that the re-sharding operation acts on are referred to as parent shards.

4. After you call a re-sharding operation, you do not need to wait for the stream to become active again

- Only 2, 3 and 4
- ✅ Only 1, 2 and 3

Explanation:-Kinesis Streams supports re-sharding which enables you to adjust the number of shards in your stream in order to adapt to changes in the rate of data flow through the stream. The shard or pair of shards that the re-sharding operation acts on are referred to as parent shards. The

shard or pair of shards that result from the re sharding operation are referred to as child shards. After you call a re-sharding operation, you need to wait for the stream to become active again. When you re-shard, data

- Only 1, 3 and 4
- Only 1, 2 and 4

Q28) In AWS Data Pipeline, an activity is (choose one)

- A set of scripts loaded at run time
- The database schema of the pipeline data
- ✓ A pipeline component that defines the work to perform
- A read/ write event from the primary database

Q29)

The operations team and the development team want a single place to view both operating system and application logs.

How should you implement this using AWS services?

1. Using AWS CloudFormation, create a CloudWatch Logs LogGroup and send the operating system and application logs of interest using the CloudWatch Logs Agent

2. Using AWS CloudFormation and configuration management, set up remote logging to send events via UDP packets to CloudTrail

3. Using configuration management, set up remote logging to send events to Amazon Kinesis and insert these into Amazon CloudSearch or Amazon Redshift, depending on available analytic tools

4. Using AWS CloudFormation, create a CloudWatch Logs LogGroup. Because the CloudWatch log agent automatically sends all operating system logs, you only have to configure the application logs for sending off-machine

- Only 2 and 4
- ✓ Only 1 and 3
- Only 2 and 3
- Only 1 and 2

Q30)

You are working with customer who has 10 TB of archival data that they want to migrate to Amazon Glacier.

The customer has a 1Mbps connection to the Internet.

Which service or feature provide the fastest method of getting the data into Amazon Glacier?

- VM Import/Export
- AWS Storage Gateway
- Amazon Glacier multipart upload
- ✓ AWS Import/Export

Q31)

A user has provisioned 2000 IOPS to the EBS volume. The application hosted on that EBS is experiencing less IOPS than provisioned.

Which of the below mentioned options does not affect the IOPS of the volume?

- The EC2 instance has 10 Gigabit Network connectivity
- The instance is EBS optimized
- The application does not have enough IO for the volume
- ✓ The volume size is too large

Q32)

You want to securely distribute credentials for your Amazon RDS instance to your fleet of web server instances. The credentials are stored in a file that is controlled by a configuration management system.

How do you securely deploy the credentials in an automated manner across the fleet of web server instances, which can number in the hundreds, while retaining the ability to roll back if needed?

- ✓ Keep credential files as a binary blob in an Amazon RDS MySQL DB instance, and have a script on each Amazon EC2 instance that pulls the files down from the RDS instance
- Insert credential files into user data and use an instance lifecycle policy to periodically refresh the files from the user data
- Store the credential files in your version-controlled repository with the rest of your code. Have a post-commit action in version control that kicks off a job in your continuous integration system which securely copies the new credentials files to all web server instances
- Store your credential files in an Amazon S3 bucket. Use Amazon S3 server-side encryption on the credential files. Have a scheduled job that pulls down the credential files into the instances every 10 minutes

---

Q33)

A us-based company is expanding their web presence into Europe.

The company wants to extend their AWS infrastructure from Northern Virginia (us-east-1) into the Dublin (eu-west-1) region.

Which of the following options would enable an equivalent experience for users on both continents?

- Use Amazon Route S3, and apply a weighted routing policy to distribute traffic across both regions
- ✓ Use Amazon Route S3, and apply a geolocation routing policy to distribution traffic across both regions
- Use a public-facing load balancer per region to load balancer web traffic, and enable sticky sessions
- Use a public-facing load balancer per region to load-balancer web traffic, and enable HTTP health checks

---

Q34)

You need to configure an Amazon S3 bucket to serve static assets for your public-facing web application.

Which methods ensure that all objects uploaded to the bucket are set to public read?

1. Set permissions on the object to public read during upload

2. Configure the bucket ACL to sell all objects to public read

3. Configure the bucket policy to set all objects to public read

4. Use AWS identity and access Management roles to set the bucket to public read

- Only 2 and 4
- Only 1 and 4
- ✓ Only 2 and 3
- Only 1 and 2

Q35)

You have started a new job and are reviewing your company's infrastructure on AWS. You notice one web application where they have an Elastic Load Balancer (&B) in front of web instances in an Auto Scaling Group. When you check the metrics for the ELB in CloudWatch you see four healthy instances in Availability Zone (AZ) A and zero in AZ B. There are zero unhealthy instances.

What do you need to fix to balance the instances across AZs?

- C. Make sure your AMI is available in both AZs
- ✓ B. Make sure Auto Scaling is configured to launch in both AZs
- A. Set the ELB to only be attached to another AZ
- D. Make sure the maximum size of the Auto Scaling Group is greater than 4

Q36)

You have a large number of web servers in an Auto Scaling group behind a load balancer. On an hourly basis, you want to filter and process the logs to collect data on unique visitors, and then put that data in a durable data store in order to run reports. Web servers in the Auto Scaling group are constantly launching and terminating based on your scaling policies, but you do not want to lose any of the log data from these servers during a stop/termination initiated by a user or by Auto Scaling.

What two approaches will meet these requirements?

1. Install an Amazon CloudWatch Logs Agent on every web server during the bootstrap process. Create a CloudWatch log group and define metric Filters to create custom metrics that track unique visitors from the streaming web server logs. Create a scheduled task on an Amazon EC2 instance that runs every hour to generate a new report based on the CloudWatch custom metrics

2. On the web servers, create a scheduled task that executes a script to rotate and transmit the logs to Amazon Glacier. Ensure that the operating system shutdown procedure triggers a logs transmission when the Amazon EC2 instance is stopped/terminated. Use Amazon Data pipeline to process data in Amazon Glacier and run reports every hour.

3. On the web servers, create a scheduled task that executes a script to rotate and transmit the logs to an Amazon S3 bucket. Ensure that the operating system shutdown process triggers a logs transmission when the Amazon EC2 instance is stopped/terminated. Use AWS Data Pipeline to move log data from the Amazon S3 bucket to Amazon Redshift in order to process and run reports every hour

4. Install an AWS Data Pipeline Logs Agent on every web server during the bootstrap process. Create a log group object in AWS Data Pipeline, and define Metric filters to move processed log data directly from the web servers to Amazon Redshift and runs reports every hour.

- ✓ Only 1 and 3
- Only 2 and 4
- Only 2 and 3
- Only 1 and 2

Q37)

In AWS, which security aspects are the customer's responsibility?

1. Life-Cycle management of IAM credentials

2. Security Group and ACL settings

3. Encryption of EBS volumes

4. Path management on the EC2 instance's operating system

- ✓ All of these
- Only 1 and 4
- Only 2 and 3
- Only 1 and 2

Q38)

A photo-sharing service stores pictures in Amazon Simple Storage Service (S3) and allows application sign-in using an opened connect-compatible identity provider.

Which AWS Security Token Service approach to temporary access should you use for the Amazon S3 operations?

- ✅ SAML-based Identity Federation
- AWS identity and Access Management roles
- Cross-Account Access
- Web identity Federation

Q39)

You have identified network throughput as a bottleneck on your m1.small EC2 instance when uploading data Into Amazon S3 In the same region.

How do you remedy this situation?

- Use DirectConnect between EC2 and S3
- ✅ Change to a larger Instance
- Add an additional ENI
- Use EBS PIOPS on the local volume

Q40)

The project you are working on currently uses a single AWS CloudFormation template to deploy its AWS infrastructure, which supports a multi-tier web application.

You have been tasked with organizing the AWS CloudFormation resources so that they can be maintained in the future, and so that different departments such as Networking and Security can review the architecture before it goes to Production.

How should you do this in a way that accommodates each department, using their existing workflows?

- Use a custom application and the AWS SDK to replicate the resources defined in the current AWS CloudFormation template, and use the existing code review system to allow other departments to approve changes before altering the application for future deployments.
- Organize the AWS CloudFormation template so that related resources are next to each other in the template for each department's use, leverage your existing continuous integration tool to constantly deploy changes from all parties to the Production environment, and then run tests for validation.
- Organize the AWS CloudFormation template so that related resources are next to each other in the template, such as VPC subnets and routing rules for Networking and Security groups and IAM information for Security
- ✅ Separate the AWS CloudFormation template into a nested structure that has individual templates for the resources that are to be governed by different departments, and use the outputs from the networking and security stacks for the application template that you control

---

Q41)

You have launched an Amazon Elastic Compute Cloud (EC2) instance into a public subnet with a primary private IP address assigned, an internet gateway is attached to the VPC, and the public route table is configured to send all internet-based internet.

Why is the internet unreachable from this instance?

- The instance "Source/Destination check" property must be enabled
- ✅ The instance does not have a public IP address
- The Internet gateway security group must allow all outbound traffic
- The instance security group must allow all inbound traffic

Q42)

A company needs to deploy virtual desktops to its customers in a virtual private cloud, leveraging existing security controls.

Which set of AWS services and features will meet the company's requirements?

- AWS Directory service, Amazon WorkSpaces, and AWS Identity and Access Management
- ✅ Virtual private network connection, AWS Directory services, and Amazon WorkSpaces
- Virtual private network connection, AWS Directory services, and ClassicLink
- Amazon Elastic Compute Cloud, and AWS identity and access management

---

Q43)

You are currently hosting multiple applications in a VPC and have logged numerous port scans coming in from a specific IP address block.

Your security team has requested that all access from the offending IP address block be denied for the next 24 hours.

Which of the following is the best method to quickly and temporarily deny access from the specified IP address block?

- Modify the Windows Firewall settings on all Amazon Machine Images (AMIs) that your organization uses in that VPC to deny access from the IP address block
- Add a rule to all of the VPC 5 Security Groups to deny access from the IP address block
- ✅ Modify the Network ACLs associated with all public subnets in the VPC to deny access from the IP address block

- Create an AD policy to modify Windows Firewall settings on all hosts in the VPC to deny access from the IP address block

---

Q44)

A us-based company is expanding their web presence into Europe.

The company wants to extend their AWS infrastructure from Northern Virginia (us-east-1) into the Dublin (eu-west-1) region.

Which of the following options would enable an equivalent experience for users on both continents?

- Use Amazon Route S3, and apply a weighted routing policy to distribute traffic across both regions
- ✓ Use Amazon Route S3, and apply a geolocation routing policy to distribution traffic across both regions
- Use a public-facing load balancer per region to load balancer web traffic, and enable sticky sessions
- Use a public-facing load balancer per region to load-balancer web traffic, and enable HTTP health checks

---

Q45)

You have started a new job and are reviewing your company's infrastructure on AWS. You notice one web application where they have an Elastic Load Balancer (&B) in front of web instances in an Auto Scaling Group When you check the metrics for the ELB in CloudWatch you see four healthy instances in Availability Zone (AZ) A and zero in AZ B There are zero unhealthy instances.

What do you need to fix to balance the instances across AZs?

- Make sure your AMI is available in both AZs
- ✓ Make sure Auto Scaling is configured to launch in both AZs
- Set the ELB to only be attached to another AZ
- Make sure the maximum size of the Auto Scaling Group is greater than 4

Q46)

You have a large number of web servers in an Auto Scaling group behind a load balancer. On an hourly basis, you want to filter and process the logs to collect data on unique visitors, and then put that data in a durable data store in order to run reports.

Web servers in the Auto Scaling group are constantly launching and terminating based on your scaling policies, but you do not want to lose any of the log data from these servers during a stop/termination initiated by a user or by Auto Scaling.

What two approaches will meet these requirements? (Choose 2 answers)

- ✓ Install an AWS Data Pipeline Logs Agent on every web server during the bootstrap process. Create a log group object in AWS Data Pipeline, and define Metric filters to move processed log data directly from the web servers to Amazon Redshift and runs reports every hour
- On the web servers, create a scheduled task that executes a script to rotate and transmit the logs to an Amazon S3 bucket. Ensure that the operating system shutdown process triggers a logs transmission when the Amazon EC2 instance is stopped/terminated. Use AWS Data Pipeline to move log data from the Amazon S3 bucket to Amazon Redshift in order to process and run reports every hour
- On the web servers, create a scheduled task that executes a script to rotate and transmit the logs to Amazon Glacier. Ensure that the operating system shutdown procedure triggers a logs transmission when the Amazon EC2 instance is stopped/terminated. Use Amazon Data pipeline to process data in Amazon Glacier and run reports every hour
- Install an Amazon CloudWatch Logs Agent on every web server during the bootstrap process. Create a CloudWatch log group and define metric Filters to create custom metrics that track unique visitors from the streaming web server logs. Create a scheduled task on an Amazon EC2 instance that runs every hour to generate a new report based on the CloudWatch custom metrics.

Q47)

In AWS, which security aspects are the customer's responsibility?

- Encryption of EBS volumes
- Security Group and ACL settings
- ✓ Life-Cycle management of IAM credentials
- Path management on the EC2 instance's operating system

Q48)

A data engineer in a manufacturing company is designing a data processing platform that receives a large volume of unstructured data.

The data engineer must populate a well-structured starschema in Amazon Redshift.

What is the most efficient architecture strategy for this purpose?

- Normalize the data using an AWS Marketplace ETL tool, persist the results to Amazon S3, anduse AWS Lambda to INSERT the data into Redshift.
- ✓ When the data is saved to Amazon S3, use S3 Event Notifications and AWS Lambda to transformthe file contents. Insert the data into the analysis schema on Redshift.
- Load the unstructured data into Redshift, and use string parsing functions to extract structureddata for inserting into the analysis schema.
- Transform the unstructured data using Amazon EMR and generate CSV datCOPY the CSVdata into the analysis schema within Redshift.

Q49)

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 Clientto run analysis against the text index.
- Use Amazon EMR to parallelize the text analysis tasks across the cluster using a streamingprogram step.
- ✓ Copy the data into Amazon ElastiCache to perform text analysis on the in-memory data and exportthe 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.

Q50)

A data engineer chooses Amazon DynamoDB as a data store for a regulated application. This application must be submitted to regulators for review. The data engineer needs to provide a control framework that lists the security controls from the process to follow to add new users down to the physical controls of the data center, including items like security guards and cameras.

How should this control mapping be achieved using AWS?

1. Request AWS third-party audit reports and/or the AWS quality addendum and map the AWS responsibilities to the controls that must be provided.

2. Request data center Temporary Auditor access to an AWS data center to verify the control mapping.

3. Request relevant SLAs and security guidelines for Amazon DynamoDB and define these guidelines within the application's architecture to map to the control framework.

4. Request Amazon DynamoDB system architecture designs to determine how to map the AWS responsibilities to the control that must be provided.

- Only 2, 3 and 4
- ✓ Only 1, 3 and 4
- Only 1, 2 and 3
- All of these

Q51)

Let us suppose 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 MapReduce 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?

1. Anonymize the personal data portions of the dataset and transfer the data files into Amazon S3 the AWS region. Have the EMR cluster read the dataset using EMRFS.

2. 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.

3. 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.

4. 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.

- Only 1 and 3
- ✓ Only 2 and 4
- Only 2 and 3
- Only 1 and 2

Q52)

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

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

In which two circumstances would choosing EVEN distribution be most appropriate? (Choose two.)

- When a new table has been loaded and it is unclear how it will be joined to dimension.
- When data transfer between nodes must be eliminated.
- ✓ When data must be grouped based on a specific key on a defined slice.
- When the tables are highly denormalized and do NOT participate in frequent joins.

Q53)

A large grocery distributor receives daily depletion reports from the field in the form of gziparchives 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.

Q54)

A web-hosting company is building a web analytics tool to capture click stream 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 KinesisPutRecord API. Use the exponential back-off algorithm for retries until a successful response isreceived.
- 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.

Q55)

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 Server Name, from which Object Name 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 Keyand 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 ServerNameas Sort Key. Define a Global Secondary Index with ServerName as partition key and TimeStampfollowed by StreamName.

Q56)

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 outputManifest 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

Q57)

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 of 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 theDynamoDB 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 DynamoDBStreams as every new item description is added to the system.
- Build a Kinesis Streams process that captures and marks the relevant items in the dangerousgoods 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 onthe DynamoDB Streams as every new item description is added to the system.

Q58)

A company receives data sets coming from external providers on Amazon S3. Data sets from different providers are dependent on one another. Data sets will arrive at different times and in no particular order.

A data architect needs to design a solution that enables the company to do the following -

Rapidly perform cross data set analysis as soon as the data becomes available

Manage dependencies between data sets that arrive at different times

Which architecture strategy offers a scalable and cost-effective solution that meets these requirements?

✅ Maintain data dependency information in an Amazon DynamoDB table. Use Amazon S3 event notifications to trigger an AWS Lambda function that maps the S3 object to the task associated with it in DynamoDB. Once all task dependencies have been resolved, process the data with Amazon EMR.

⚪ Maintain data dependency information in an Amazon ElastiCache Redis cluster. Use Amazon S3event notifications to trigger an AWS Lambda function that maps the S3 object to Redis. Once the task dependencies have been resolved, process the data with Amazon EMR.

⚪ Maintain data dependency information in Amazon RDS for MySQL. Use an AWS Data Pipeline job to load an Amazon EMR Hive table based on task dependencies and event notification triggers in Amazon S3.

⚪ Maintain data dependency information in an Amazon DynamoDB table. Use Amazon SNS andevent notifications to publish data to fleet of Amazon EC2 workers. Once the task dependencies have been resolved, process the data with Amazon EMR.

Q59)

A media advertising company handles a large number of real-time messages sourced from over 200 websites in real time. Processing latency must be kept low. Based on calculations, a 60-shard Amazon Kinesis stream is more than sufficient to handle the maximum data through put, even with traffic spikes. The company also uses an Amazon Kinesis Client Library (KCL) application running on Amazon Elastic Compute Cloud (EC2) managed by an Auto Scaling group.

Amazon Cloud Watch indicates an average of 25% CPU and a modest level of network traffic across all running servers. The company reports a 150% to 200% increase in latency of processing messages from Amazon Kinesis during peak times. There are NO reports of delay from the sites publishing to Amazon Kinesis.

What is the appropriate solution to address the latency?

⚪ Increase the minimum number of instances in the Auto Scaling group.
⚪ Increase the size of the Amazon EC2 instances to increase network throughput.
⚪ Increase the number of shards in the Amazon Kinesis stream to 80 for greater concurrency.
✅ Increase Amazon DynamoDB throughput on the checkpoint table.

Q60)

A Redshift data warehouse has different user teams that need to query the same table with very different query types.

These user teams are experiencing poor performance.

Which action improves performance for the user teams in this situation?

⚪ Maintain team-specific copies of the table.
✅ Add interleaved sort keys per team.
⚪ Create custom table views.
⚪ Add support for workload management queue hopping.
