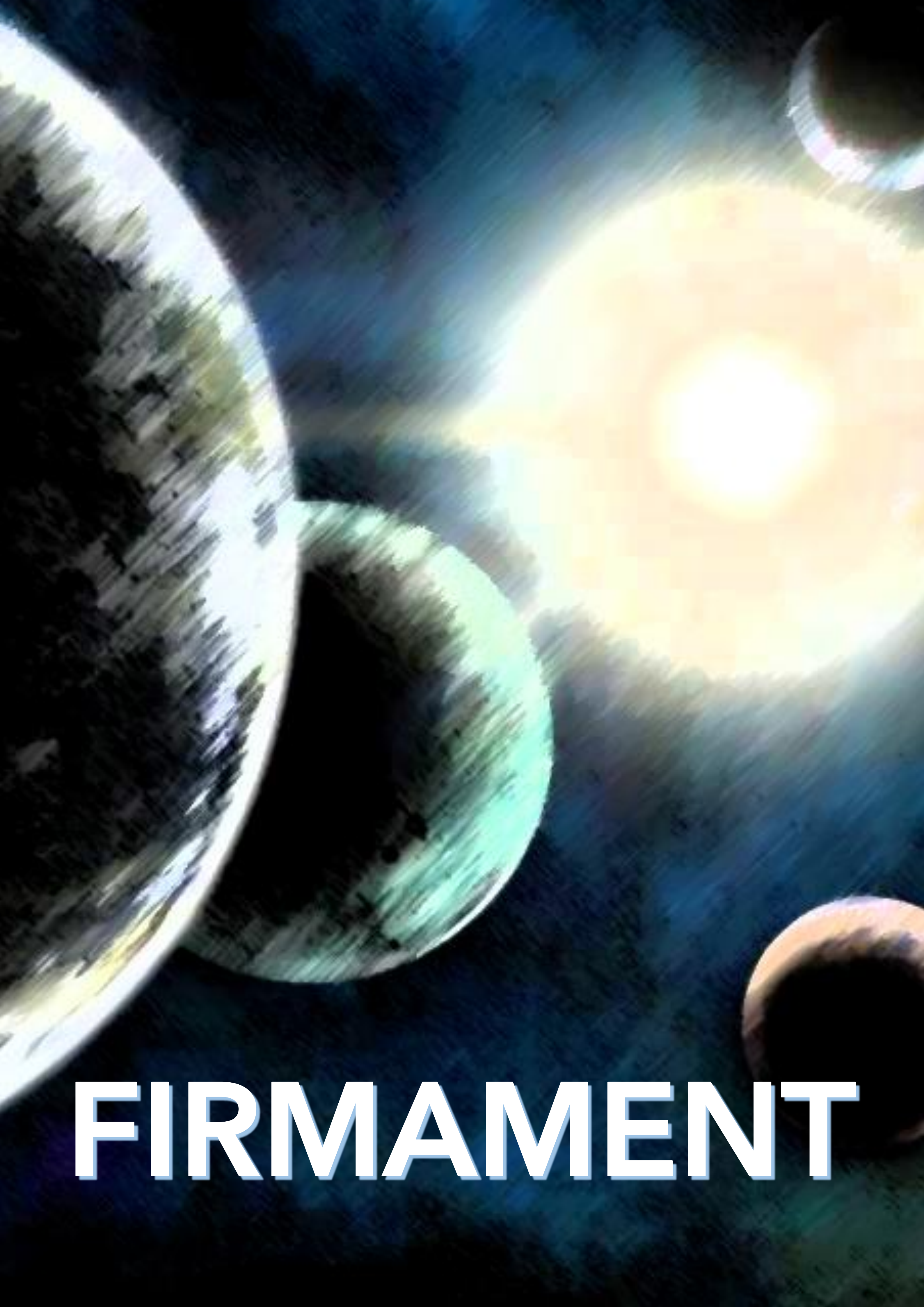


FIRMAMENT

Contents

2013	Management Console Mobile Application	Elastic Transcoder	OpsWorks	CloudHSM
	AppStream	CloudTrail		
	WorkSpaces	AWS Certification	Kinesis	

Cover image credit: (Getty Images/iStockphoto)

This image was, for example, used in an article in The Independent newspaper on April 12th 2022. The article was entitled “China’s ‘Earth 2.0’ Spacecraft could finally find life on alien worlds”.

Management Console

Bibliography

I. Official

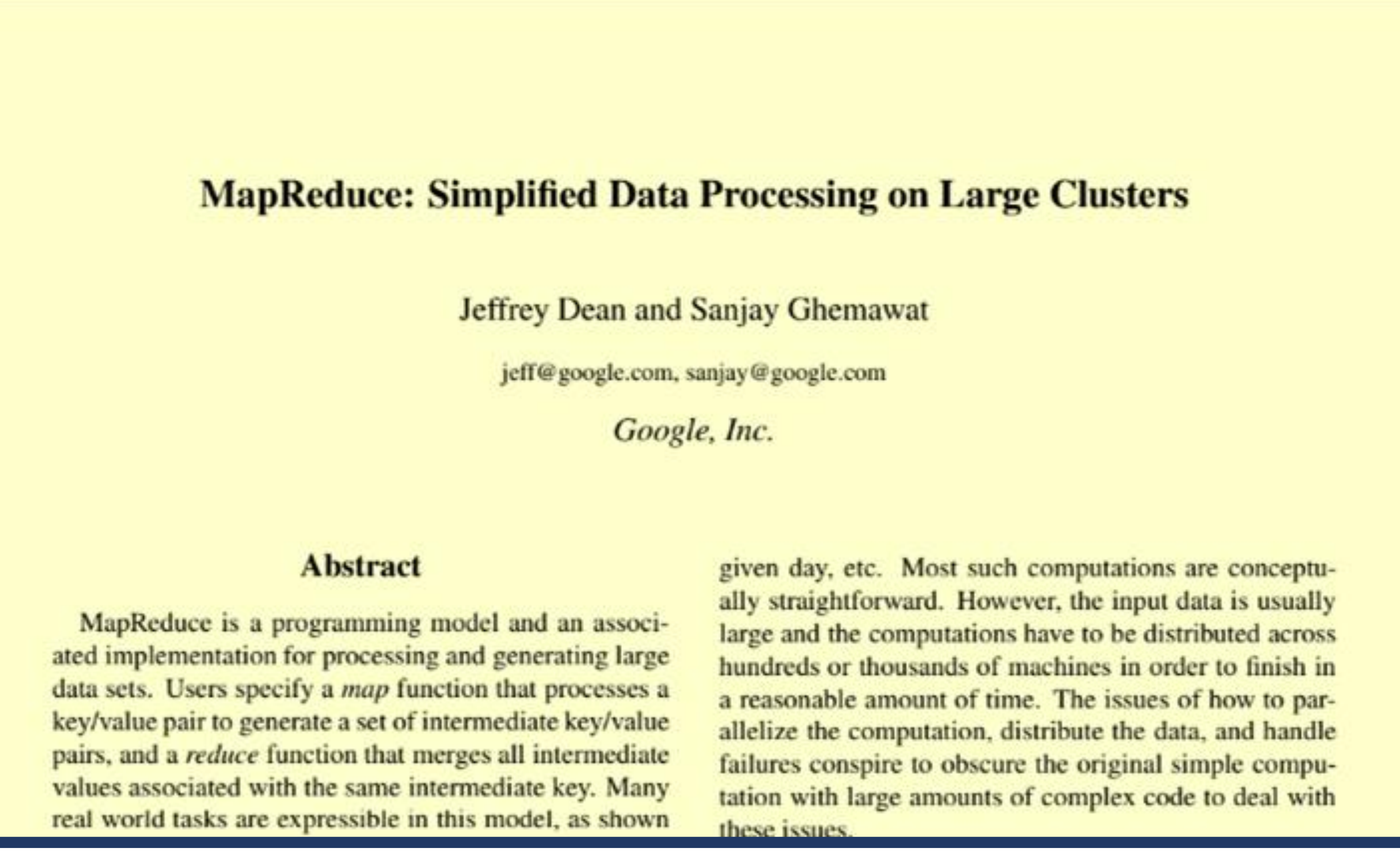
II. Unofficial

https://jakehendy.com/2023/02/05/What-does-the-console-actually-do/?ck_subscriber_id=1560524742

Elastic Map Reduce

CONTENTS

Text goes here.



MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

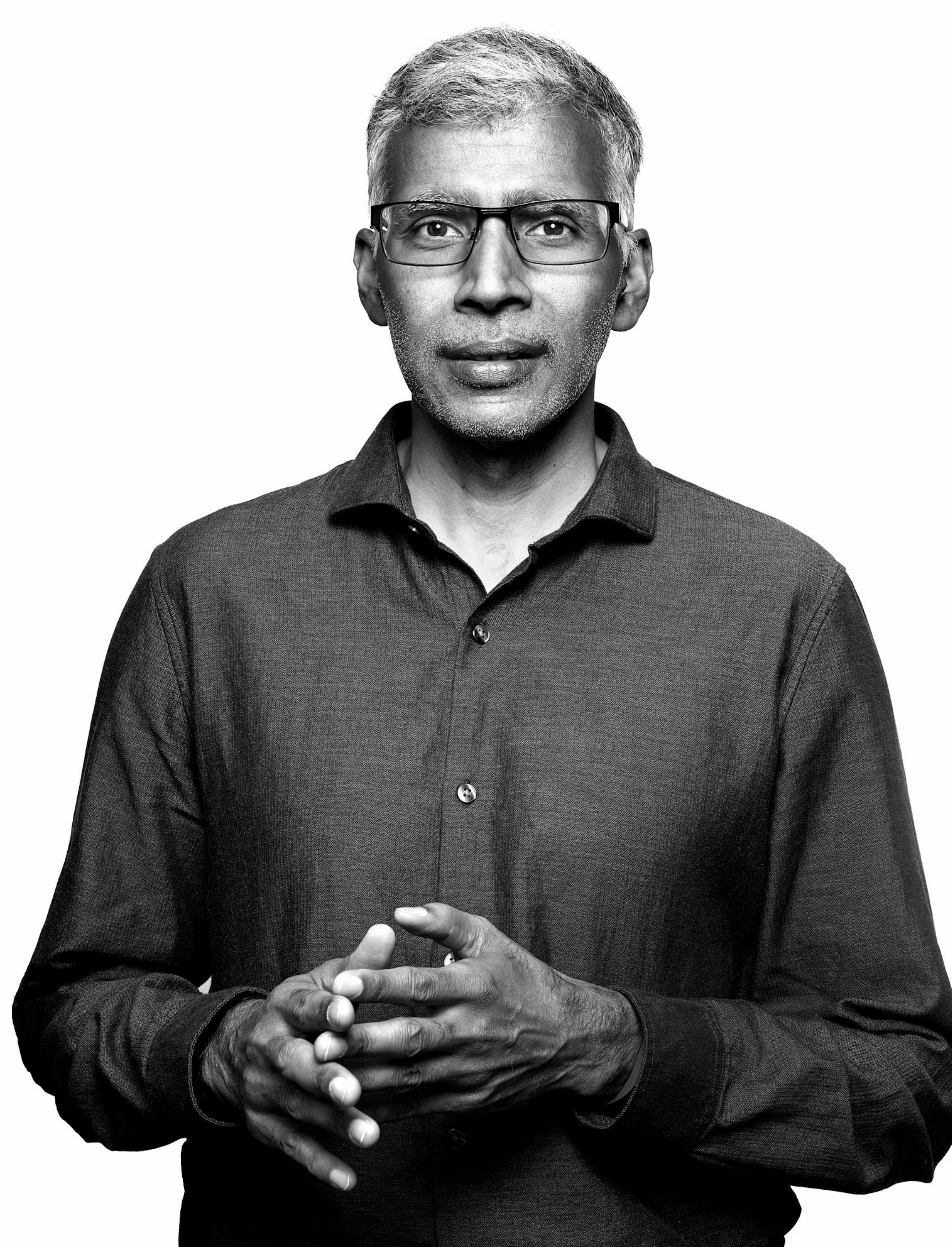
Google, Inc.

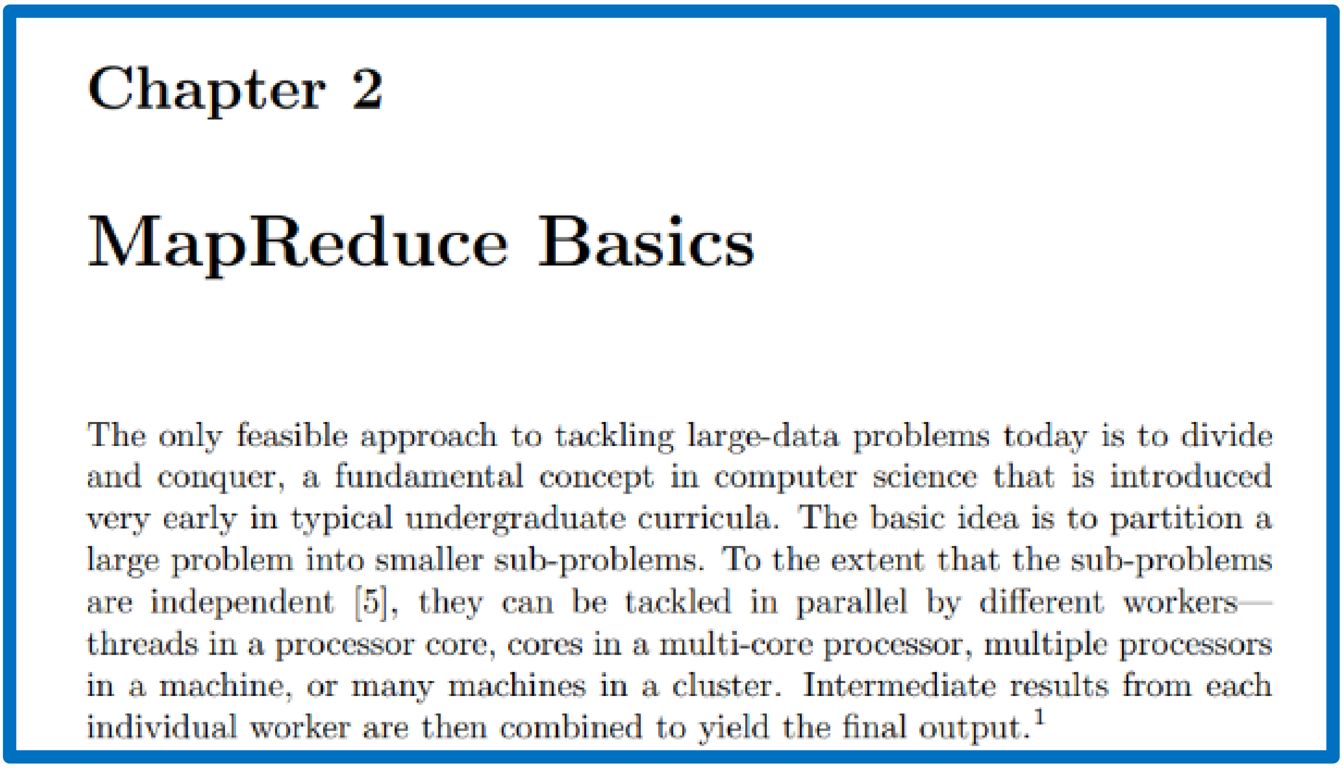
Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

Sanjay Ghemawat

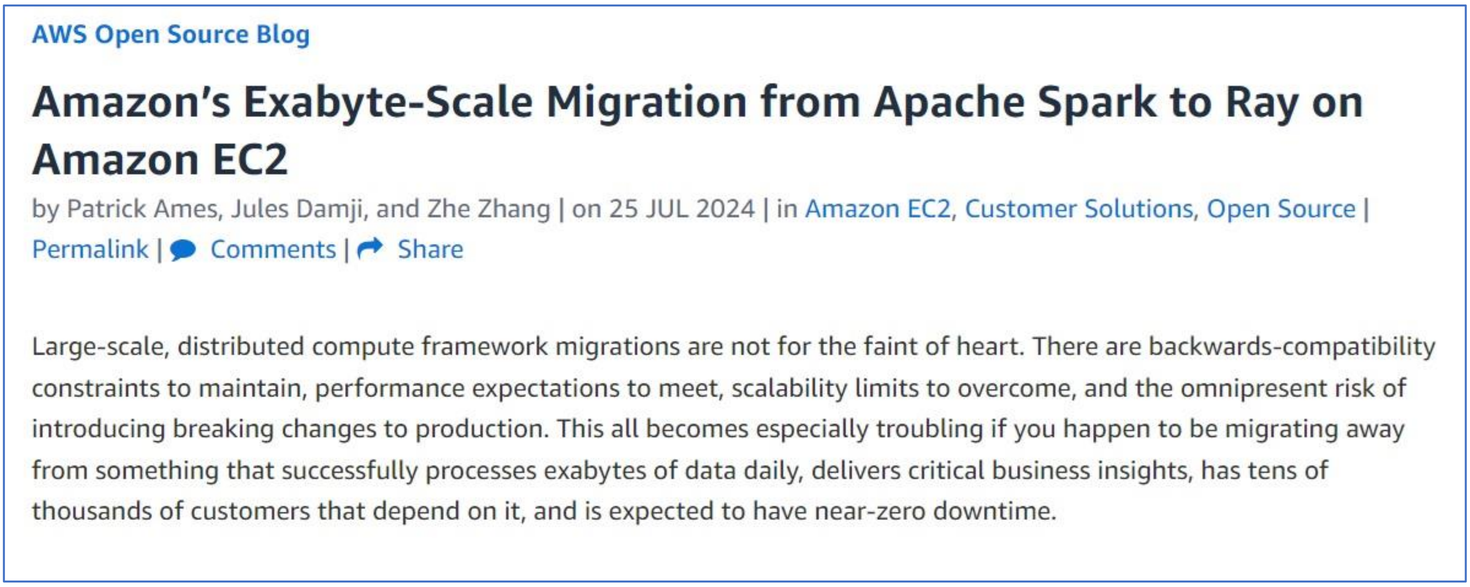




Chapter 2

MapReduce Basics

The only feasible approach to tackling large-data problems today is to divide and conquer, a fundamental concept in computer science that is introduced very early in typical undergraduate curricula. The basic idea is to partition a large problem into smaller sub-problems. To the extent that the sub-problems are independent [5], they can be tackled in parallel by different workers—threads in a processor core, cores in a multi-core processor, multiple processors in a machine, or many machines in a cluster. Intermediate results from each individual worker are then combined to yield the final output.[1]

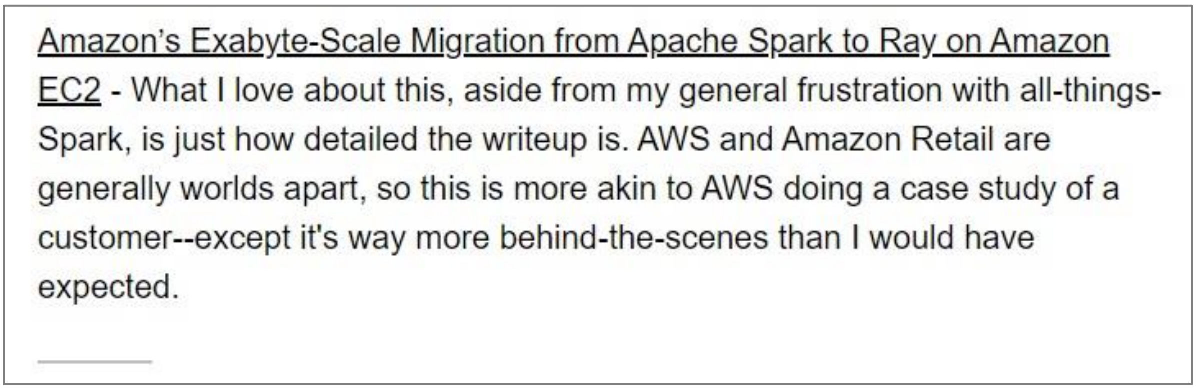


AWS Open Source Blog

Amazon’s Exabyte-Scale Migration from Apache Spark to Ray on Amazon EC2

by Patrick Ames, Jules Damji, and Zhe Zhang | on 25 JUL 2024 | in Amazon EC2, Customer Solutions, Open Source | Permalink | Comments | Share

Large-scale, distributed compute framework migrations are not for the faint of heart. There are backwards-compatibility constraints to maintain, performance expectations to meet, scalability limits to overcome, and the omnipresent risk of introducing breaking changes to production. This all becomes especially troubling if you happen to be migrating away from something that successfully processes exabytes of data daily, delivers critical business insights, has tens of thousands of customers that depend on it, and is expected to have near-zero downtime.



Amazon’s Exabyte-Scale Migration from Apache Spark to Ray on Amazon EC2 - What I love about this, aside from my general frustration with all-things-Spark, is just how detailed the writeup is. AWS and Amazon Retail are generally worlds apart, so this is more akin to AWS doing a case study of a customer--except it's way more behind-the-scenes than I would have expected.

Corey Quinn (or his stand in, while he in on vacation) writing on July 29th 2024

TPN

1. ***Phenomenon1*** – the tendency of X to Y.
2. ***Phen2*** – the tendency of X to Y.
3. ***Phen3*** – the tendency of X to Y.
4. ***Phen4*** – the tendency of X to Y.
5. ***Phen5*** – the tendency of X to Y.
6. ***Phen6*** – the tendency of X to Y.
7. ***Phen7*** – the tendency of X to Y.
8. ***Phen8*** – the tendency of X to Y.
9. ***Phen9*** – the tendency of X to Y.
10. ***Phen10*** – the tendency of X to Y.

Glossary

Term1

Description of what term means here.

Term2

Description of what term means here.

Term3

Description of what term means here.

Bibliography

I. Official
II. Unofficial
III. Critical
IV. General

I. Official

[Ames 2024]

Ames, Patrick and Jules Damji and Zhe Zhang (2024). Amazon’s Exabyte-Scale Migration from Apache Spark to Ray on Amazon EC2. *AWS Open Source Blog* [Blog]. July 25th 2024. Available at:

<https://aws.amazon.com/blogs/opensource/amazons-exabyte-scale-migration-from-apache-spark-to-ray-on-amazon-ec2/>

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

II. Unofficial

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

IV. General

[Surname1]

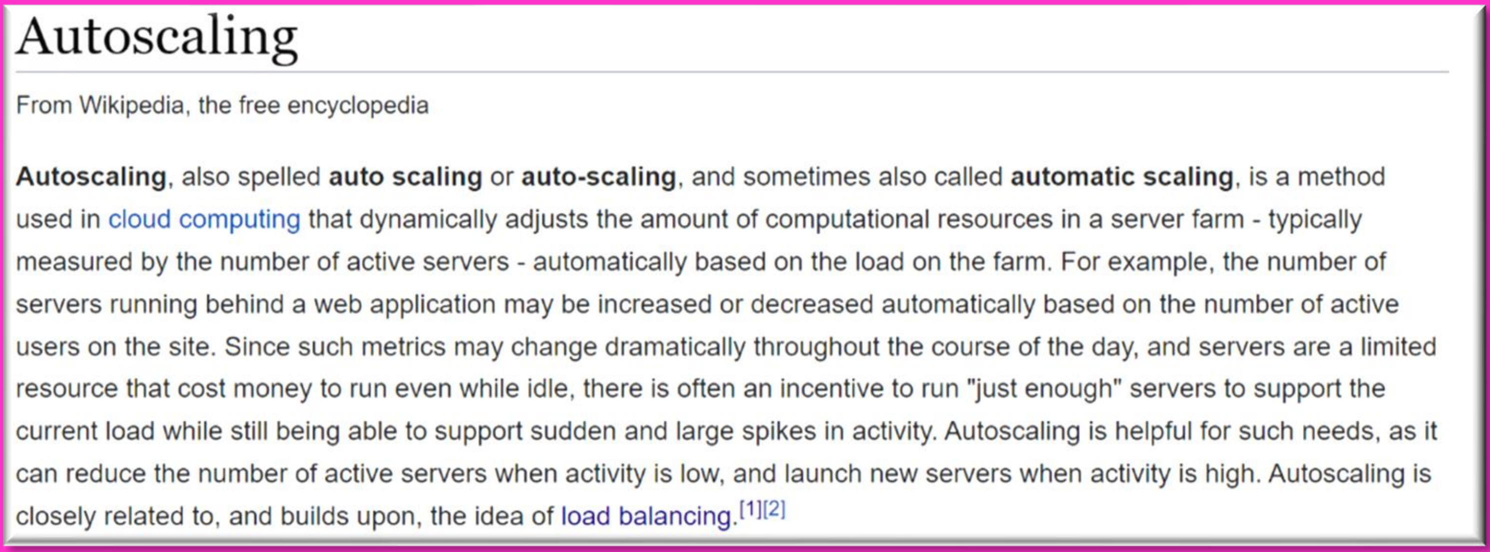
Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

Auto Scaling

Notice how AWS write it as two, distinct words. There is a space in between the two words. (It is not a funky brand, like GuardDuty, CloudWatch, or WorkMail).

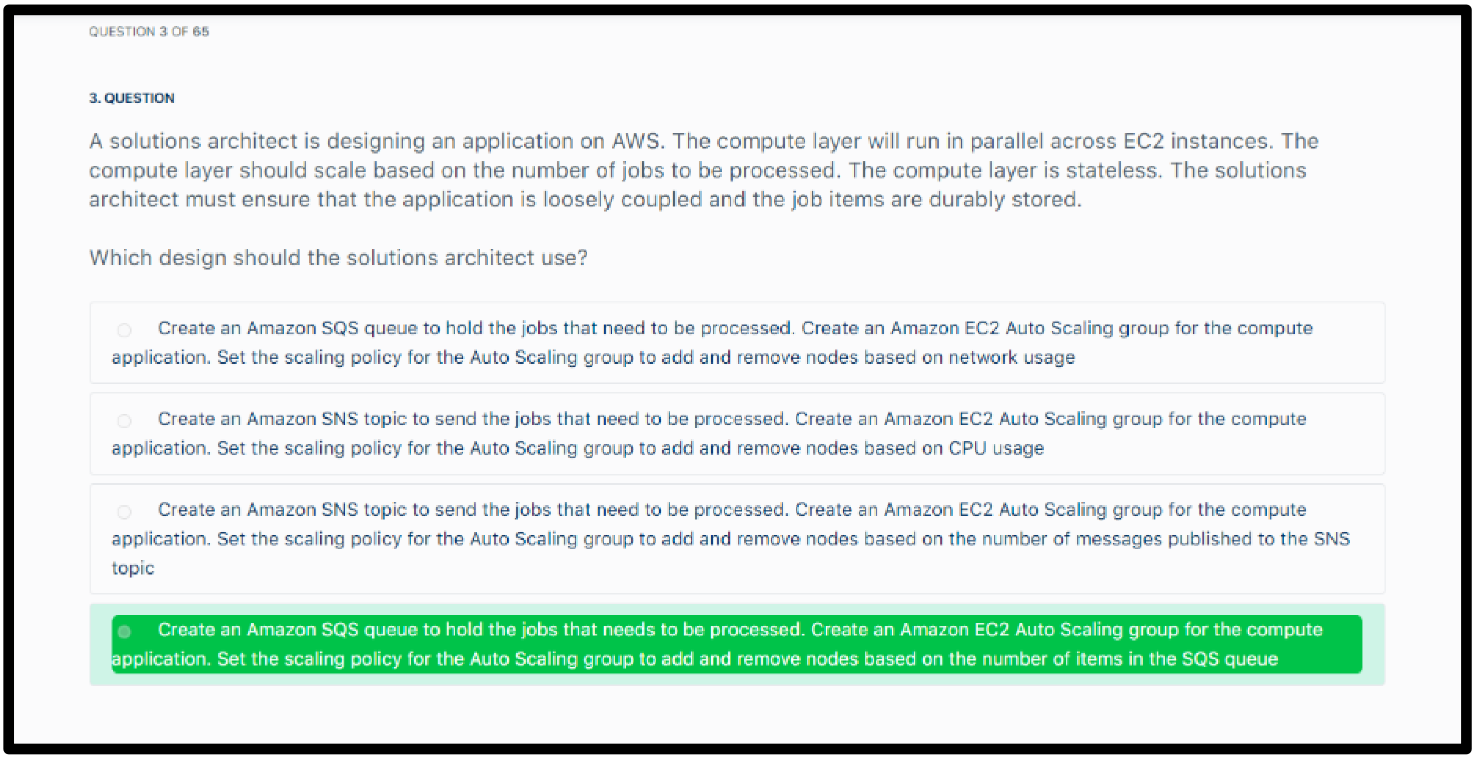


Autoscaling

From Wikipedia, the free encyclopedia

Autoscaling, also spelled **auto scaling** or **auto-scaling**, and sometimes also called **automatic scaling**, is a method used in cloud computing that dynamically adjusts the amount of computational resources in a server farm - typically measured by the number of active servers - automatically based on the load on the farm. For example, the number of servers running behind a web application may be increased or decreased automatically based on the number of active users on the site. Since such metrics may change dramatically throughout the course of the day, and servers are a limited resource that cost money to run even while idle, there is often an incentive to run "just enough" servers to support the current load while still being able to support sudden and large spikes in activity. Autoscaling is helpful for such needs, as it can reduce the number of active servers when activity is low, and launch new servers when activity is high. Autoscaling is closely related to, and builds upon, the idea of load balancing.[1][2]

AUTO SCALING



QUESTION 3 OF 65

3. QUESTION

A solutions architect is designing an application on AWS. The compute layer will run in parallel across EC2 instances. The compute layer should scale based on the number of jobs to be processed. The compute layer is stateless. The solutions architect must ensure that the application is loosely coupled and the job items are durably stored.

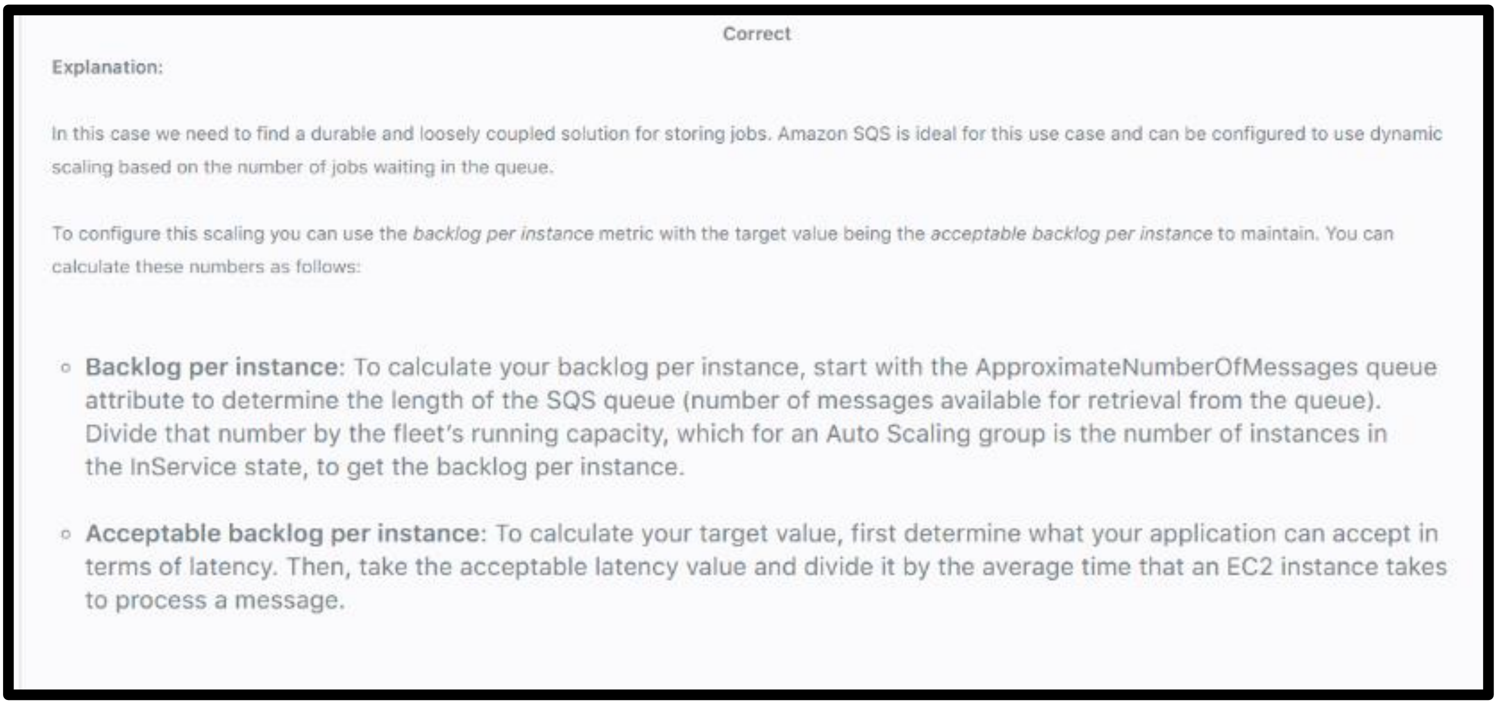
Which design should the solutions architect use?

○ Create an Amazon SQS queue to hold the jobs that need to be processed. Create an Amazon EC2 Auto Scaling group for the compute application. Set the scaling policy for the Auto Scaling group to add and remove nodes based on network usage

○ Create an Amazon SNS topic to send the jobs that need to be processed. Create an Amazon EC2 Auto Scaling group for the compute application. Set the scaling policy for the Auto Scaling group to add and remove nodes based on CPU usage

○ Create an Amazon SNS topic to send the jobs that need to be processed. Create an Amazon EC2 Auto Scaling group for the compute application. Set the scaling policy for the Auto Scaling group to add and remove nodes based on the number of messages published to the SNS topic

● Create an Amazon SQS queue to hold the jobs that needs to be processed. Create an Amazon EC2 Auto Scaling group for the compute application. Set the scaling policy for the Auto Scaling group to add and remove nodes based on the number of items in the SQS queue



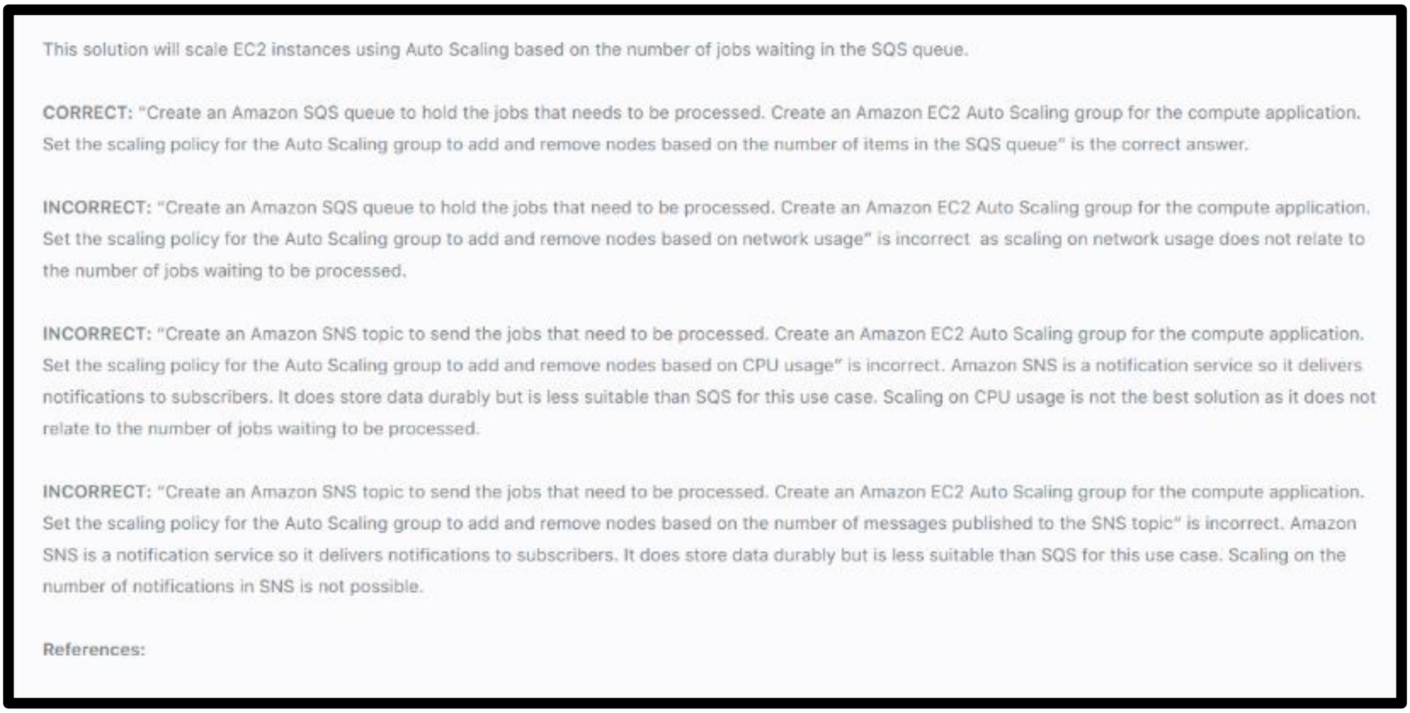
Correct

Explanation:

In this case we need to find a durable and loosely coupled solution for storing jobs. Amazon SQS is ideal for this use case and can be configured to use dynamic scaling based on the number of jobs waiting in the queue.

To configure this scaling you can use the *backlog per instance* metric with the target value being the *acceptable backlog per instance* to maintain. You can calculate these numbers as follows:

- **Backlog per instance:** To calculate your backlog per instance, start with the ApproximateNumberOfMessages queue attribute to determine the length of the SQS queue (number of messages available for retrieval from the queue). Divide that number by the fleet's running capacity, which for an Auto Scaling group is the number of instances in the InService state, to get the backlog per instance.
- **Acceptable backlog per instance:** To calculate your target value, first determine what your application can accept in terms of latency. Then, take the acceptable latency value and divide it by the average time that an EC2 instance takes to process a message.



This solution will scale EC2 instances using Auto Scaling based on the number of jobs waiting in the SQS queue.

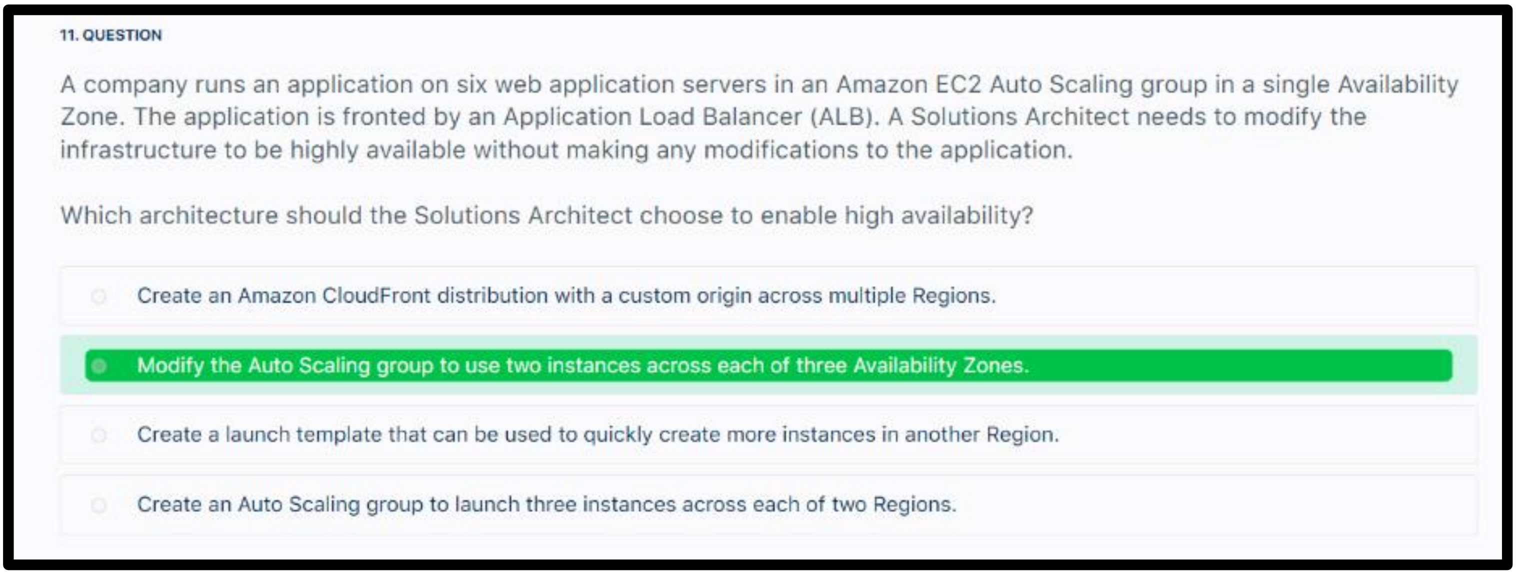
CORRECT: "Create an Amazon SQS queue to hold the jobs that needs to be processed. Create an Amazon EC2 Auto Scaling group for the compute application. Set the scaling policy for the Auto Scaling group to add and remove nodes based on the number of items in the SQS queue" is the correct answer.

INCORRECT: "Create an Amazon SQS queue to hold the jobs that need to be processed. Create an Amazon EC2 Auto Scaling group for the compute application. Set the scaling policy for the Auto Scaling group to add and remove nodes based on network usage" is incorrect as scaling on network usage does not relate to the number of jobs waiting to be processed.

INCORRECT: "Create an Amazon SNS topic to send the jobs that need to be processed. Create an Amazon EC2 Auto Scaling group for the compute application. Set the scaling policy for the Auto Scaling group to add and remove nodes based on CPU usage" is incorrect. Amazon SNS is a notification service so it delivers notifications to subscribers. It does store data durably but is less suitable than SQS for this use case. Scaling on CPU usage is not the best solution as it does not relate to the number of jobs waiting to be processed.

INCORRECT: "Create an Amazon SNS topic to send the jobs that need to be processed. Create an Amazon EC2 Auto Scaling group for the compute application. Set the scaling policy for the Auto Scaling group to add and remove nodes based on the number of messages published to the SNS topic" is incorrect. Amazon SNS is a notification service so it delivers notifications to subscribers. It does store data durably but is less suitable than SQS for this use case. Scaling on the number of notifications in SNS is not possible.

References:



11. QUESTION

A company runs an application on six web application servers in an Amazon EC2 Auto Scaling group in a single Availability Zone. The application is fronted by an Application Load Balancer (ALB). A Solutions Architect needs to modify the infrastructure to be highly available without making any modifications to the application.

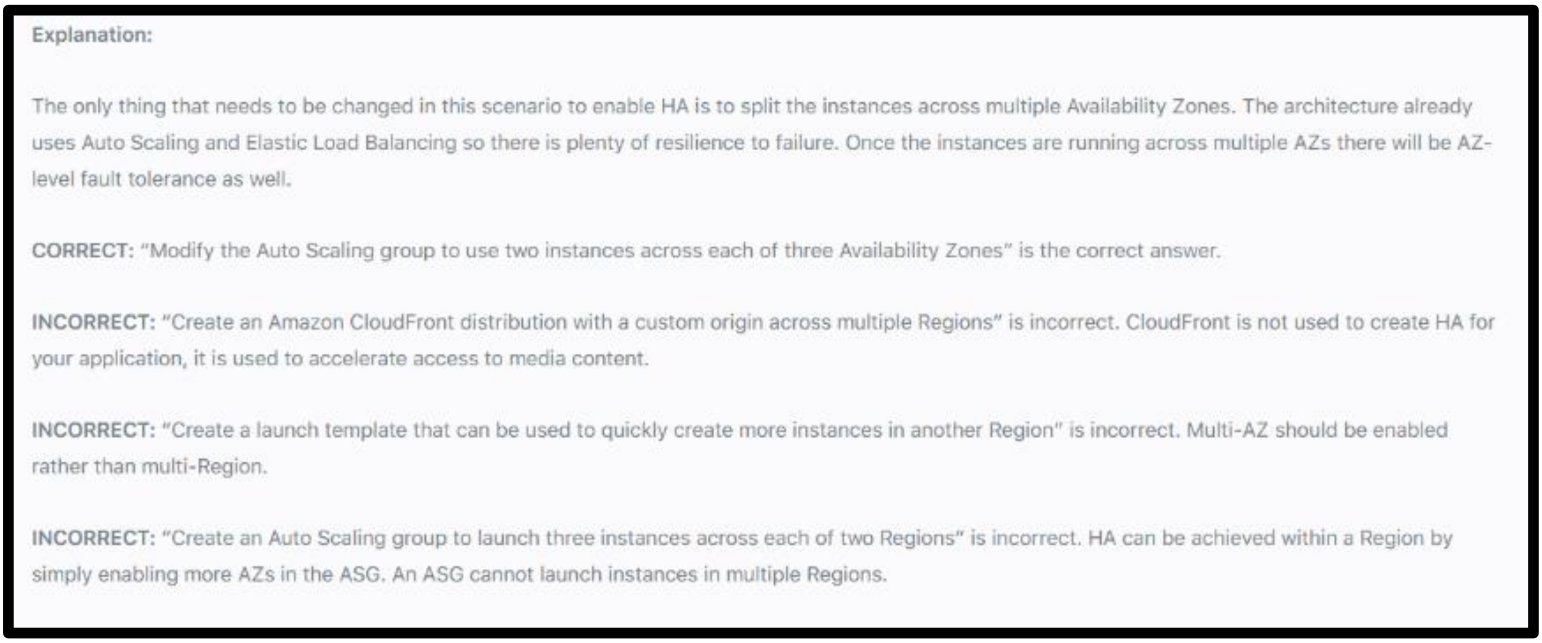
Which architecture should the Solutions Architect choose to enable high availability?

○ Create an Amazon CloudFront distribution with a custom origin across multiple Regions.

● Modify the Auto Scaling group to use two instances across each of three Availability Zones.

○ Create a launch template that can be used to quickly create more instances in another Region.

○ Create an Auto Scaling group to launch three instances across each of two Regions.



Explanation:

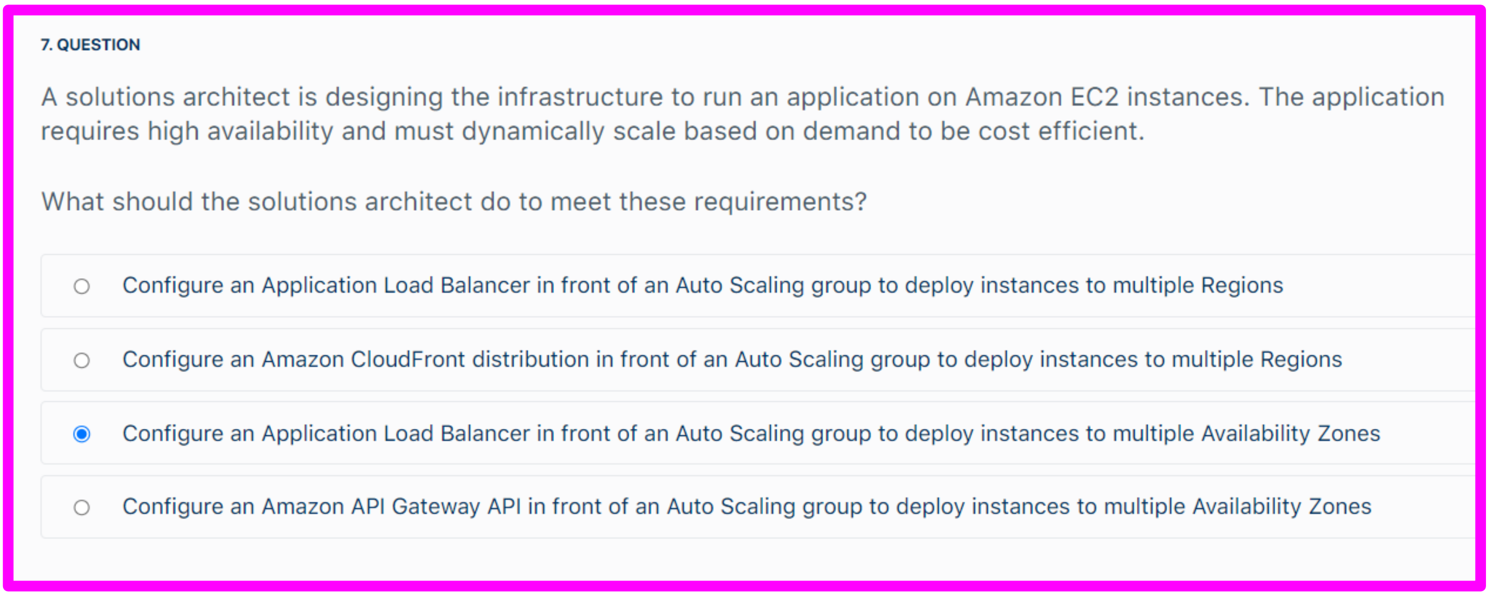
The only thing that needs to be changed in this scenario to enable HA is to split the instances across multiple Availability Zones. The architecture already uses Auto Scaling and Elastic Load Balancing so there is plenty of resilience to failure. Once the instances are running across multiple AZs there will be AZ-level fault tolerance as well.

CORRECT: "Modify the Auto Scaling group to use two instances across each of three Availability Zones" is the correct answer.

INCORRECT: "Create an Amazon CloudFront distribution with a custom origin across multiple Regions" is incorrect. CloudFront is not used to create HA for your application, it is used to accelerate access to media content.

INCORRECT: "Create a launch template that can be used to quickly create more instances in another Region" is incorrect. Multi-AZ should be enabled rather than multi-Region.

INCORRECT: "Create an Auto Scaling group to launch three instances across each of two Regions" is incorrect. HA can be achieved within a Region by simply enabling more AZs in the ASG. An ASG cannot launch instances in multiple Regions.



7. QUESTION

A solutions architect is designing the infrastructure to run an application on Amazon EC2 instances. The application requires high availability and must dynamically scale based on demand to be cost efficient.

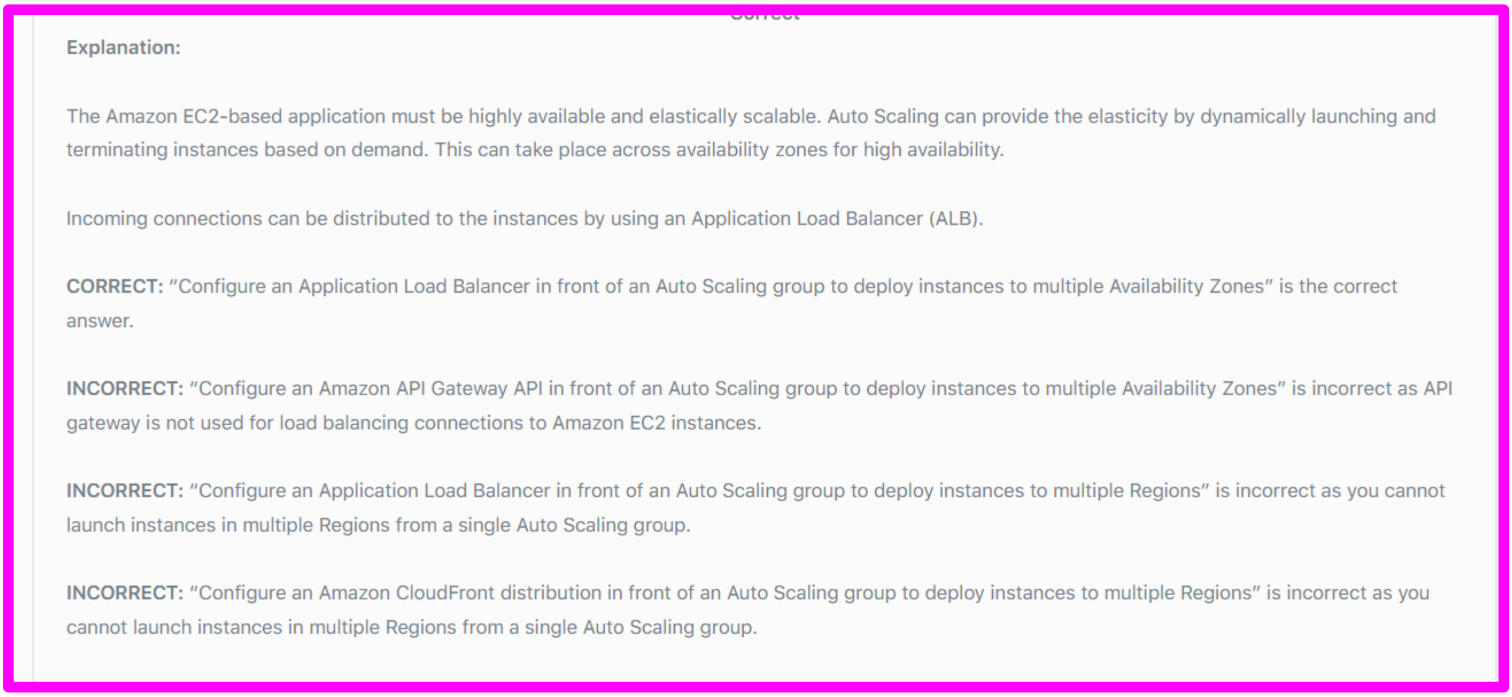
What should the solutions architect do to meet these requirements?

○ Configure an Application Load Balancer in front of an Auto Scaling group to deploy instances to multiple Regions

○ Configure an Amazon CloudFront distribution in front of an Auto Scaling group to deploy instances to multiple Regions

● Configure an Application Load Balancer in front of an Auto Scaling group to deploy instances to multiple Availability Zones

○ Configure an Amazon API Gateway API in front of an Auto Scaling group to deploy instances to multiple Availability Zones



Correct

Explanation:

The Amazon EC2-based application must be highly available and elastically scalable. Auto Scaling can provide the elasticity by dynamically launching and terminating instances based on demand. This can take place across availability zones for high availability.

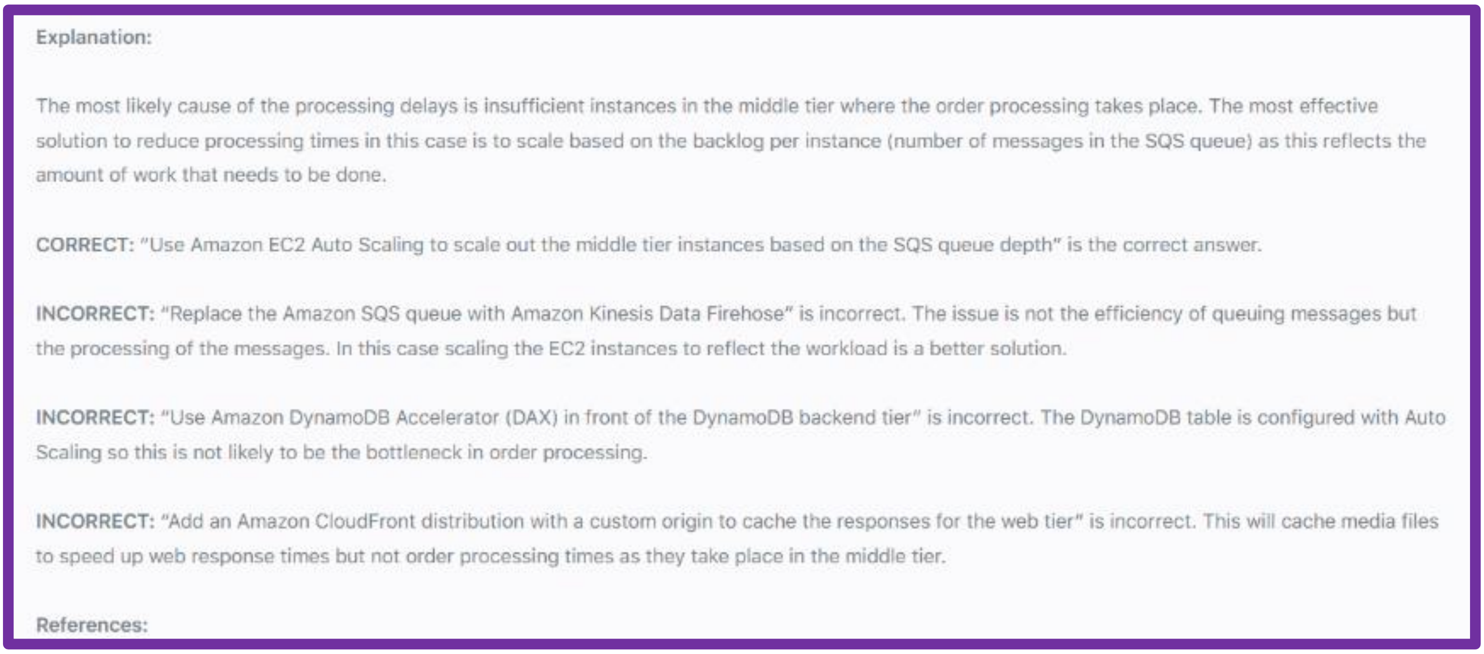
Incoming connections can be distributed to the instances by using an Application Load Balancer (ALB).

CORRECT: "Configure an Application Load Balancer in front of an Auto Scaling group to deploy instances to multiple Availability Zones" is the correct answer.

INCORRECT: "Configure an Amazon API Gateway API in front of an Auto Scaling group to deploy instances to multiple Availability Zones" is incorrect as API gateway is not used for load balancing connections to Amazon EC2 instances.

INCORRECT: "Configure an Application Load Balancer in front of an Auto Scaling group to deploy instances to multiple Regions" is incorrect as you cannot launch instances in multiple Regions from a single Auto Scaling group.

INCORRECT: "Configure an Amazon CloudFront distribution in front of an Auto Scaling group to deploy instances to multiple Regions" is incorrect as you cannot launch instances in multiple Regions from a single Auto Scaling group.



Explanation:

The most likely cause of the processing delays is insufficient instances in the middle tier where the order processing takes place. The most effective solution to reduce processing times in this case is to scale based on the backlog per instance (number of messages in the SQS queue) as this reflects the amount of work that needs to be done.

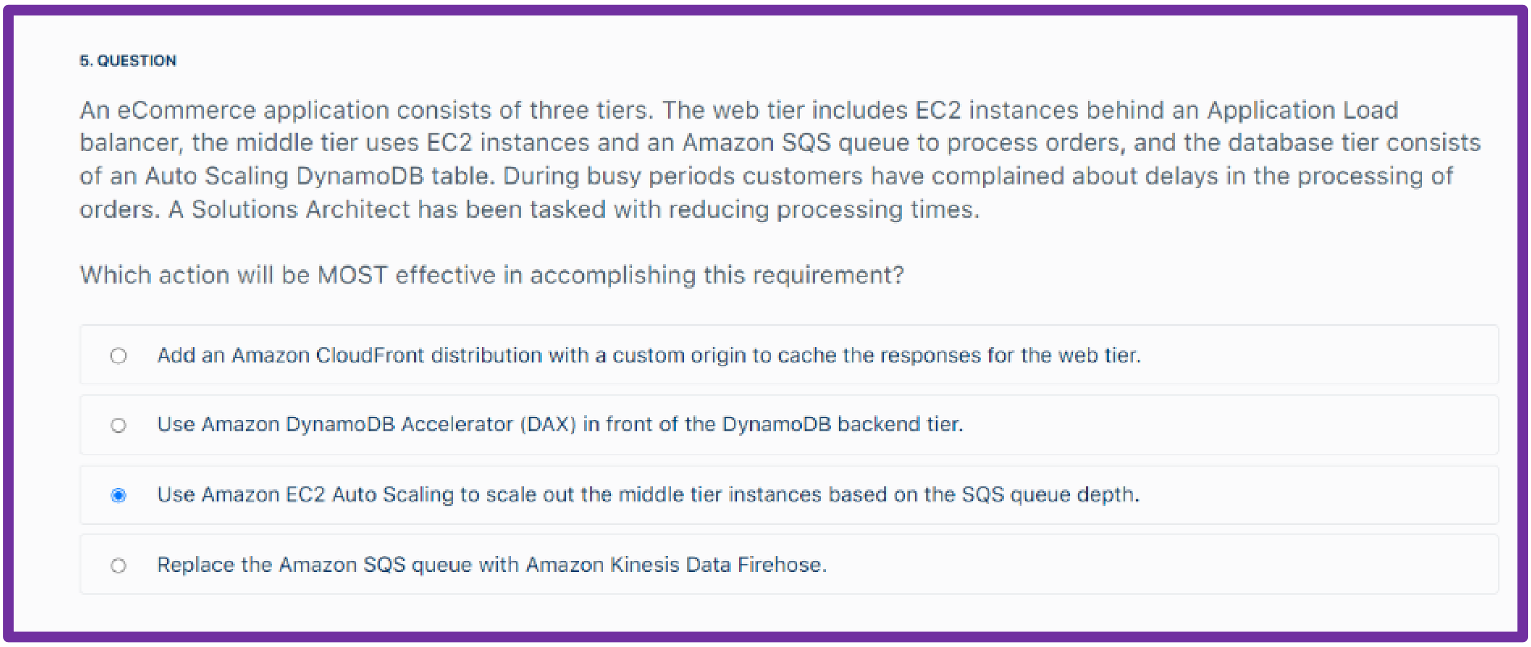
CORRECT: "Use Amazon EC2 Auto Scaling to scale out the middle tier instances based on the SQS queue depth" is the correct answer.

INCORRECT: "Replace the Amazon SQS queue with Amazon Kinesis Data Firehose" is incorrect. The issue is not the efficiency of queuing messages but the processing of the messages. In this case scaling the EC2 instances to reflect the workload is a better solution.

INCORRECT: "Use Amazon DynamoDB Accelerator (DAX) in front of the DynamoDB backend tier" is incorrect. The DynamoDB table is configured with Auto Scaling so this is not likely to be the bottleneck in order processing.

INCORRECT: "Add an Amazon CloudFront distribution with a custom origin to cache the responses for the web tier" is incorrect. This will cache media files to speed up web response times but not order processing times as they take place in the middle tier.

References:



5. QUESTION

An eCommerce application consists of three tiers. The web tier includes EC2 instances behind an Application Load balancer, the middle tier uses EC2 instances and an Amazon SQS queue to process orders, and the database tier consists of an Auto Scaling DynamoDB table. During busy periods customers have complained about delays in the processing of orders. A Solutions Architect has been tasked with reducing processing times.

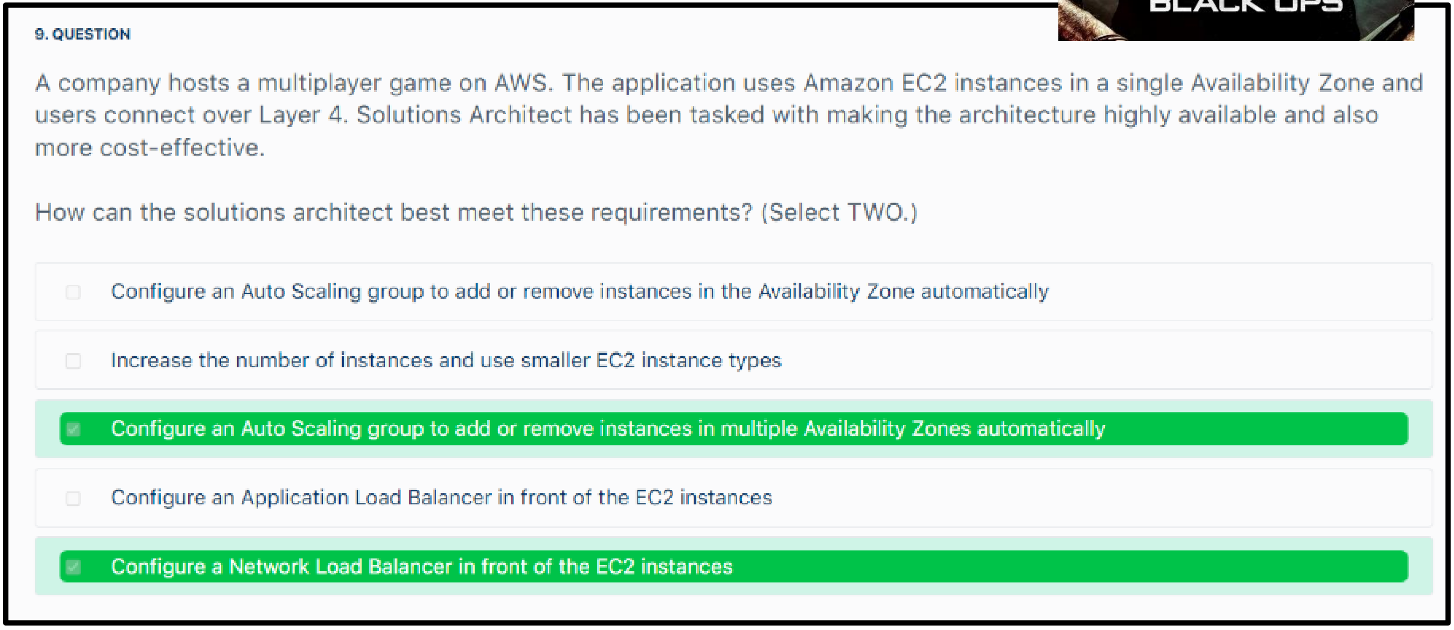
Which action will be MOST effective in accomplishing this requirement?

○ Add an Amazon CloudFront distribution with a custom origin to cache the responses for the web tier.

○ Use Amazon DynamoDB Accelerator (DAX) in front of the DynamoDB backend tier.

● Use Amazon EC2 Auto Scaling to scale out the middle tier instances based on the SQS queue depth.

○ Replace the Amazon SQS queue with Amazon Kinesis Data Firehose.



9. QUESTION

A company hosts a multiplayer game on AWS. The application uses Amazon EC2 instances in a single Availability Zone and users connect over Layer 4. Solutions Architect has been tasked with making the architecture highly available and also more cost-effective.

How can the solutions architect best meet these requirements? (Select TWO.)

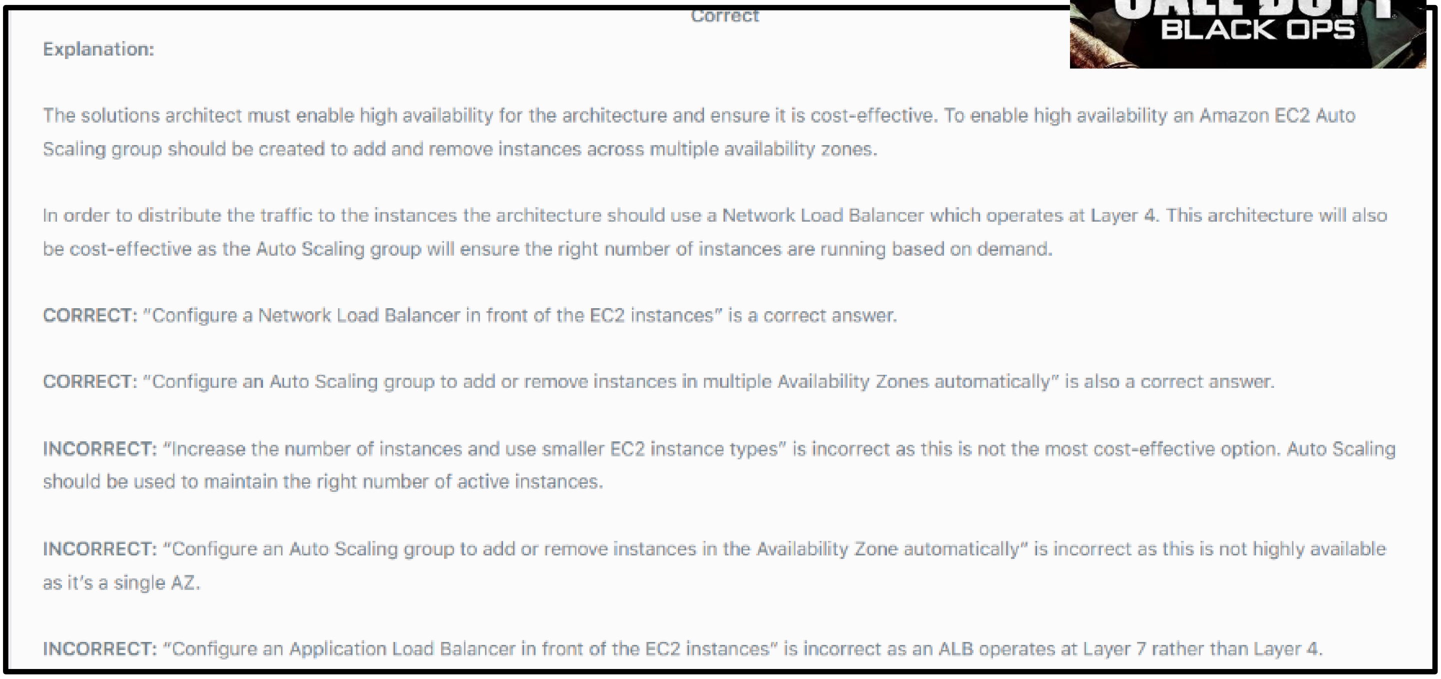
□ Configure an Auto Scaling group to add or remove instances in the Availability Zone automatically

□ Increase the number of instances and use smaller EC2 instance types

☑ Configure an Auto Scaling group to add or remove instances in multiple Availability Zones automatically

□ Configure an Application Load Balancer in front of the EC2 instances

☑ Configure a Network Load Balancer in front of the EC2 instances



Correct

Explanation:

The solutions architect must enable high availability for the architecture and ensure it is cost-effective. To enable high availability an Amazon EC2 Auto Scaling group should be created to add and remove instances across multiple availability zones.

In order to distribute the traffic to the instances the architecture should use a Network Load Balancer which operates at Layer 4. This architecture will also be cost-effective as the Auto Scaling group will ensure the right number of instances are running based on demand.

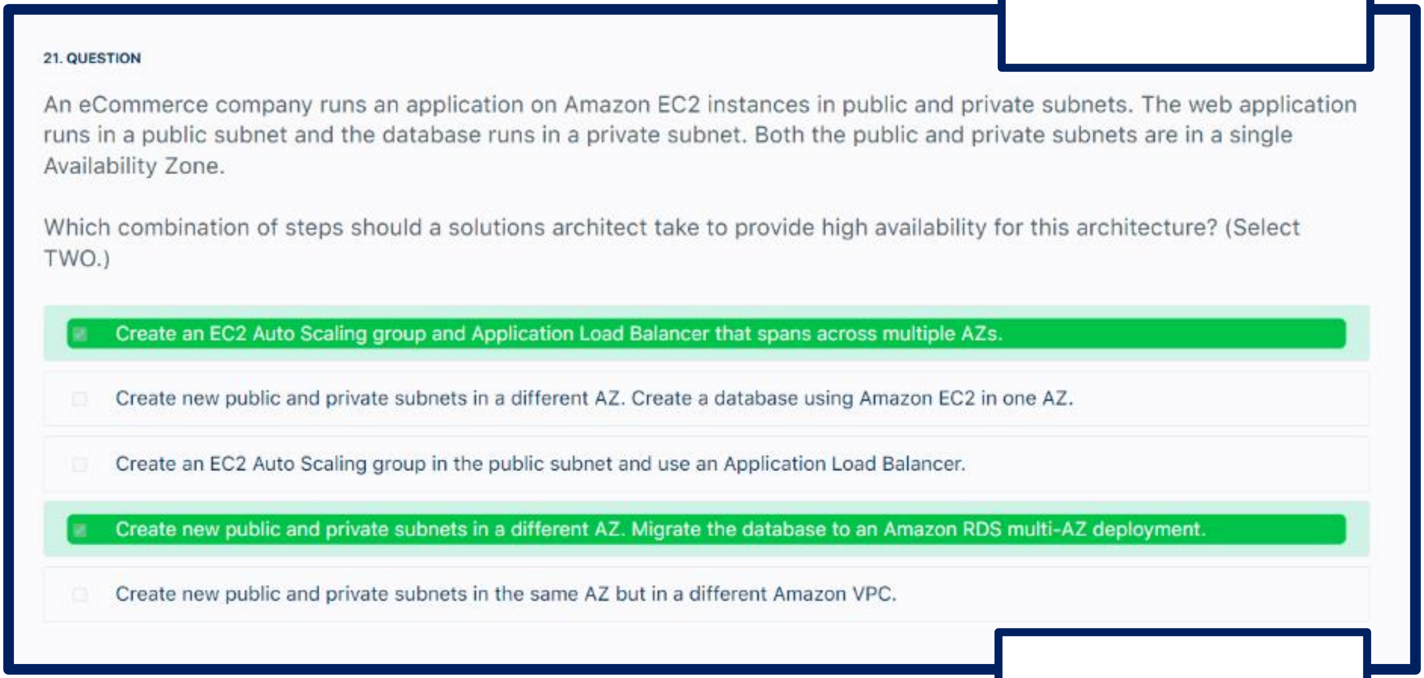
CORRECT: "Configure a Network Load Balancer in front of the EC2 instances" is a correct answer.

CORRECT: "Configure an Auto Scaling group to add or remove instances in multiple Availability Zones automatically" is also a correct answer.

INCORRECT: "Increase the number of instances and use smaller EC2 instance types" is incorrect as this is not the most cost-effective option. Auto Scaling should be used to maintain the right number of active instances.

INCORRECT: "Configure an Auto Scaling group to add or remove instances in the Availability Zone automatically" is incorrect as this is not highly available as it's a single AZ.

INCORRECT: "Configure an Application Load Balancer in front of the EC2 instances" is incorrect as an ALB operates at Layer 7 rather than Layer 4.

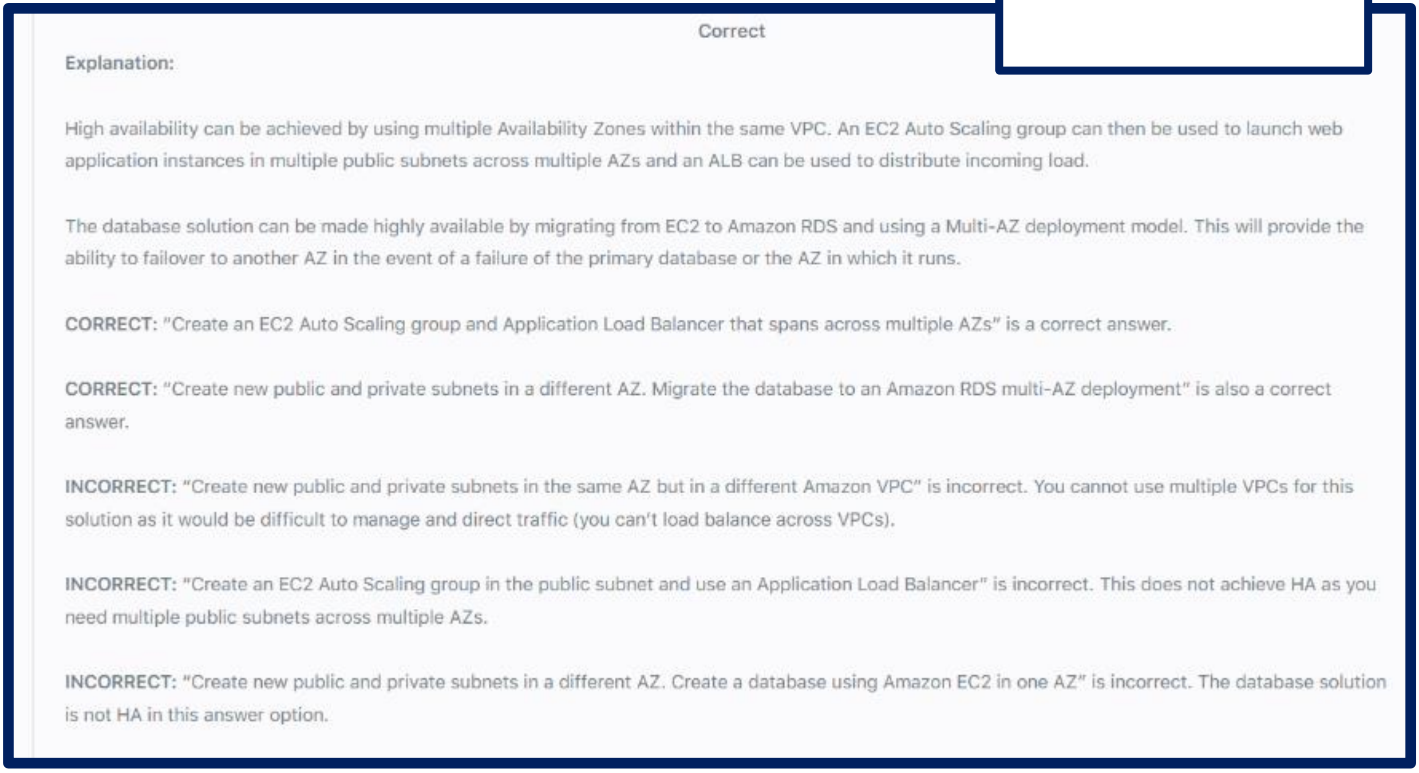


21. QUESTION

An eCommerce company runs an application on Amazon EC2 instances in public and private subnets. The web application runs in a public subnet and the database runs in a private subnet. Both the public and private subnets are in a single Availability Zone.

Which combination of steps should a solutions architect take to provide high availability for this architecture? (Select TWO.)

- ☑ Create an EC2 Auto Scaling group and Application Load Balancer that spans across multiple AZs.
- ☐ Create new public and private subnets in a different AZ. Create a database using Amazon EC2 in one AZ.
- ☐ Create an EC2 Auto Scaling group in the public subnet and use an Application Load Balancer.
- ☑ Create new public and private subnets in a different AZ. Migrate the database to an Amazon RDS multi-AZ deployment.
- ☐ Create new public and private subnets in the same AZ but in a different Amazon VPC.



Correct

Explanation:

High availability can be achieved by using multiple Availability Zones within the same VPC. An EC2 Auto Scaling group can then be used to launch web application instances in multiple public subnets across multiple AZs and an ALB can be used to distribute incoming load.

The database solution can be made highly available by migrating from EC2 to Amazon RDS and using a Multi-AZ deployment model. This will provide the ability to failover to another AZ in the event of a failure of the primary database or the AZ in which it runs.

CORRECT: "Create an EC2 Auto Scaling group and Application Load Balancer that spans across multiple AZs" is a correct answer.

CORRECT: "Create new public and private subnets in a different AZ. Migrate the database to an Amazon RDS multi-AZ deployment" is also a correct answer.

INCORRECT: "Create new public and private subnets in the same AZ but in a different Amazon VPC" is incorrect. You cannot use multiple VPCs for this solution as it would be difficult to manage and direct traffic (you can't load balance across VPCs).

INCORRECT: "Create an EC2 Auto Scaling group in the public subnet and use an Application Load Balancer" is incorrect. This does not achieve HA as you need multiple public subnets across multiple AZs.

INCORRECT: "Create new public and private subnets in a different AZ. Create a database using Amazon EC2 in one AZ" is incorrect. The database solution is not HA in this answer option.

TPN

11. ***Phenomenon1*** – the tendency of X to Y.
12. ***Phen2*** – the tendency of X to Y.
13. ***Phen3*** – the tendency of X to Y.
14. ***Phen4*** – the tendency of X to Y.
15. ***Phen5*** – the tendency of X to Y.
16. ***Phen6*** – the tendency of X to Y.
17. ***Phen7*** – the tendency of X to Y.
18. ***Phen8*** – the tendency of X to Y.
19. ***Phen9*** – the tendency of X to Y.
20. ***Phen10*** – the tendency of X to Y.

Glossary

Term1

Description of what term means here.

Term2

Description of what term means here.

Term3

Description of what term means here.

Bibliography

I. Official

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

II. Unofficial

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

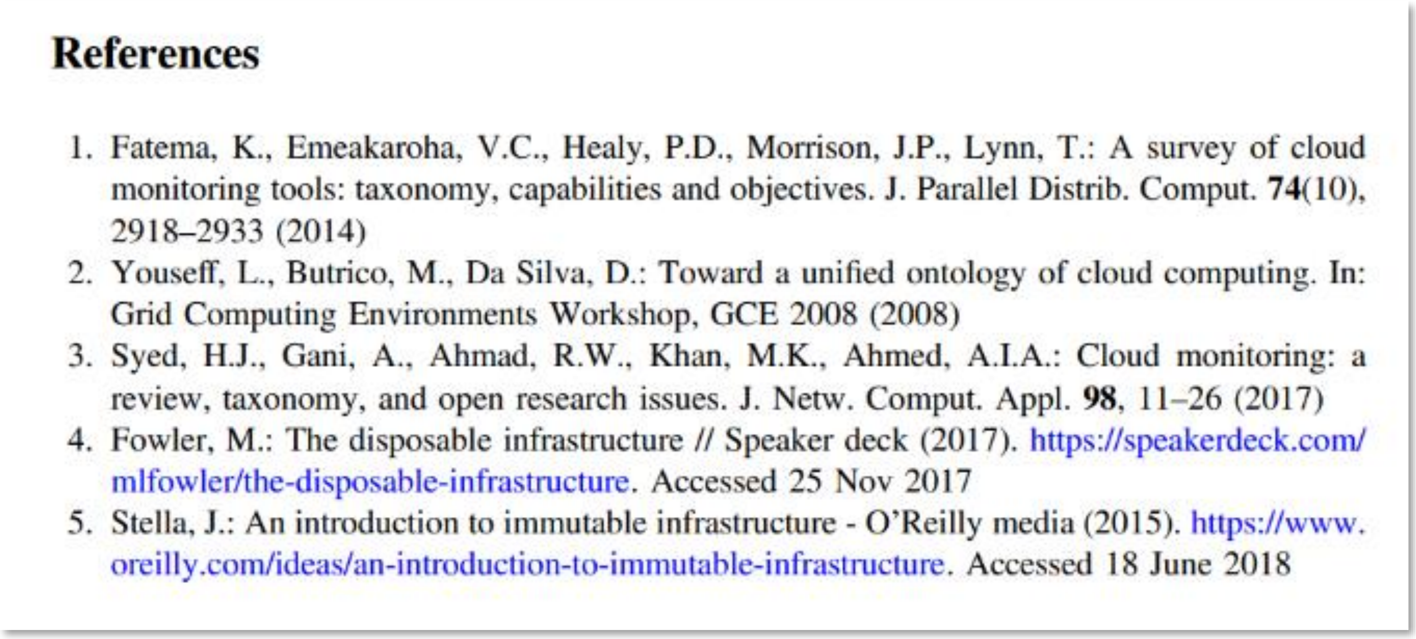
Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

IV. General

[Surname1]

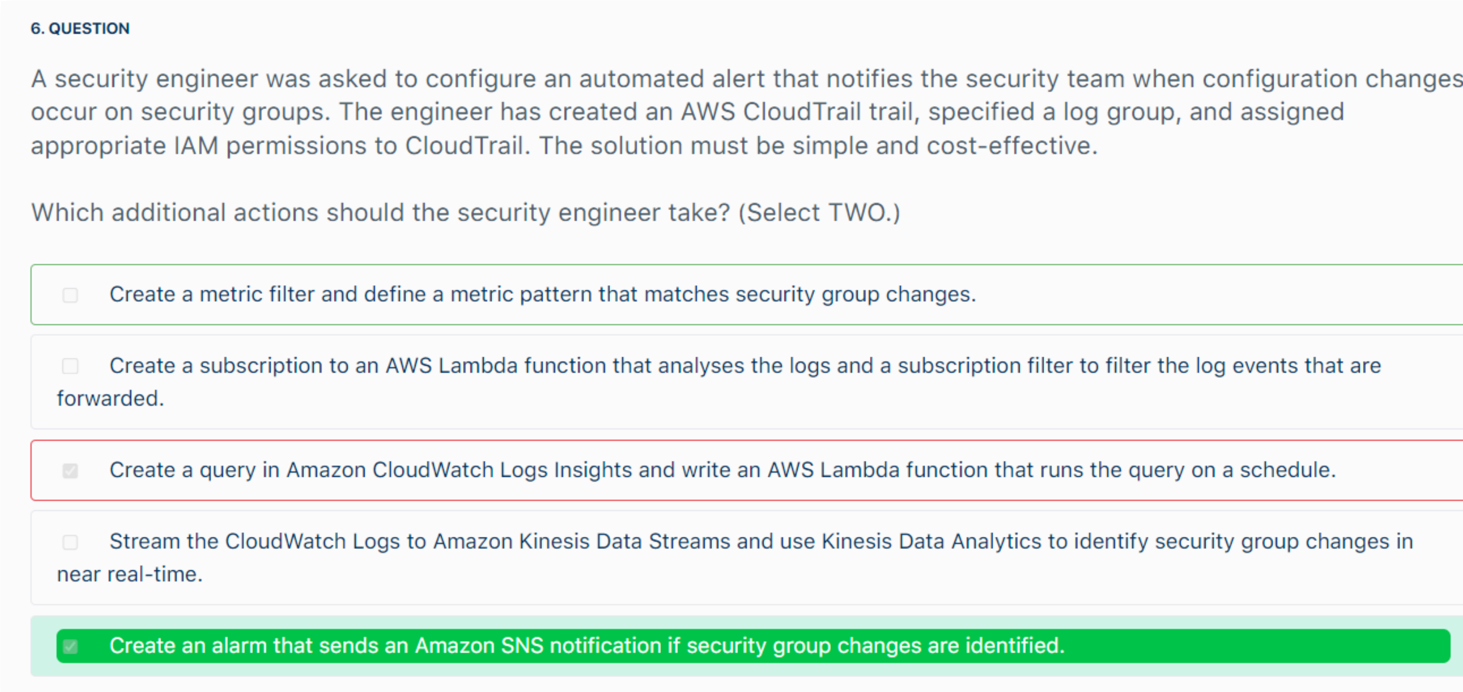
Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

CloudWatch



References

1. Fatema, K., Emeakaroha, V.C., Healy, P.D., Morrison, J.P., Lynn, T.: A survey of cloud monitoring tools: taxonomy, capabilities and objectives. J. Parallel Distrib. Comput. **74**(10), 2918–2933 (2014)
2. Youseff, L., Butrico, M., Da Silva, D.: Toward a unified ontology of cloud computing. In: Grid Computing Environments Workshop, GCE 2008 (2008)
3. Syed, H.J., Gani, A., Ahmad, R.W., Khan, M.K., Ahmed, A.I.A.: Cloud monitoring: a review, taxonomy, and open research issues. J. Netw. Comput. Appl. **98**, 11–26 (2017)
4. Fowler, M.: The disposable infrastructure // Speaker deck (2017). https://speakerdeck.com/mlfowler/the-disposable-infrastructure. Accessed 25 Nov 2017
5. Stella, J.: An introduction to immutable infrastructure - O'Reilly media (2015). https://www.oreilly.com/ideas/an-introduction-to-immutable-infrastructure. Accessed 18 June 2018



6. QUESTION

A security engineer was asked to configure an automated alert that notifies the security team when configuration changes occur on security groups. The engineer has created an AWS CloudTrail trail, specified a log group, and assigned appropriate IAM permissions to CloudTrail. The solution must be simple and cost-effective.

Which additional actions should the security engineer take? (Select TWO.)

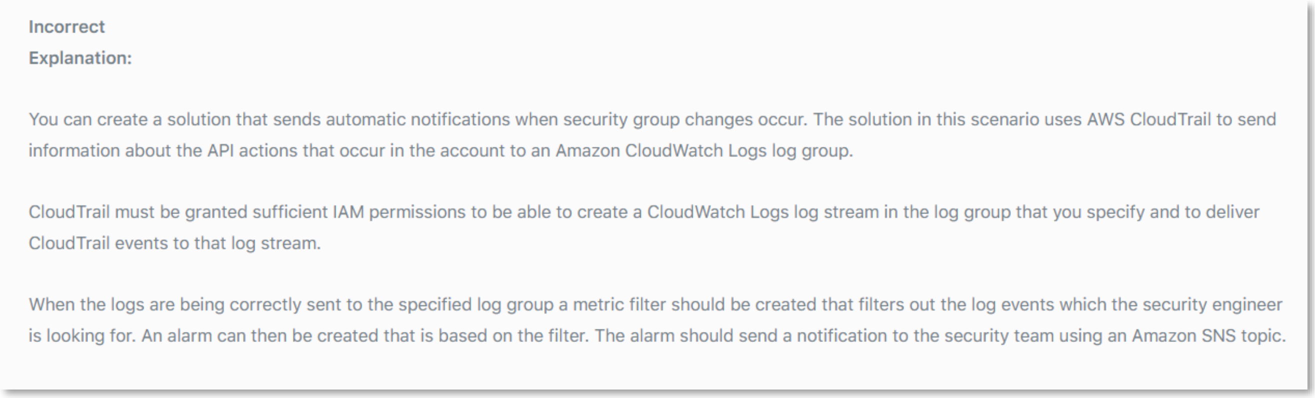
☐ Create a metric filter and define a metric pattern that matches security group changes.

☐ Create a subscription to an AWS Lambda function that analyses the logs and a subscription filter to filter the log events that are forwarded.

☑ Create a query in Amazon CloudWatch Logs Insights and write an AWS Lambda function that runs the query on a schedule.

☐ Stream the CloudWatch Logs to Amazon Kinesis Data Streams and use Kinesis Data Analytics to identify security group changes in near real-time.

☑ Create an alarm that sends an Amazon SNS notification if security group changes are identified.



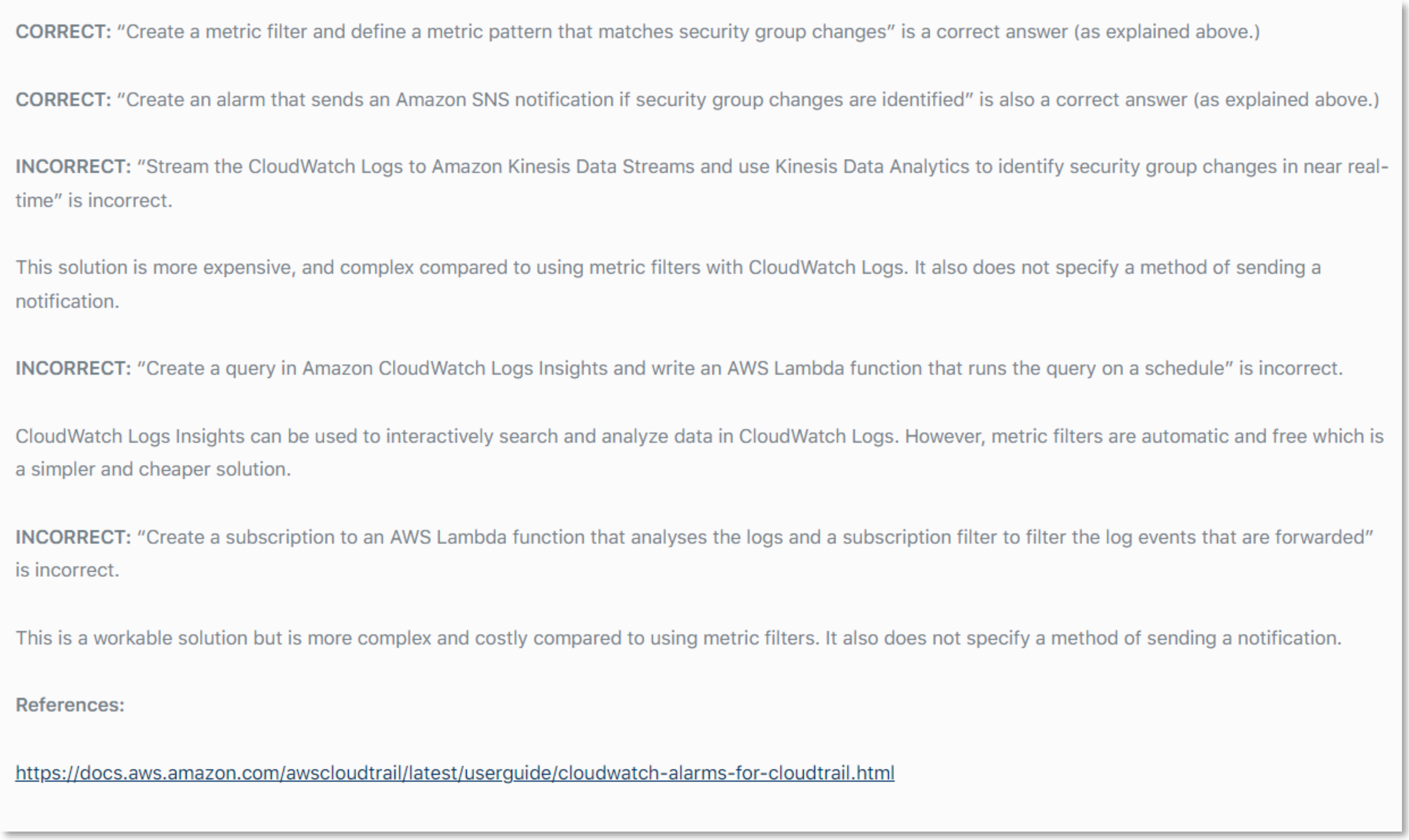
Incorrect

Explanation:

You can create a solution that sends automatic notifications when security group changes occur. The solution in this scenario uses AWS CloudTrail to send information about the API actions that occur in the account to an Amazon CloudWatch Logs log group.

CloudTrail must be granted sufficient IAM permissions to be able to create a CloudWatch Logs log stream in the log group that you specify and to deliver CloudTrail events to that log stream.

When the logs are being correctly sent to the specified log group a metric filter should be created that filters out the log events which the security engineer is looking for. An alarm can then be created that is based on the filter. The alarm should send a notification to the security team using an Amazon SNS topic.



CORRECT: "Create a metric filter and define a metric pattern that matches security group changes" is a correct answer (as explained above.)

CORRECT: "Create an alarm that sends an Amazon SNS notification if security group changes are identified" is also a correct answer (as explained above.)

INCORRECT: "Stream the CloudWatch Logs to Amazon Kinesis Data Streams and use Kinesis Data Analytics to identify security group changes in near real-time" is incorrect.

This solution is more expensive, and complex compared to using metric filters with CloudWatch Logs. It also does not specify a method of sending a notification.

INCORRECT: "Create a query in Amazon CloudWatch Logs Insights and write an AWS Lambda function that runs the query on a schedule" is incorrect.

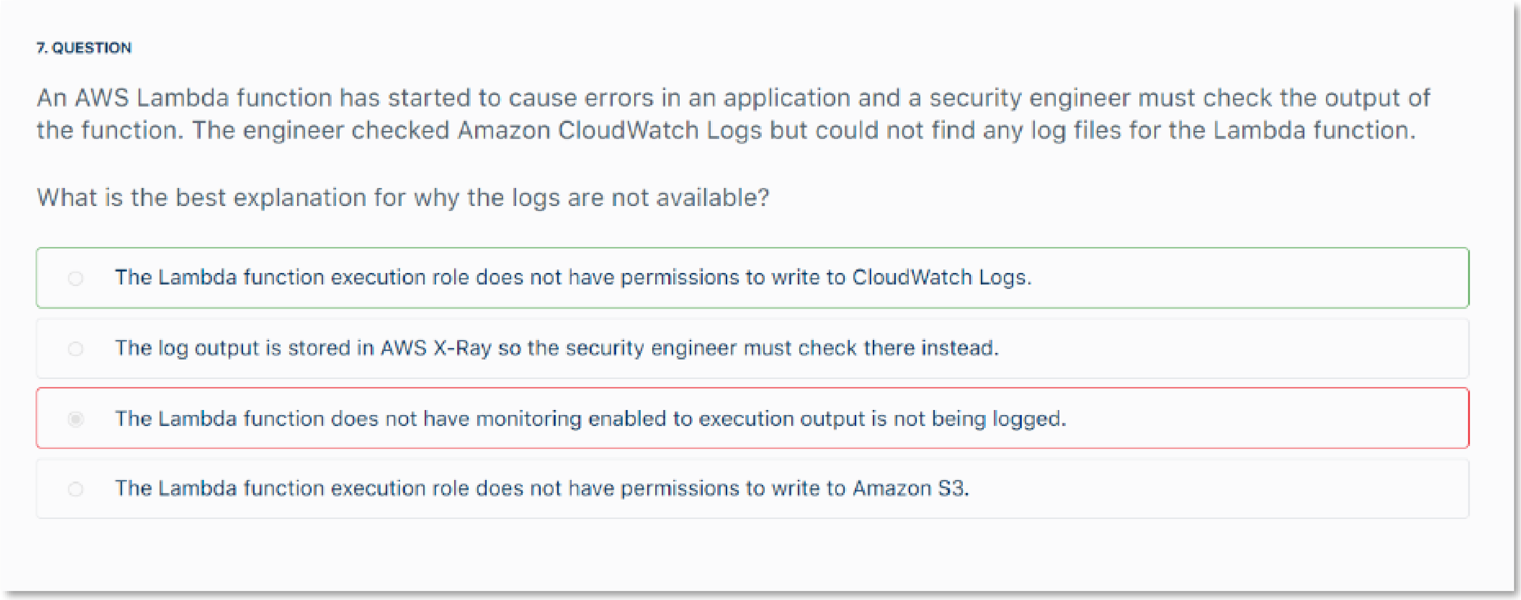
CloudWatch Logs Insights can be used to interactively search and analyze data in CloudWatch Logs. However, metric filters are automatic and free which is a simpler and cheaper solution.

INCORRECT: "Create a subscription to an AWS Lambda function that analyses the logs and a subscription filter to filter the log events that are forwarded" is incorrect.

This is a workable solution but is more complex and costly compared to using metric filters. It also does not specify a method of sending a notification.

References:

https://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudwatch-alarms-for-cloudtrail.html



7. QUESTION

An AWS Lambda function has started to cause errors in an application and a security engineer must check the output of the function. The engineer checked Amazon CloudWatch Logs but could not find any log files for the Lambda function.

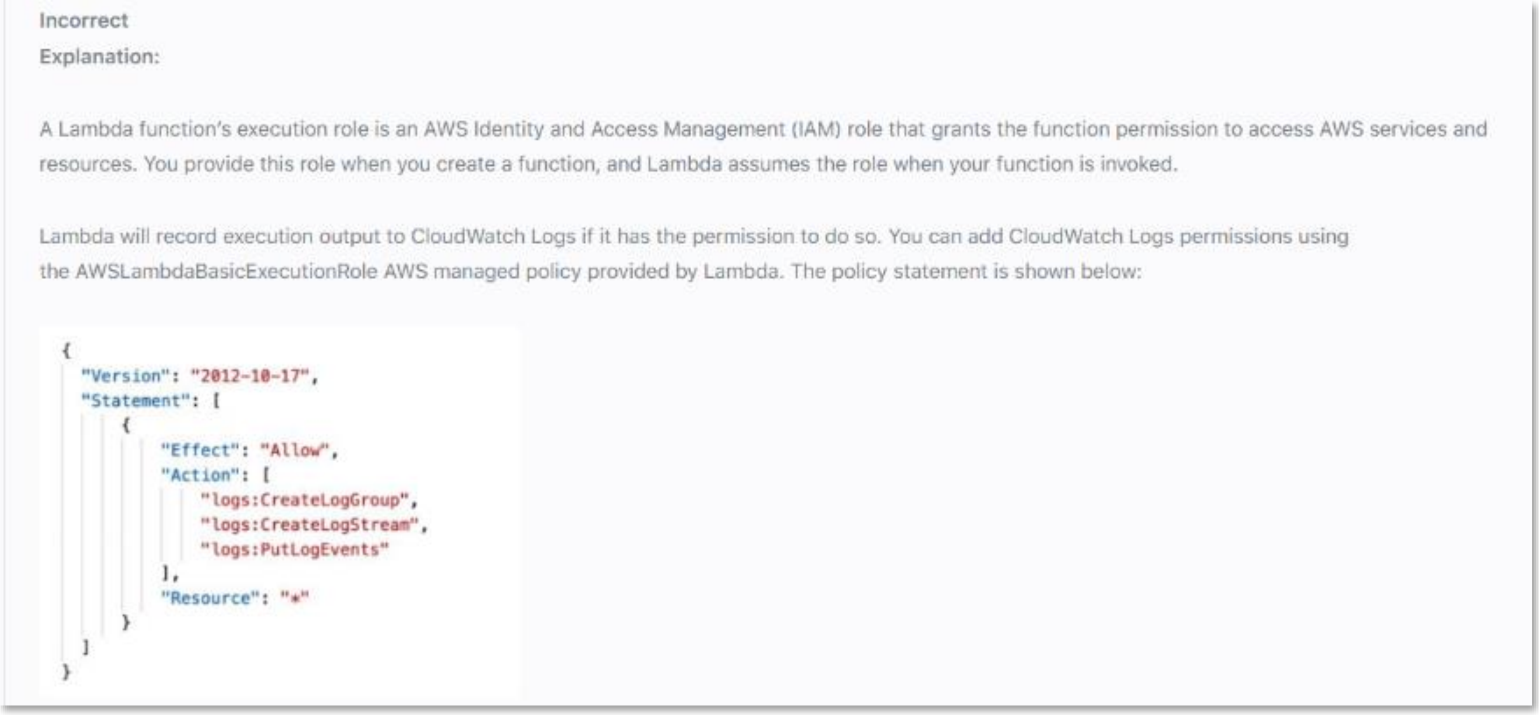
What is the best explanation for why the logs are not available?

○ The Lambda function execution role does not have permissions to write to CloudWatch Logs.

○ The log output is stored in AWS X-Ray so the security engineer must check there instead.

● The Lambda function does not have monitoring enabled to execution output is not being logged.

○ The Lambda function execution role does not have permissions to write to Amazon S3.



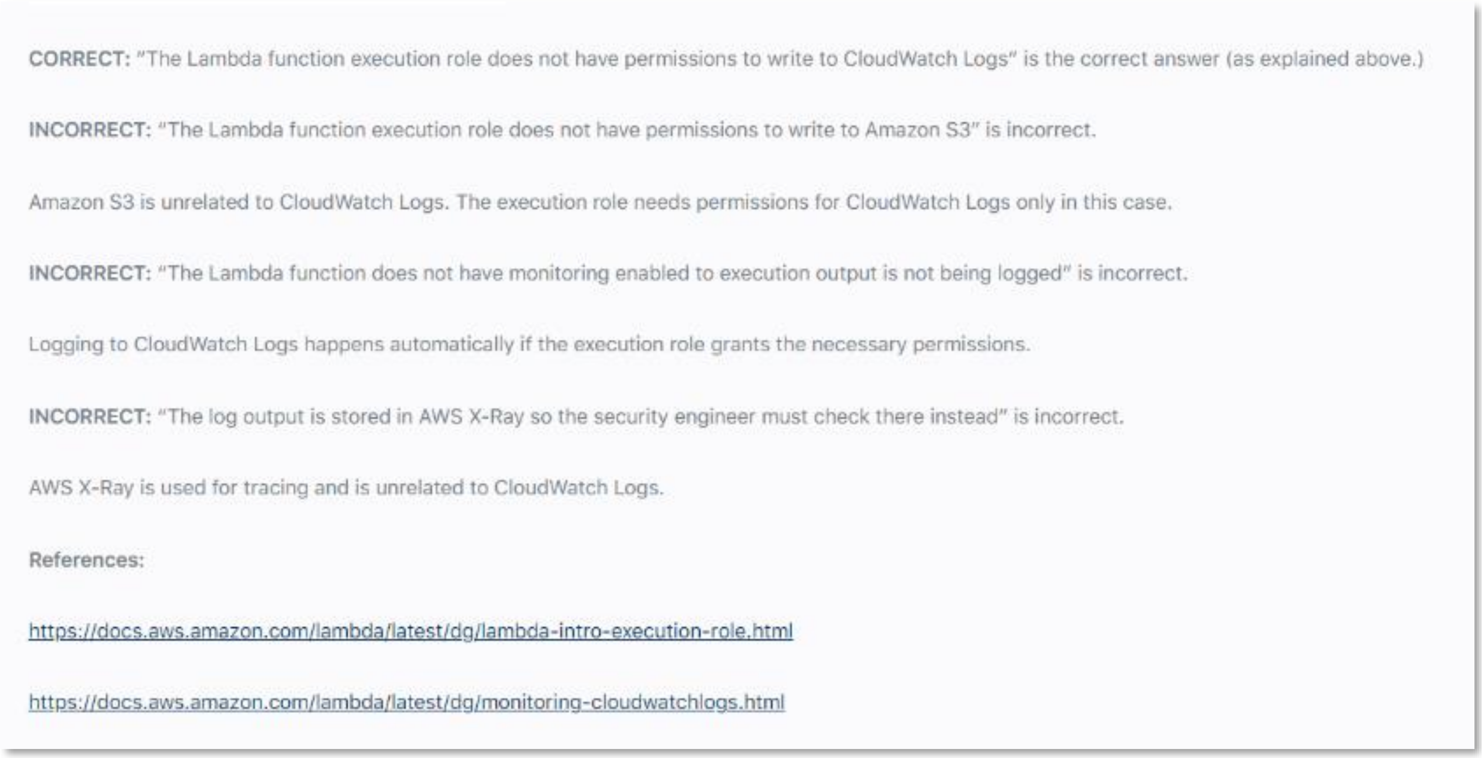
Incorrect

Explanation:

A Lambda function's execution role is an AWS Identity and Access Management (IAM) role that grants the function permission to access AWS services and resources. You provide this role when you create a function, and Lambda assumes the role when your function is invoked.

Lambda will record execution output to CloudWatch Logs if it has the permission to do so. You can add CloudWatch Logs permissions using the AWSLambdaBasicExecutionRole AWS managed policy provided by Lambda. The policy statement is shown below:

```
{
  "Version": "2012-10-17",
  "Statement": [
      {
          "Effect": "Allow",
          "Action": [
              "logs:CreateLogGroup",
              "logs:CreateLogStream",
              "logs:PutLogEvents"
          ],
          "Resource": "*"
      }
  ]
}
```



CORRECT: "The Lambda function execution role does not have permissions to write to CloudWatch Logs" is the correct answer (as explained above.)

INCORRECT: "The Lambda function execution role does not have permissions to write to Amazon S3" is incorrect.

Amazon S3 is unrelated to CloudWatch Logs. The execution role needs permissions for CloudWatch Logs only in this case.

INCORRECT: "The Lambda function does not have monitoring enabled to execution output is not being logged" is incorrect.

Logging to CloudWatch Logs happens automatically if the execution role grants the necessary permissions.

INCORRECT: "The log output is stored in AWS X-Ray so the security engineer must check there instead" is incorrect.

AWS X-Ray is used for tracing and is unrelated to CloudWatch Logs.

References:

https://docs.aws.amazon.com/lambda/latest/dg/lambda-intro-execution-role.html

https://docs.aws.amazon.com/lambda/latest/dg/monitoring-cloudwatchlogs.html

Amazon CloudWatch now supports dashboard variables

Posted On: Jun 30, 2023

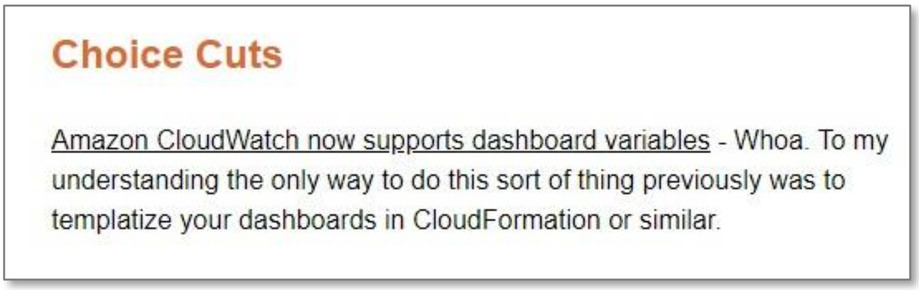
We are excited to announce Amazon CloudWatch dashboard variables, a new experience that makes it easier for you to quickly navigate between different operational views by using variables as navigation to view different data sets for the same resource.

With dashboard variables, you can create drill-down filters that enable you to build multiple views within a single dashboard. You can create a single dashboard with multiple custom variables, where you can switch between data sets, and more efficiently reuse and manage a more efficient fleet of dashboards.

Custom labels become navigation on your dashboard that can be used to toggle between data sets in the same view. You can now define set of labels such as 'application environment', 'region' or a 'customer id' and switch between those data sets based on the selection of label by using dropdown selectors, radio buttons or input boxes.

The new dashboard variable experience is now available in all AWS commercial regions at no additional cost and you can start using it immediately. You can use a step by step guide in the CloudWatch custom dashboard user interface to configure and manage your variables, or directly from a CloudWatch custom dashboard JSON, by adding an array of "variables" to the dashboard definition.

To learn more about creating dashboard variables, please refer to our documentation.



Choice Cuts

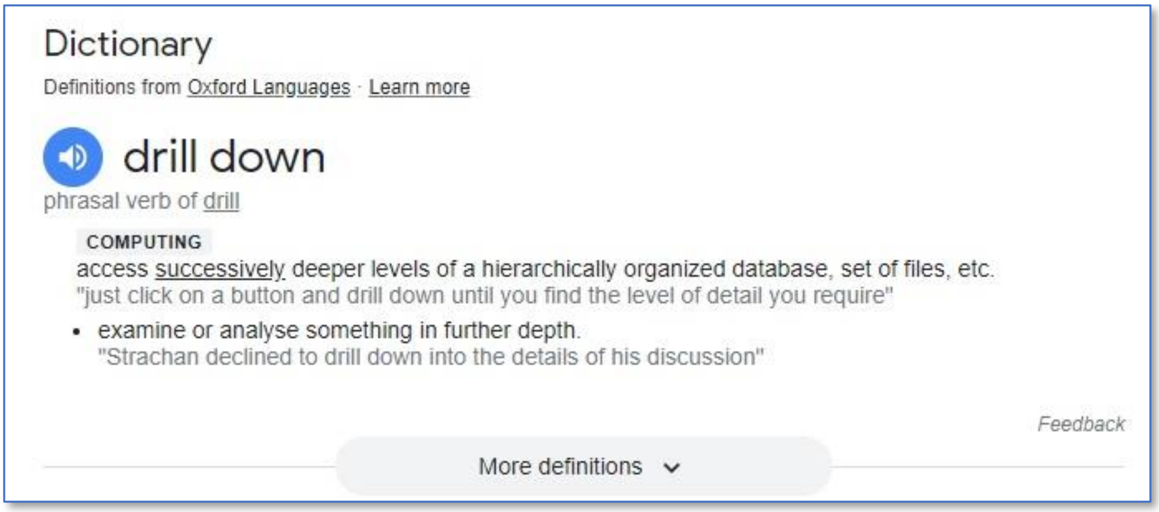
Amazon CloudWatch now supports dashboard variables - Whoa. To my understanding the only way to do this sort of thing previously was to templatize your dashboards in CloudFormation or similar.

Email from Corey Quinn on 10th July 2023

What does this mean?

Well, we can now "build multiple views within a single dashboard". It is not clear, however, what this means. It would be better if a concrete example was provided in this press release.

What is meant by "drill-down filters"? I find this term to be confusing. I have no idea what a drill-down filter is.



Dictionary

Definitions from Oxford Languages · Learn more

drill down

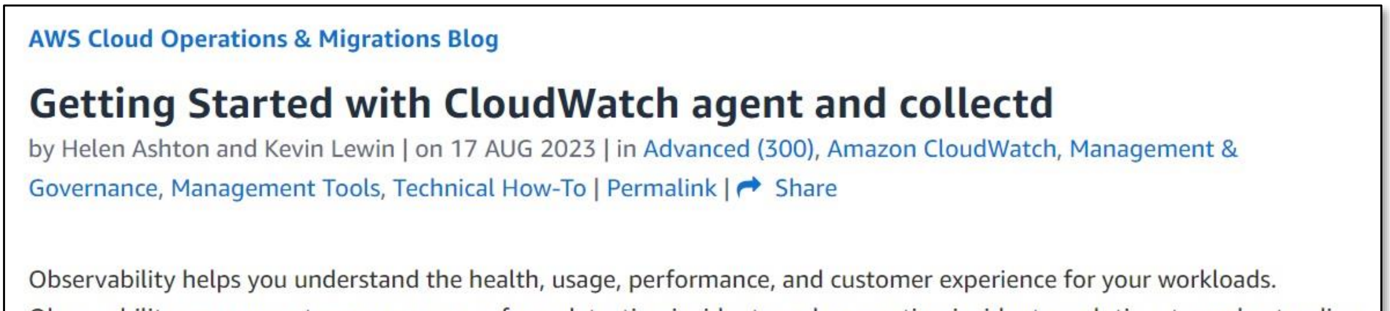
phrasal verb of drill

COMPUTING

access successively deeper levels of a hierarchically organized database, set of files, etc.
"just click on a button and drill down until you find the level of detail you require"

- examine or analyse something in further depth.
  "Strachan declined to drill down into the details of his discussion"

I understand what it is to drill down. But how can we have a drill down *filter*? A filter is just some property. We consider results and see whether each result has the property. There are no hierarchical levels involved. So—what on earth is a drill down filter?

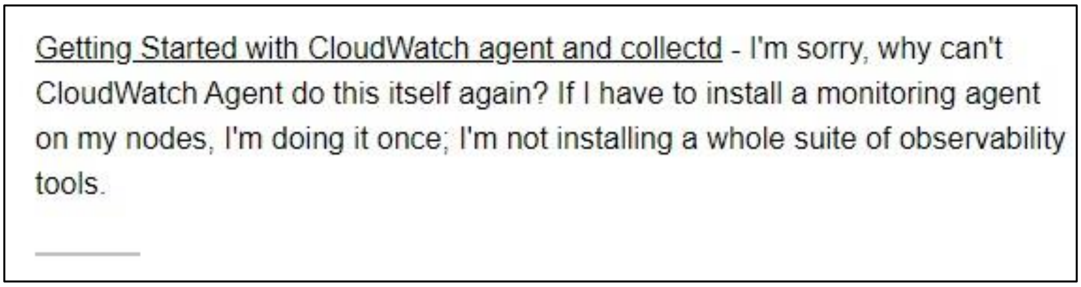


AWS Cloud Operations & Migrations Blog

Getting Started with CloudWatch agent and collectd

by Helen Ashton and Kevin Lewin | on 17 AUG 2023 | in Advanced (300), Amazon CloudWatch, Management & Governance, Management Tools, Technical How-To | Permalink | Share

Observability helps you understand the health, usage, performance, and customer experience for your workloads.



Getting Started with CloudWatch agent and collectd - I'm sorry, why can't CloudWatch Agent do this itself again? If I have to install a monitoring agent on my nodes, I'm doing it once; I'm not installing a whole suite of observability tools.

Email from Corey Quinn on 21st August 2023

Announcing Live Tail in Amazon CloudWatch Logs, providing real-time exploration of logs

Posted On: Jun 6, 2023

We are excited to announce Amazon CloudWatch Logs Live Tail, a new interactive log analytics experience feature that helps you detect and debug anomalies in applications. You can now view your logs interactively in real-time as they're ingested, which helps you to analyze and resolve issues across your systems and applications.

Live Tail provides customers a rich out-of-the-box experience to view and detect issues in their incoming logs. Additionally, it provides fine-grained controls to filter, highlight attributes of interest, and pause/replay logs while troubleshooting issues. Using Live Tail, DevOps teams can quickly validate if a process has correctly started, impact of configuration changes, or if a new deployment has gone smoothly.

Amazon CloudWatch Logs Live Tail is available in all AWS Commercial regions.

Start exploring and analyzing your logs in real-time using Amazon CloudWatch Logs Live Tail. To learn more, visit Amazon CloudWatch features or read Amazon CloudWatch Logs Live Tail Documentation. For pricing details, check Amazon CloudWatch Pricing - view and analyze your logs using CloudWatch Logs Live

Tail for an example of pricing. Check out the the AWS Cloud Operations blog to discover more about Live Tail.

Bibliography

I. Official
II. Unofficial
III. Critical
IV. General

I. Official

Amazon CloudWatch now supports dashboard variables. [Announcement]. Available at: https://aws.amazon.com/about-aws/whats-new/2023/06/amazon-cloudwatch-dashboard-variables/?utm_source=substack&utm_medium=email

Elastic Load Balancing

LOAD BALANCER

So, we have two distinct services:

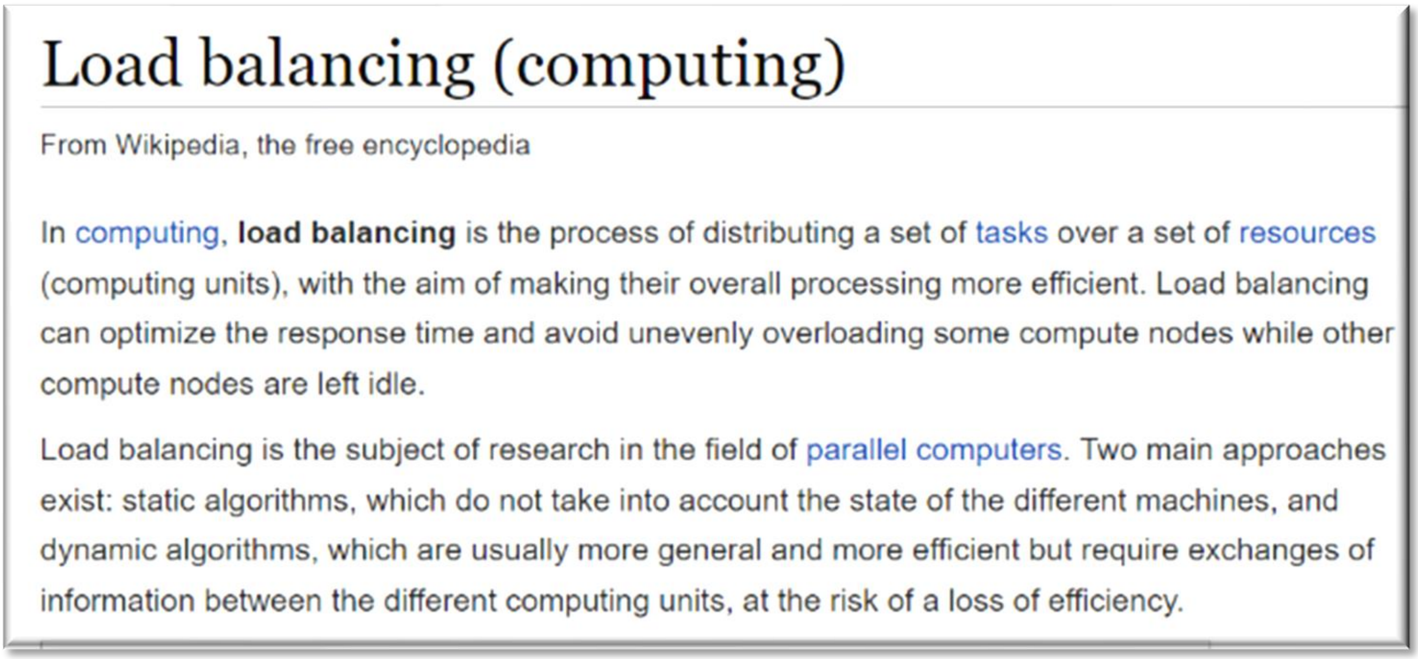
Auto Scaling and load balancing are distinct activities. I will associate Auto Scaling with pink and Load Balancing with blue to reinforce this. (The image of a hot air balloon and traffic light are somewhat arbitrary—do not read too much into them.) We often talk of there being an auto scaling group *behind* a load balancer.

Let me use a motorway analogy (a *highway*, for American readers). Auto Scaling involves altering capacity, by opening or closing lanes on the motorway. In the cloud, we add instances or terminate them.

Load balancing involves directing cars (traffic) to the particular lanes. Usually, we want to spread the traffic equally across the lanes. Either way, we work with what we've got (Auto Scaling deals with adding or removing lanes). Such distributing of traffic does not usually happen on motorways, so imagine that a police officer has to do this because there is an accident or something. He has to stand in a high-visibility jacket and gesture to the motorists, telling which lane to go into. In the cloud, requests from clients are distributed (balanced) across the instances.

We can also make an analogy with the checkouts in supermarkets. Sometimes, the tills become overwhelmed. Two things help the supermarket to make their service as available as possible. First, a speaker tells queuing customers which checkout they should go to. It says: "checkout number 2 please", when this becomes available. This is load balancing. The supermarket manager might also make the decision to open more tills. This is Auto Scaling. Since we are increasing the number of employees on the tills, we are scaling out. Giving energy drinks to the employees already present would be scaling *up* (increasing the capacity of the employees we already have).

What on earth is "load balancing"?

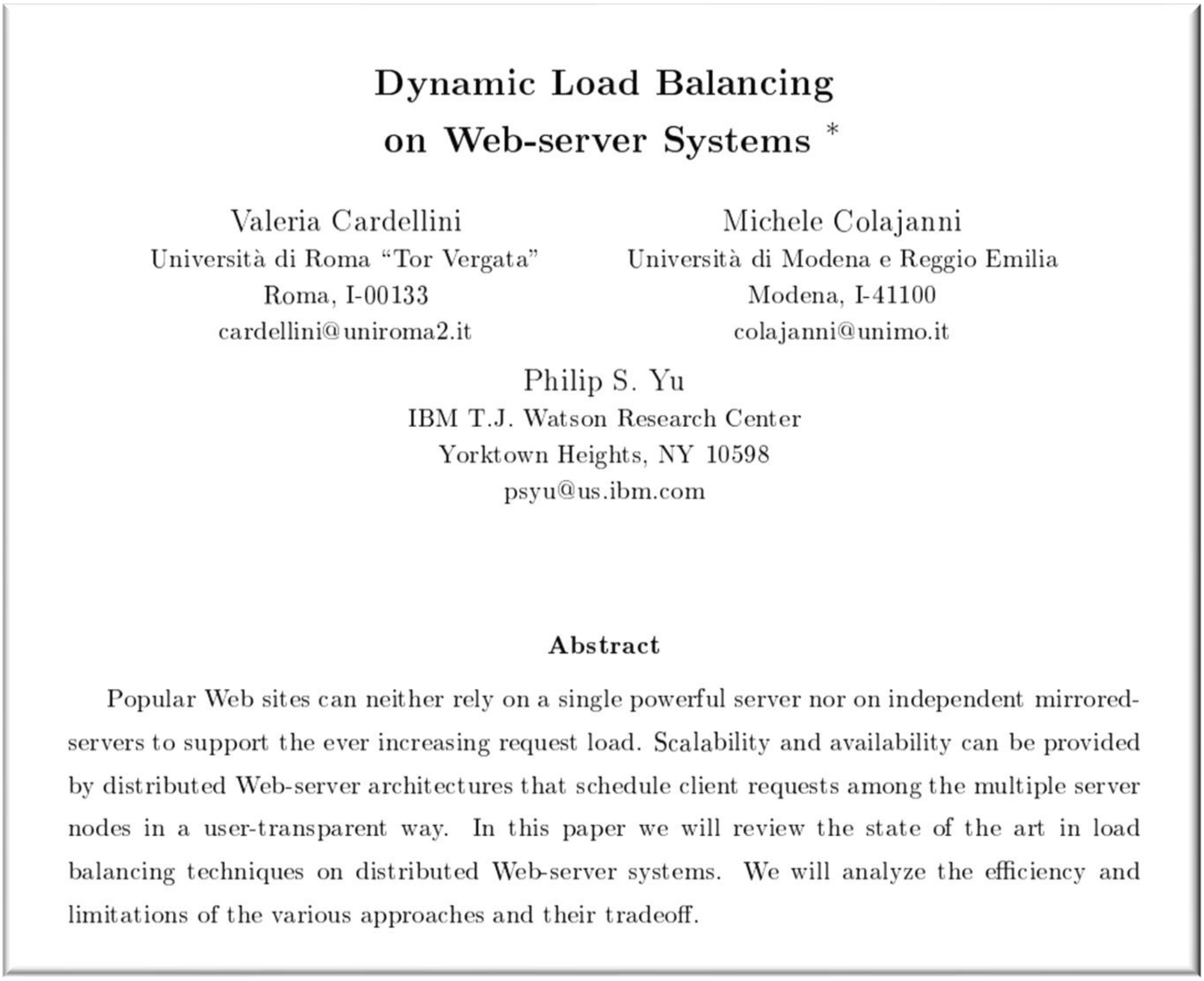


Load balancing (computing)

From Wikipedia, the free encyclopedia

In computing, **load balancing** is the process of distributing a set of tasks over a set of resources (computing units), with the aim of making their overall processing more efficient. Load balancing can optimize the response time and avoid unevenly overloading some compute nodes while other compute nodes are left idle.

Load balancing is the subject of research in the field of parallel computers. Two main approaches exist: static algorithms, which do not take into account the state of the different machines, and dynamic algorithms, which are usually more general and more efficient but require exchanges of information between the different computing units, at the risk of a loss of efficiency.



Dynamic Load Balancing on Web-server Systems *

Valeria Cardellini
Università di Roma "Tor Vergata"
Roma, I-00133
cardellini@uniroma2.it

Michele Colajanni
Università di Modena e Reggio Emilia
Modena, I-41100
colajanni@unimo.it

Philip S. Yu
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
psyu@us.ibm.com

Abstract

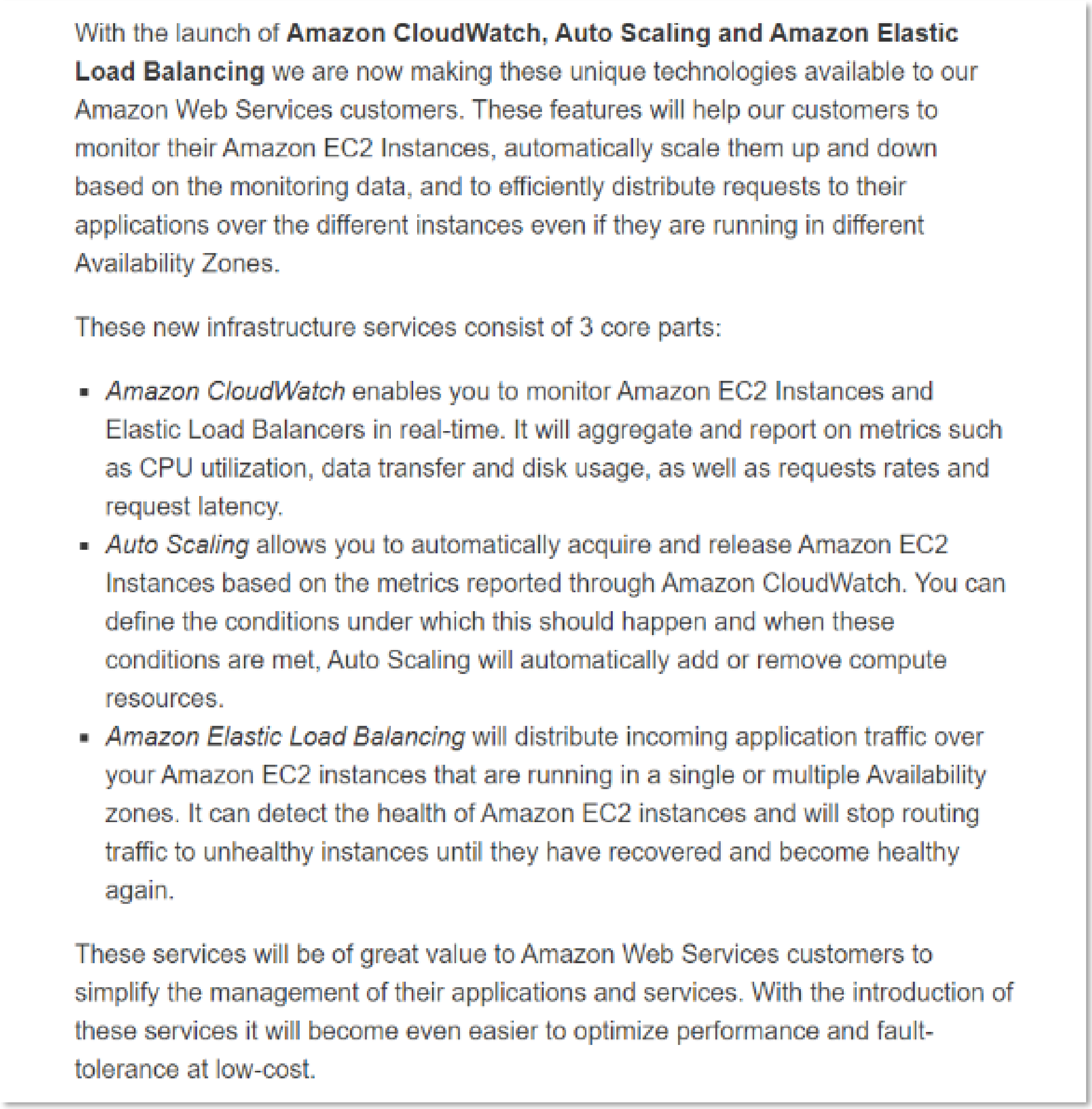
Popular Web sites can neither rely on a single powerful server nor on independent mirrored-servers to support the ever increasing request load. Scalability and availability can be provided by distributed Web-server architectures that schedule client requests among the multiple server nodes in a user-transparent way. In this paper we will review the state of the art in load balancing techniques on distributed Web-server systems. We will analyze the efficiency and limitations of the various approaches and their tradeoff.

There are now a number of different types of load balancer available on AWS. Note that the basic task—of balancing the load—is relatively easy to accomplish. Our doorman greets the incoming clients, and sends them on their way to a server. There is not really much more to say about this. We simply use a roundrobin algorithm to make sure each EC2 instance is given a fair share of the traffic. What's happened, however, is that AWS have realised the load balancer is a sort of "piggy in the middle". There are therefore a lot of functions which it can fulfil, as a mediator of traffic—or a sort of "Janus-headed go-between faculty". Most of our discussion is in fact going to be related to the *way* the load balancer distributes the traffic.

For instance, there's a question over whether the traffic is distributed *securely* or not. There's a question over whether the load balancer should be very sympathetic to the desires of client, with the many requests is has. Perhaps the client has a "favourite" server it likes to be served by. There's a question over whether the client should be sent to the same server each time.

Our load balancer can even make the lives of the servers behind it more easy or more difficult. If servers haven't quite finished initialising, the load balancer has a choice to allow them a grace period. It is sympathetic to the *server*, waiting for a period, before sending any traffic towards it. So, these humble "doormen" act with the utmost discretion, in ways configurable by you, such that their role in your deployment is *essential.*

AWS have a product, which provides load balancing. In exam questions, we will often talk as if there is a load balancer *in front of* a group of EC2 instances, or *in front of* an Auto Scaling group. The load balancer has the instances *behind* it, as if like a bouncer on a nightclub facing outwards onto the street. AWS call their service "Elastic Load Balancing" (ELB). I was announced in 2009, part of the "golden triad" of services. See this blog post from Werner Vogels:



With the launch of **Amazon CloudWatch, Auto Scaling and Amazon Elastic Load Balancing** we are now making these unique technologies available to our Amazon Web Services customers. These features will help our customers to monitor their Amazon EC2 Instances, automatically scale them up and down based on the monitoring data, and to efficiently distribute requests to their applications over the different instances even if they are running in different Availability Zones.

These new infrastructure services consist of 3 core parts:

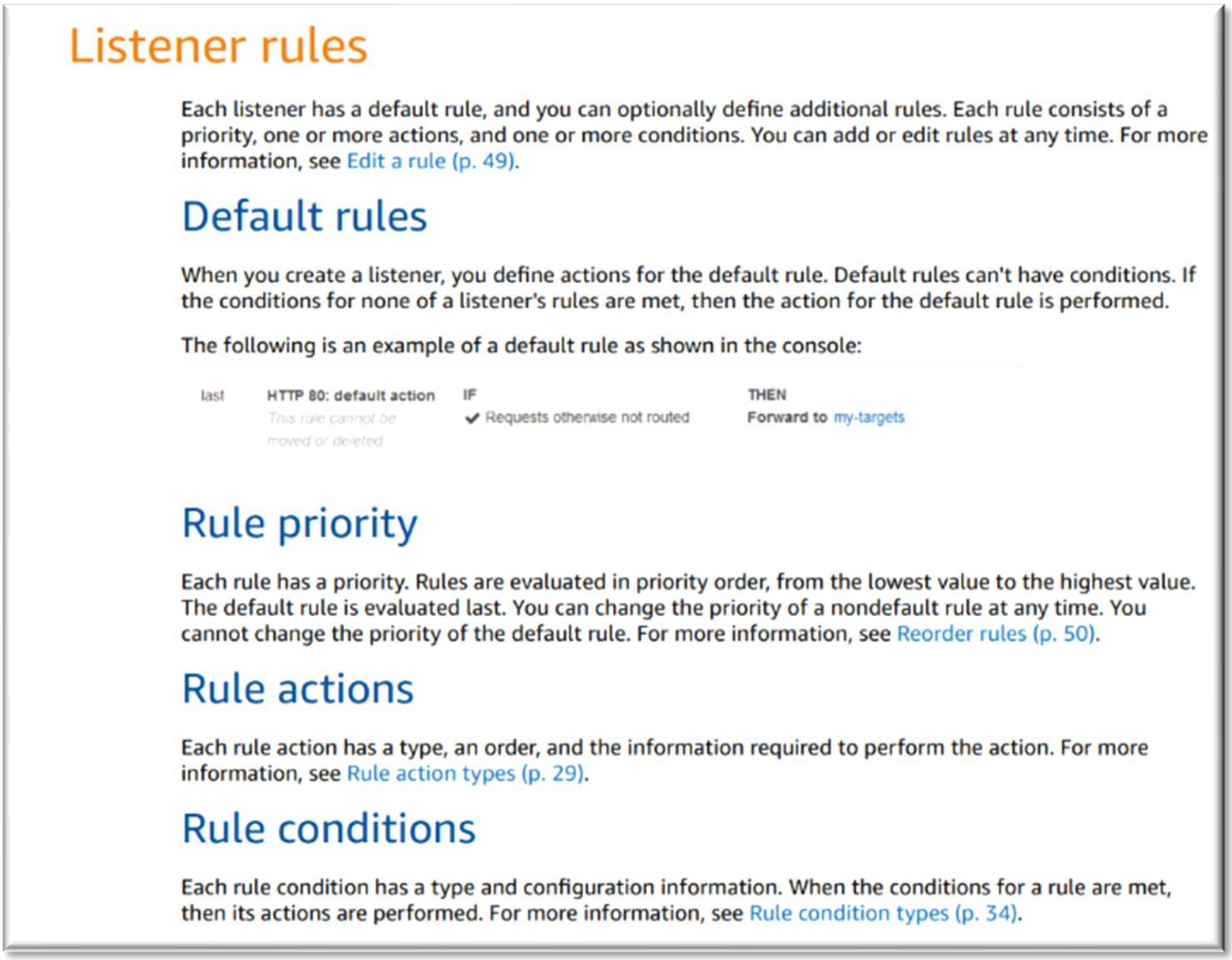
- *Amazon CloudWatch* enables you to monitor Amazon EC2 Instances and Elastic Load Balancers in real-time. It will aggregate and report on metrics such as CPU utilization, data transfer and disk usage, as well as requests rates and request latency.
- *Auto Scaling* allows you to automatically acquire and release Amazon EC2 Instances based on the metrics reported through Amazon CloudWatch. You can define the conditions under which this should happen and when these conditions are met, Auto Scaling will automatically add or remove compute resources.
- *Amazon Elastic Load Balancing* will distribute incoming application traffic over your Amazon EC2 instances that are running in a single or multiple Availability zones. It can detect the health of Amazon EC2 instances and will stop routing traffic to unhealthy instances until they have recovered and become healthy again.

These services will be of great value to Amazon Web Services customers to simplify the management of their applications and services. With the introduction of these services it will become even easier to optimize performance and fault-tolerance at low-cost.

The concepts of ELB

Target group

Listener



Listener rules

Each listener has a default rule, and you can optionally define additional rules. Each rule consists of a priority, one or more actions, and one or more conditions. You can add or edit rules at any time. For more information, see Edit a rule (p. 49).

Default rules

When you create a listener, you define actions for the default rule. Default rules can't have conditions. If the conditions for none of a listener's rules are met, then the action for the default rule is performed.

The following is an example of a default rule as shown in the console:

last	HTTP 80: default action This rule cannot be moved or deleted	IF ✔ Requests otherwise not routed	THEN Forward to my-targets

Rule priority

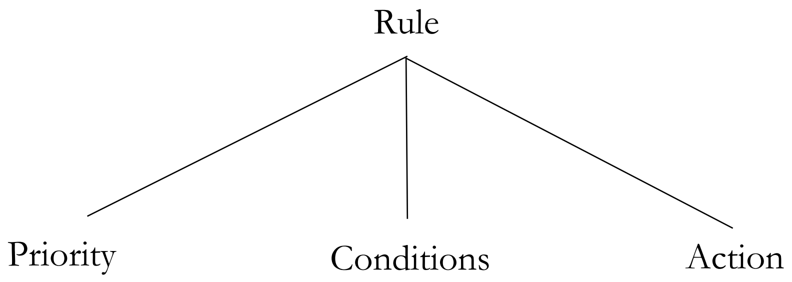
Each rule has a priority. Rules are evaluated in priority order, from the lowest value to the highest value. The default rule is evaluated last. You can change the priority of a nondefault rule at any time. You cannot change the priority of the default rule. For more information, see Reorder rules (p. 50).

Rule actions

Each rule action has a type, an order, and the information required to perform the action. For more information, see Rule action types (p. 29).

Rule conditions

Each rule condition has a type and configuration information. When the conditions for a rule are met, then its actions are performed. For more information, see Rule condition types (p. 34).

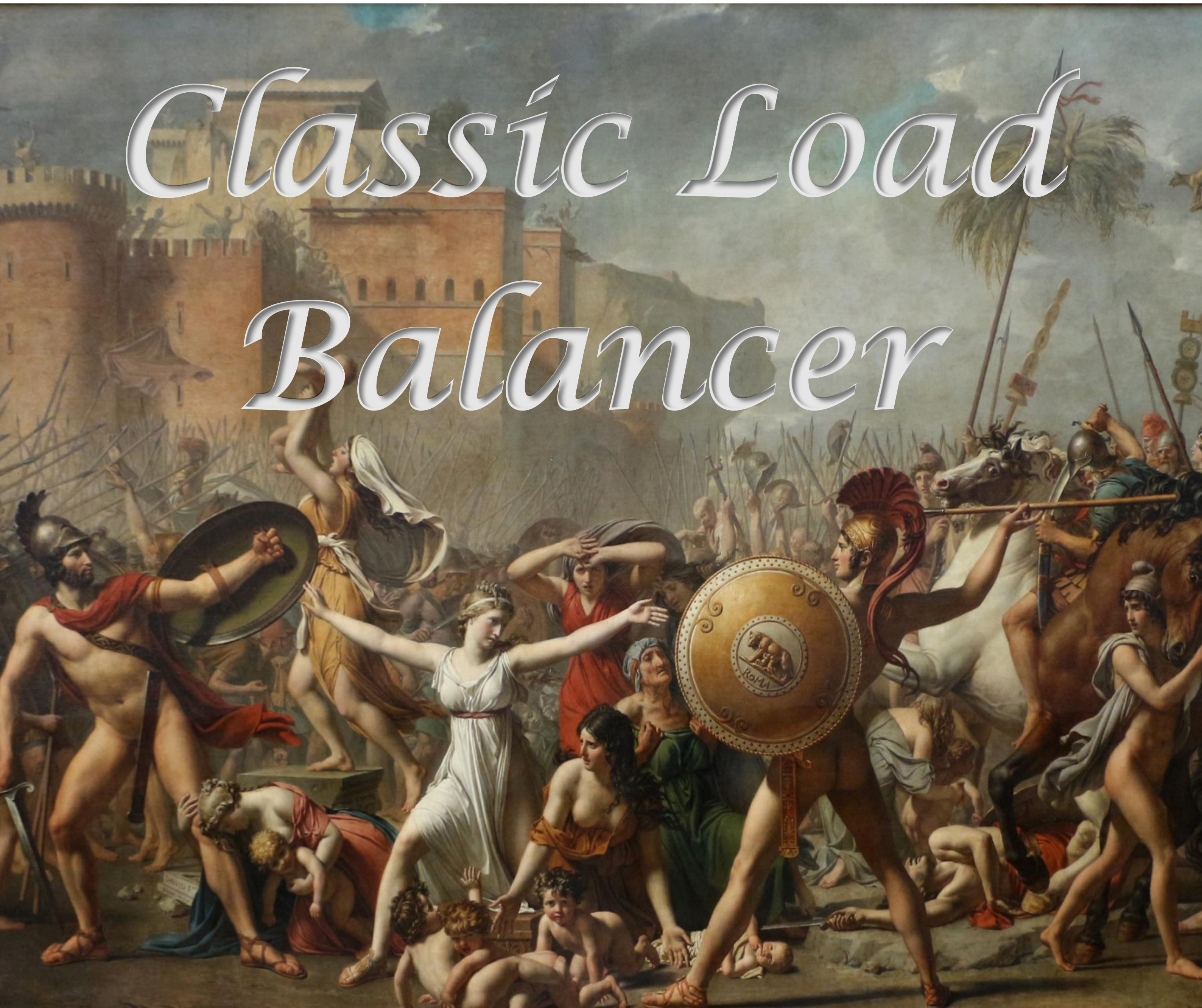


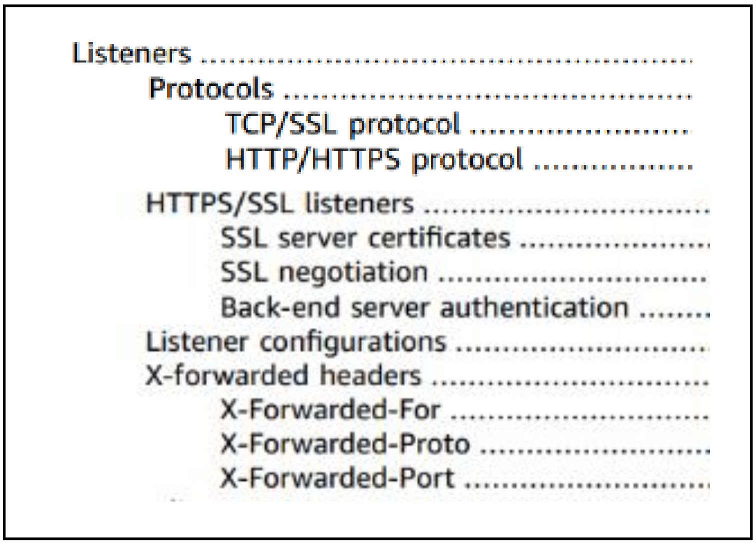
Brief Timeline

2009	AWS announce a package of three services: Auto Scaling, CloudWatch, and ELB (*Elastic Load Balancing*).
11th Aug 2016	AWS announce their “Application load balancer”.
7th September 2017	AWS announce their “Network load balancer”

The Classic Load Balancer

Obviously, this is a retrospective term. They didn't call it classic when they created it in 2009.

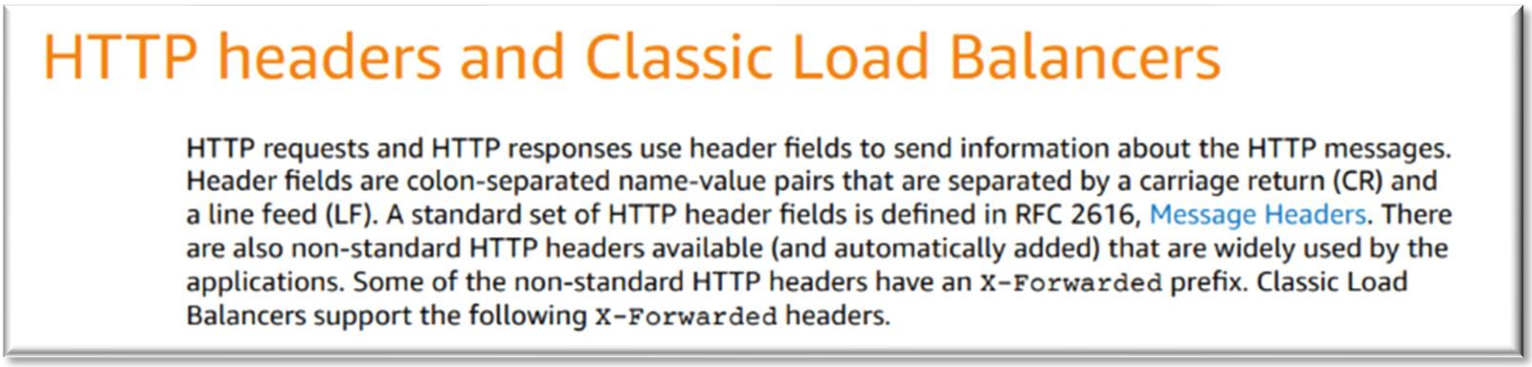




Listeners ....
Protocols ....
TCP/SSL protocol ....
HTTP/HTTPS protocol ....
HTTPS/SSL listeners ....
SSL server certificates ....
SSL negotiation ....
Back-end server authentication ....
Listener configurations ....
X-forwarded headers ....
X-Forwarded-For ....
X-Forwarded-Proto ....
X-Forwarded-Port ....

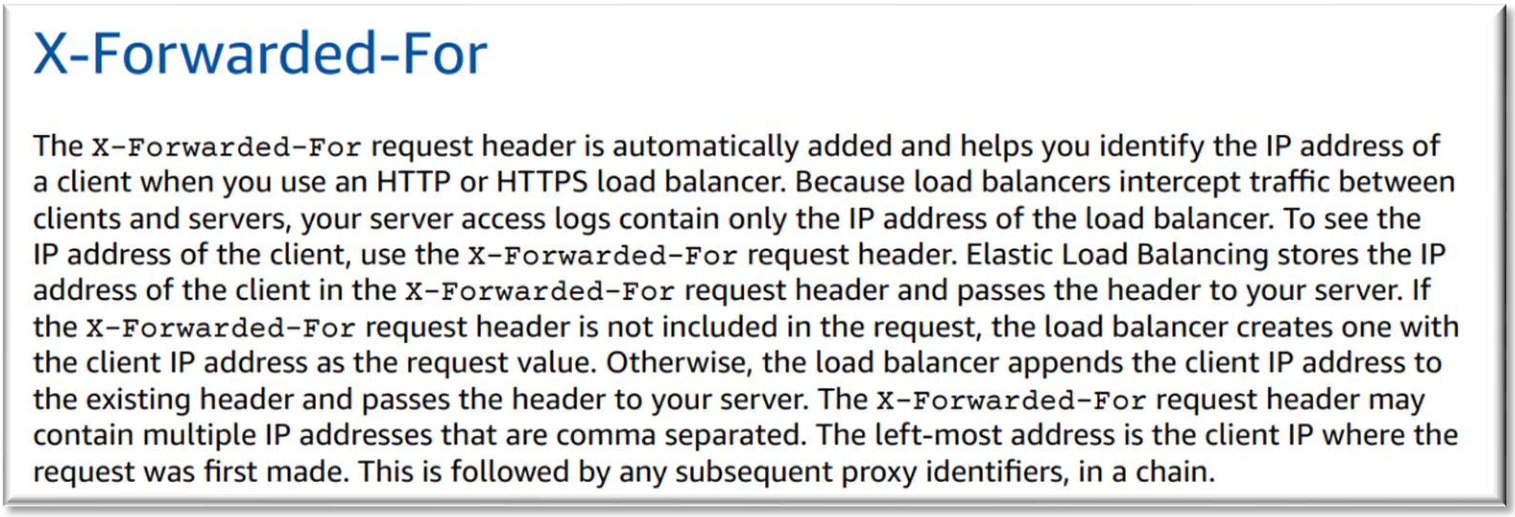
Screenshot from the documentation on the Classic Load Balancer. Chapter 6 is called "Listeners" and contains a discussion of X-forwarded headers.

What on earth is "X-Forwarded-For"?



HTTP headers and Classic Load Balancers

HTTP requests and HTTP responses use header fields to send information about the HTTP messages. Header fields are colon-separated name-value pairs that are separated by a carriage return (CR) and a line feed (LF). A standard set of HTTP header fields is defined in RFC 2616, Message Headers. There are also non-standard HTTP headers available (and automatically added) that are widely used by the applications. Some of the non-standard HTTP headers have an `X-Forwarded` prefix. Classic Load Balancers support the following `X-Forwarded` headers.



X-Forwarded-For

The `X-Forwarded-For` request header is automatically added and helps you identify the IP address of a client when you use an HTTP or HTTPS load balancer. Because load balancers intercept traffic between clients and servers, your server access logs contain only the IP address of the load balancer. To see the IP address of the client, use the `X-Forwarded-For` request header. Elastic Load Balancing stores the IP address of the client in the `X-Forwarded-For` request header and passes the header to your server. If the `X-Forwarded-For` request header is not included in the request, the load balancer creates one with the client IP address as the request value. Otherwise, the load balancer appends the client IP address to the existing header and passes the header to your server. The `X-Forwarded-For` request header may contain multiple IP addresses that are comma separated. The left-most address is the client IP where the request was first made. This is followed by any subsequent proxy identifiers, in a chain.

Eight funky things

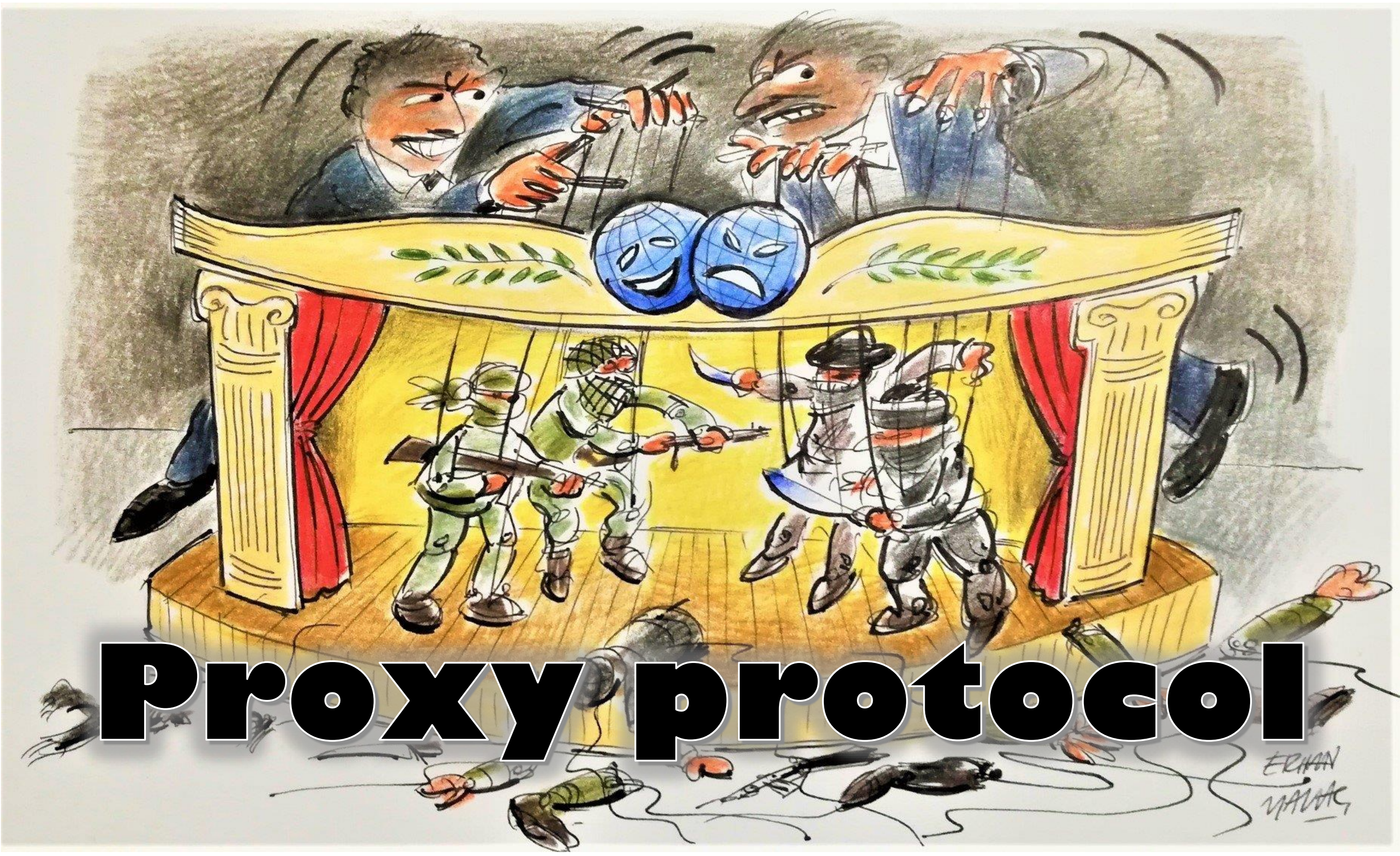
Chapter Eight of the CLB (Classic Load Balancer) documentation is entitled "**Configure your load balancer**". The discussion turns to a number of funky settings you can choose. There are eight of these funky, configurable settings in total:

1. The idle timeout
2. Cross-zone load balancing
3. Connection draining
4. Proxy protocol

5. Sticky sessions
6. Desync mitigation mode
7. Tag your load balancer
8. Domain name

Figure 1 - Classical depiction of a quarrel breaking out, after prolonged disagreement over the Eight Funky Things which can be configured on the Classic Load Balancer. Some have suggested that the central couple represents the server and client with the eight surrounding people represent the Eight Things which can be configured.

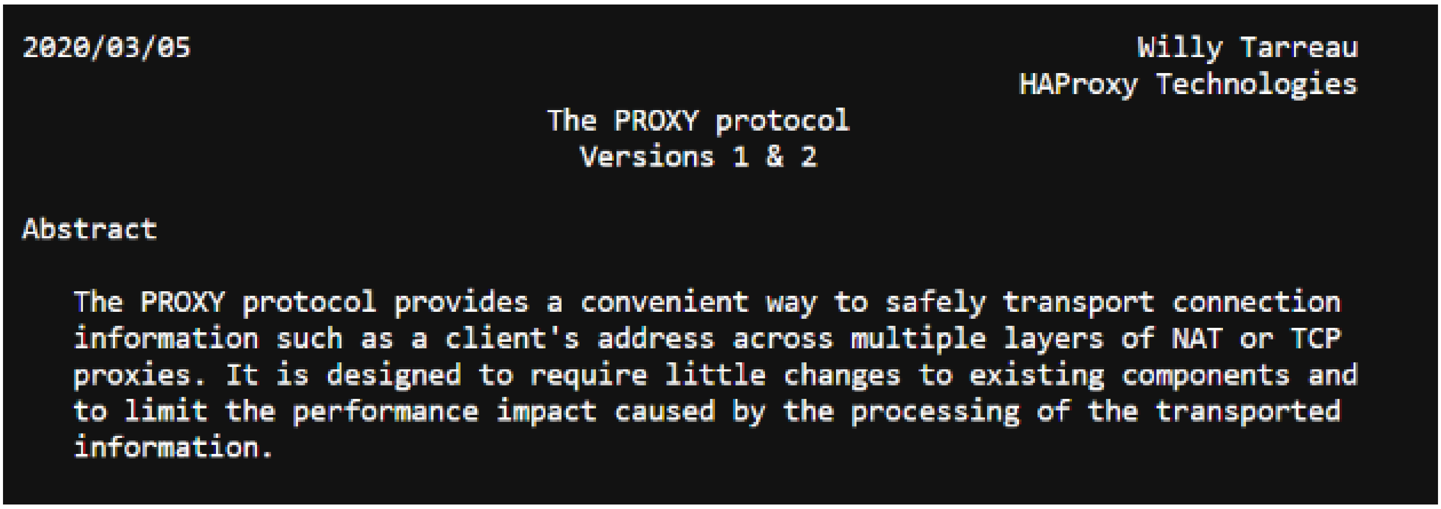
What on earth is “proxy protocol”?



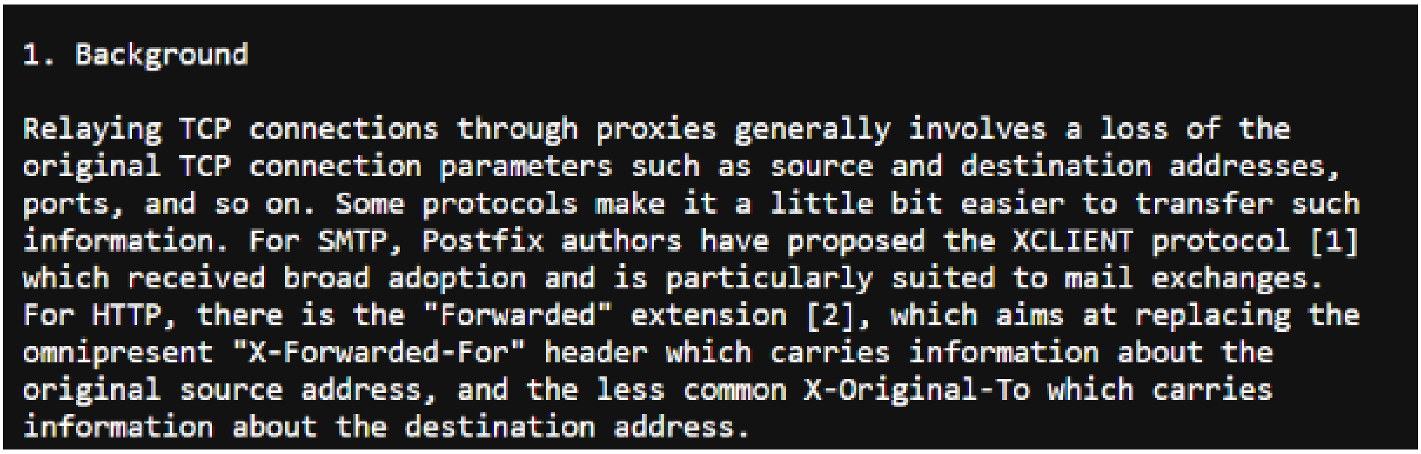
If something acts as a proxy, it acts on behalf of another entity, or represents another entity. You can vote by proxy (get someone to represent you at the ballot box) and nation states can fight wars by proxy (get some other state to defeat some enemy). Here, we're talking about clients and servers, with a load balancer in between.

When the proxy protocol in place, the client is represented by a header, bolted on by the load balancer. The packet, or perhaps the load balancer (since it latched on the header) acts on behalf of the client. With proxy protocol, the destination server can work out where the traffic *really* came from, as opposed to thinking it came from the load balancer itself.

Below is an RFC published in 2020. Willy Tarreau (shown right) is the author.



```
2020/03/05                                                  Willy Tarreau
                                                     HAProxy Technologies
                           The PROXY protocol
                             Versions 1 & 2

Abstract

   The PROXY protocol provides a convenient way to safely transport connection
   information such as a client's address across multiple layers of NAT or TCP
   proxies. It is designed to require little changes to existing components and
   to limit the performance impact caused by the processing of the transported
   information.
```



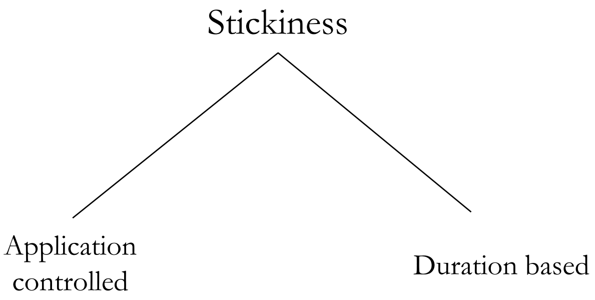
```
1. Background

Relaying TCP connections through proxies generally involves a loss of the
original TCP connection parameters such as source and destination addresses,
ports, and so on. Some protocols make it a little bit easier to transfer such
information. For SMTP, Postfix authors have proposed the XCLIENT protocol [1]
which received broad adoption and is particularly suited to mail exchanges.
For HTTP, there is the "Forwarded" extension [2], which aims at replacing the
omnipresent "X-Forwarded-For" header which carries information about the
original source address, and the less common X-Original-To which carries
information about the destination address.
```

What on earth is “connection draining”?

I'’ve started
so I’ll finish
MASTERMIND

STICKY SESSIONS

The documentation makes a distinction between "application controlled" stickiness and "duration based" stickiness.

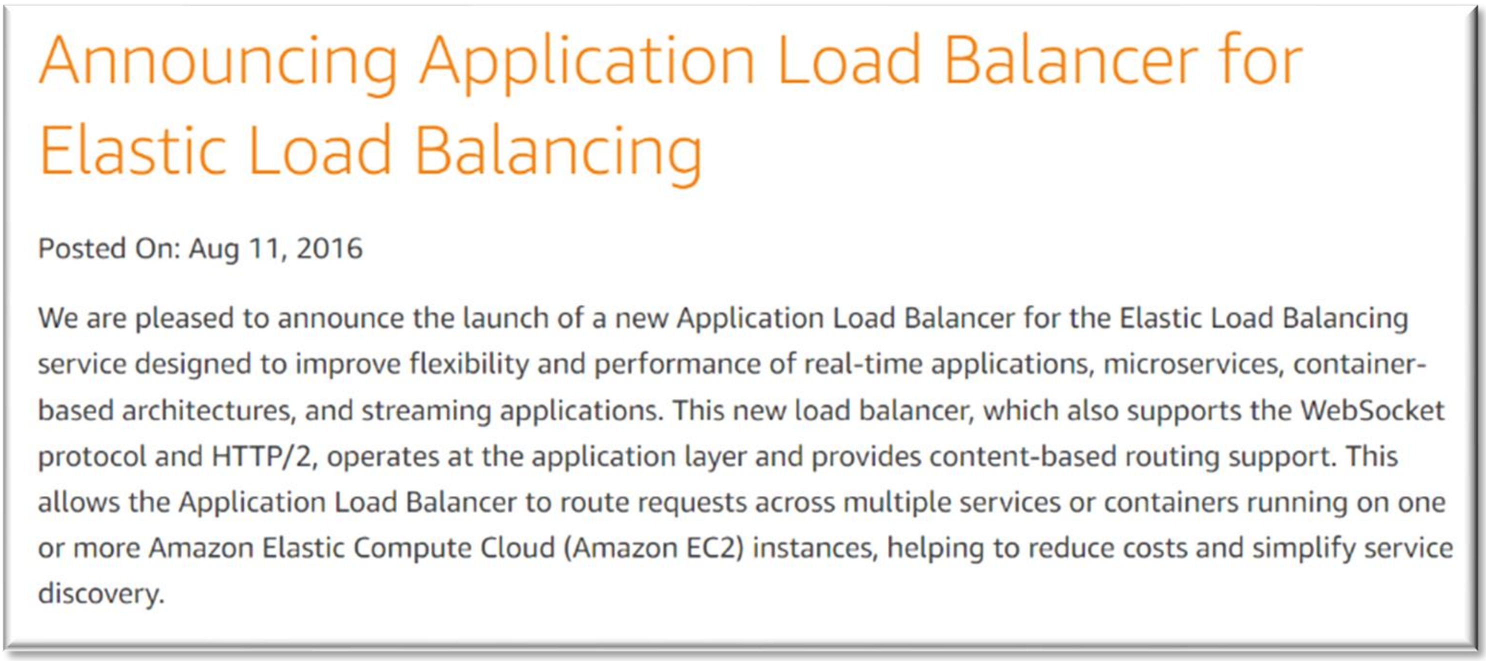


Stickiness

Application controlled

Duration based

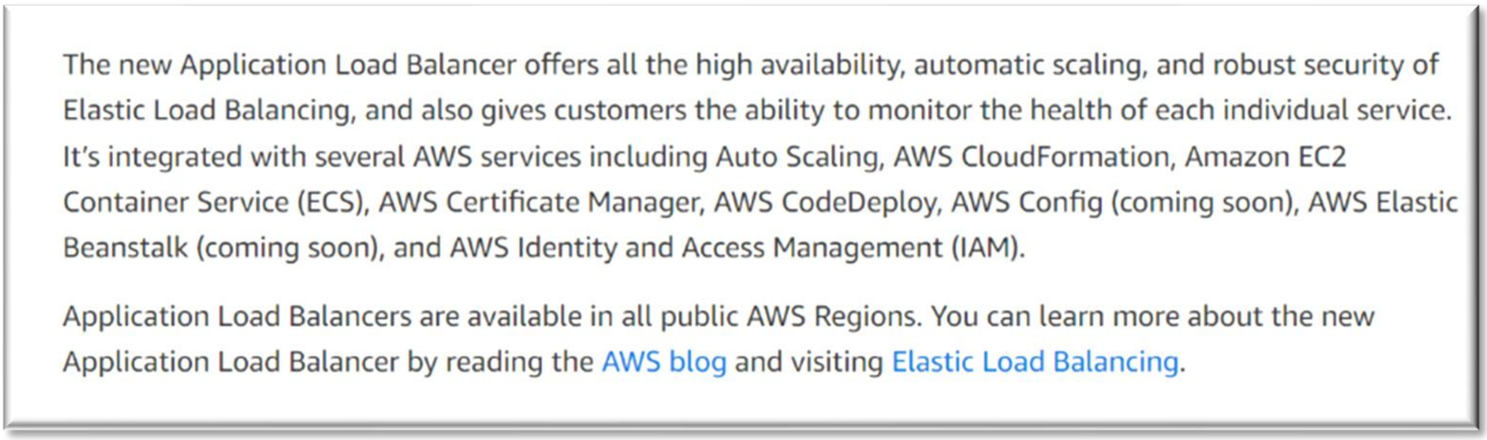
Application Load Balancer



Announcing Application Load Balancer for Elastic Load Balancing

Posted On: Aug 11, 2016

We are pleased to announce the launch of a new Application Load Balancer for the Elastic Load Balancing service designed to improve flexibility and performance of real-time applications, microservices, container-based architectures, and streaming applications. This new load balancer, which also supports the WebSocket protocol and HTTP/2, operates at the application layer and provides content-based routing support. This allows the Application Load Balancer to route requests across multiple services or containers running on one or more Amazon Elastic Compute Cloud (Amazon EC2) instances, helping to reduce costs and simplify service discovery.



The new Application Load Balancer offers all the high availability, automatic scaling, and robust security of Elastic Load Balancing, and also gives customers the ability to monitor the health of each individual service. It's integrated with several AWS services including Auto Scaling, AWS CloudFormation, Amazon EC2 Container Service (ECS), AWS Certificate Manager, AWS CodeDeploy, AWS Config (coming soon), AWS Elastic Beanstalk (coming soon), and AWS Identity and Access Management (IAM).

Application Load Balancers are available in all public AWS Regions. You can learn more about the new Application Load Balancer by reading the AWS blog and visiting Elastic Load Balancing.

Matt Wood announcing the Application Load Balancer in 2016

What capabilities did the ALB bring?

When AWS announced their Application Load Balancer (ALB) in 2016, what things could it do which could not previously be done with Elastic Load Balancing? The pre-ALB capabilities are generally referred to as the "Classic Load Balancer".

When Matt Wood announced the ALB in 2016, he described it as:

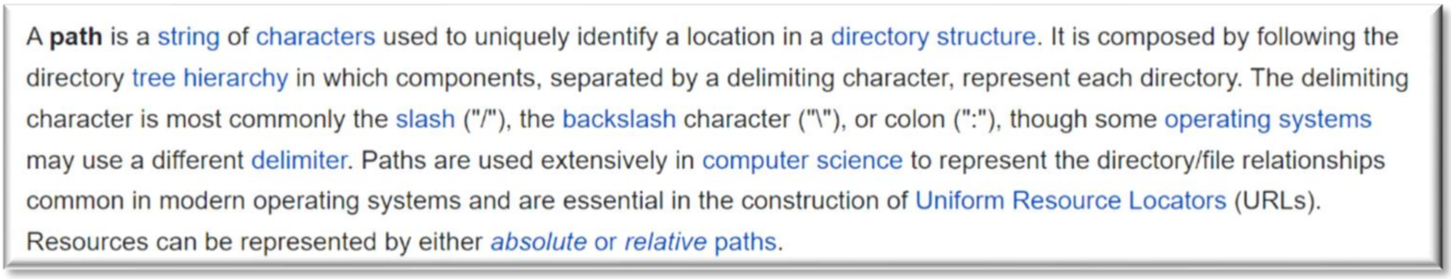
> a higher order routing abstraction, which gets the right traffic to the right application component.
>
> The service also supports enhanced metrics and health checks, to monitor traffic to each application service behind the load balancer.

Wood says that the ALB will be particularly helpful for container-based applications. We can be quite confident that the ALB was a Layer 7 load balancer, which supported both EC2 instances and containers as targets, from the offset. It has other features, but these may have been added later—working out which were added later is in fact quite difficult, so I will leave this historical task for now.

What is undoubtedly true is that a defining feature of the ALB is "content based routing". What does this mean? Well, we are

What is path-based routing?

First, we need to gain a precise understanding of a "path".



A **path** is a string of characters used to uniquely identify a location in a directory structure. It is composed by following the directory tree hierarchy in which components, separated by a delimiting character, represent each directory. The delimiting character is most commonly the slash ("/"), the backslash character ("\"), or colon (":"), though some operating systems may use a different delimiter. Paths are used extensively in computer science to represent the directory/file relationships common in modern operating systems and are essential in the construction of Uniform Resource Locators (URLs). Resources can be represented by either *absolute* or *relative* paths.

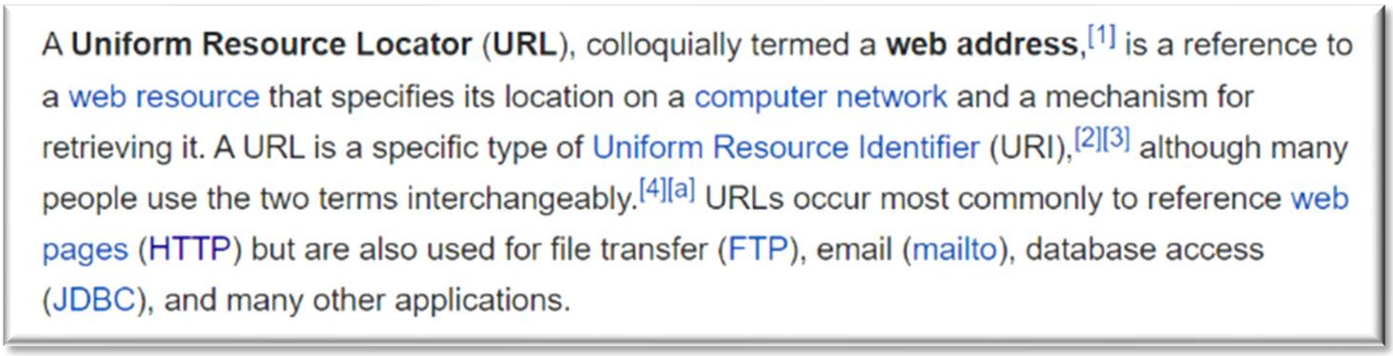
Figure 1 The Wikipedia article entitled "Path (computing)"

We are clearly not working in the domain of storage here (Elastic Load Balancing is usually part of *networking* discussions). So we are surely not concerned with files within directories. What do we mean by ‘path’, then?

Well, as the article above tells us, paths “are essential in the construction of Uniform Resource Locators (URLS)”. Notice how strong the claim is. We’re being told:

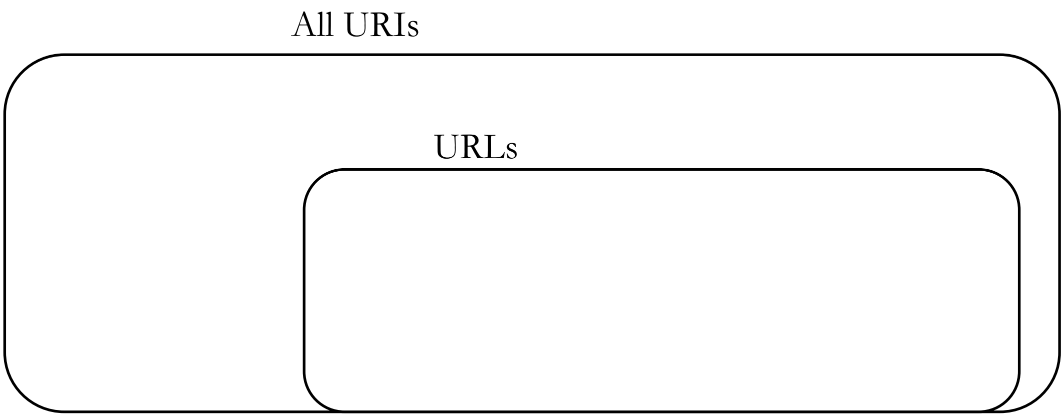
> If there is a URL, then there is a path.

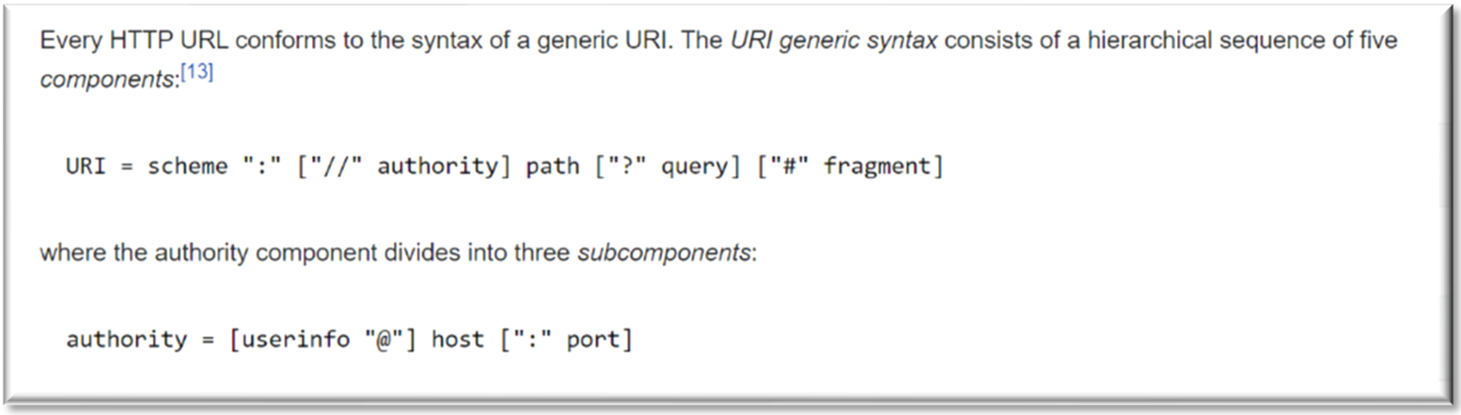
We’re not saying the URL *is* the path. Rather, the path is one part of the URL. We’re going to have to dive slightly deeper into URLs to understand.



A **Uniform Resource Locator** (**URL**), colloquially termed a **web address**,[1] is a reference to a web resource that specifies its location on a computer network and a mechanism for retrieving it. A URL is a specific type of Uniform Resource Identifier (URI),[2][3] although many people use the two terms interchangeably.[4][a] URLs occur most commonly to reference web pages (HTTP) but are also used for file transfer (FTP), email (mailto), database access (JDBC), and many other applications.

This tells us that a URL is just one specific kind of URI. In other words:





Every HTTP URL conforms to the syntax of a generic URI. The *URI generic syntax* consists of a hierarchical sequence of five *components*:[13]

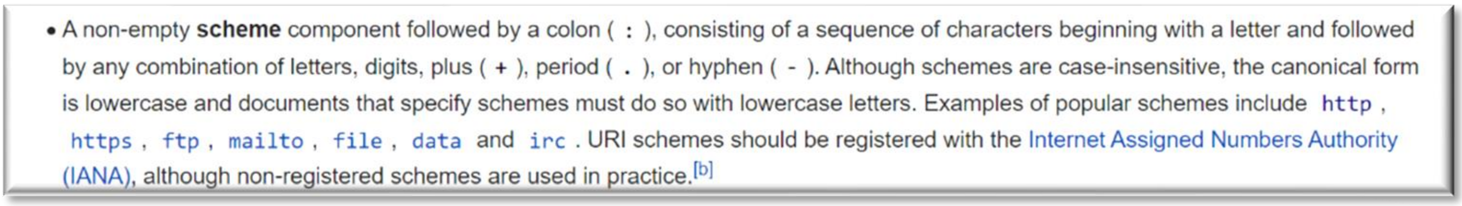
```
URI = scheme ":" ["//" authority] path ["?" query] ["#" fragment]
```

where the authority component divides into three *subcomponents*:

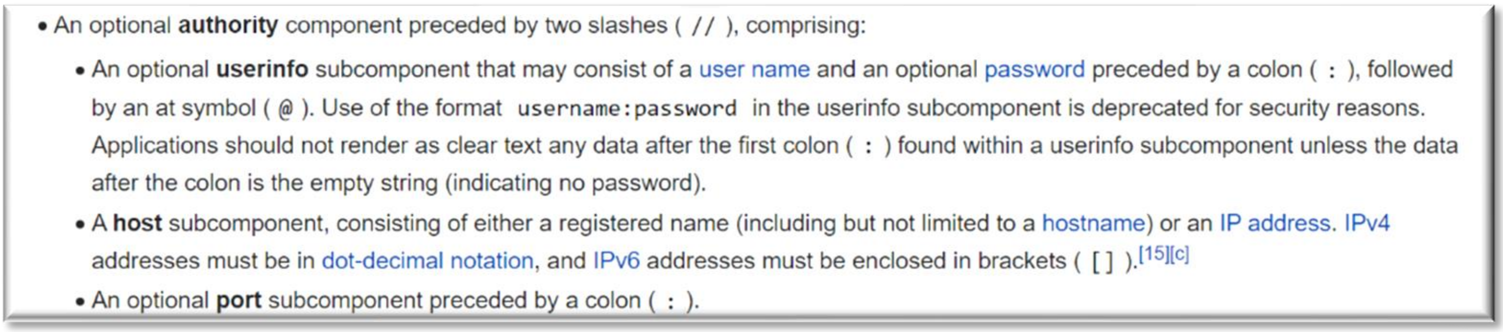
```
authority = [userinfo "@"] host [":" port]
```

Breaking down a URI

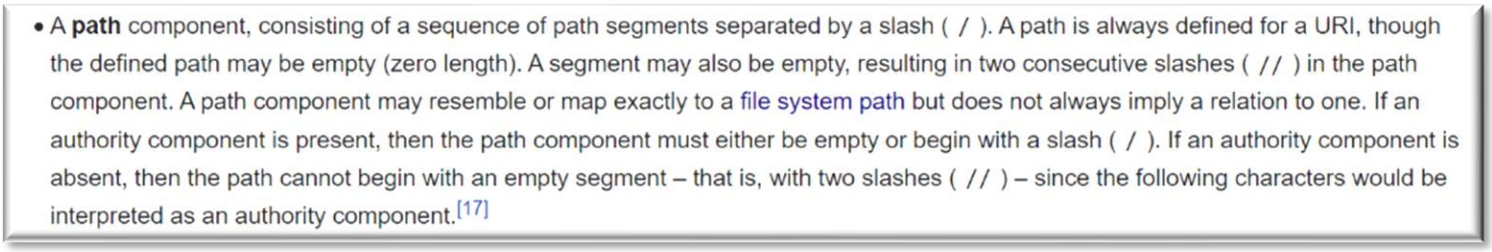
We're now going to break down the parts of a URI.



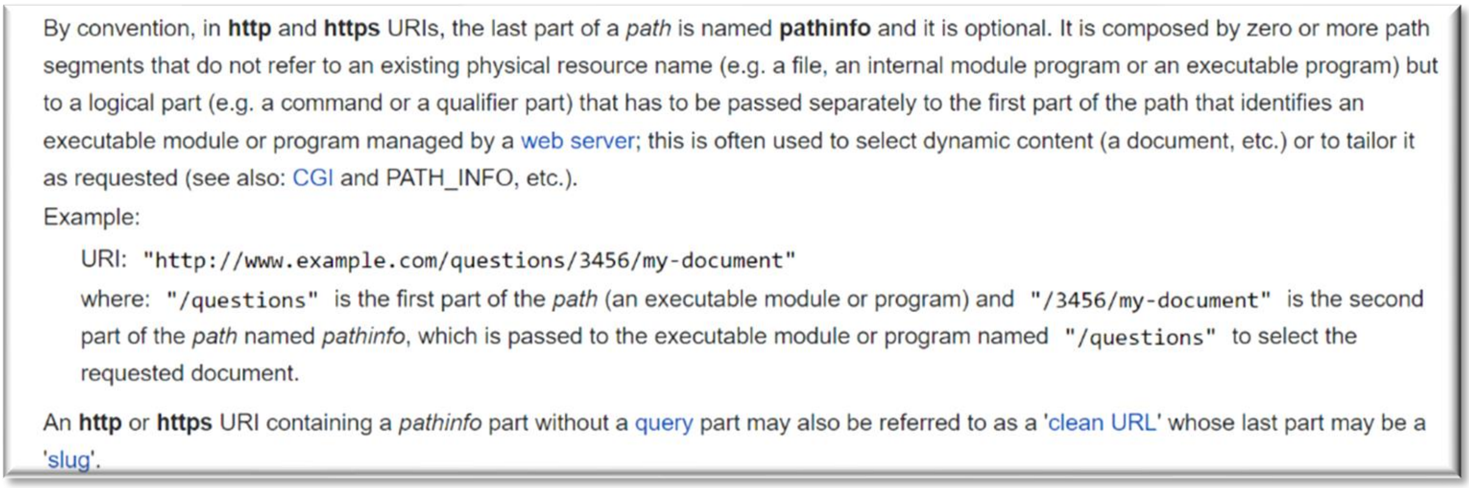
- A non-empty **scheme** component followed by a colon (`:`), consisting of a sequence of characters beginning with a letter and followed by any combination of letters, digits, plus (`+`), period (`.`), or hyphen (`-`). Although schemes are case-insensitive, the canonical form is lowercase and documents that specify schemes must do so with lowercase letters. Examples of popular schemes include `http` , `https` , `ftp` , `mailto` , `file` , `data` and `irc` . URI schemes should be registered with the Internet Assigned Numbers Authority (IANA), although non-registered schemes are used in practice.[b]



- An optional **authority** component preceded by two slashes (`//`), comprising:
  - An optional **userinfo** subcomponent that may consist of a user name and an optional password preceded by a colon (`:`), followed by an at symbol (`@`). Use of the format `username:password` in the userinfo subcomponent is deprecated for security reasons. Applications should not render as clear text any data after the first colon (`:`) found within a userinfo subcomponent unless the data after the colon is the empty string (indicating no password).
  - A **host** subcomponent, consisting of either a registered name (including but not limited to a hostname) or an IP address. IPv4 addresses must be in dot-decimal notation, and IPv6 addresses must be enclosed in brackets (`[ ]`).[15][c]
  - An optional **port** subcomponent preceded by a colon (`:`).



- A **path** component, consisting of a sequence of path segments separated by a slash (`/`). A path is always defined for a URI, though the defined path may be empty (zero length). A segment may also be empty, resulting in two consecutive slashes (`//`) in the path component. A path component may resemble or map exactly to a file system path but does not always imply a relation to one. If an authority component is present, then the path component must either be empty or begin with a slash (`/`). If an authority component is absent, then the path cannot begin with an empty segment – that is, with two slashes (`//`) – since the following characters would be interpreted as an authority component.[17]



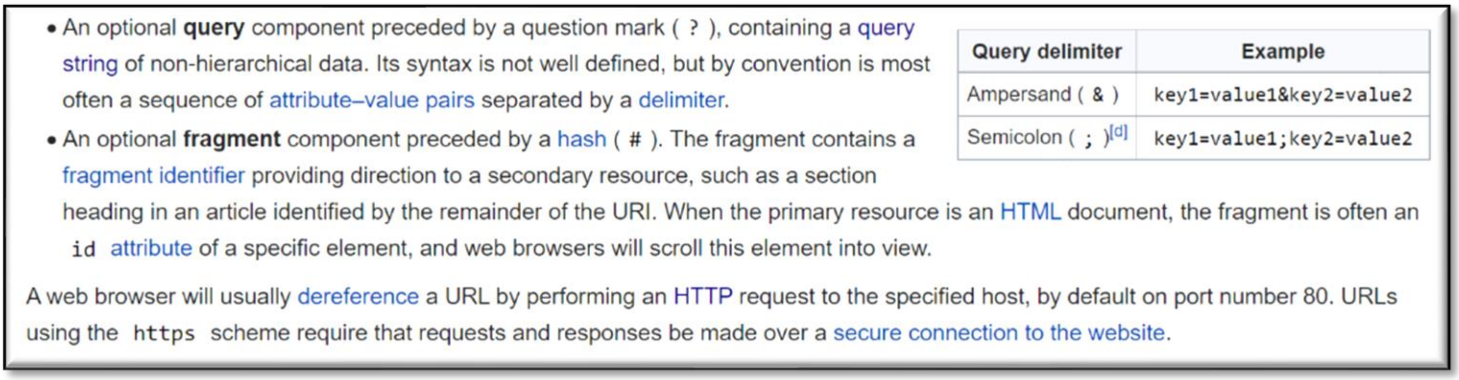
By convention, in **http** and **https** URIs, the last part of a *path* is named **pathinfo** and it is optional. It is composed by zero or more path segments that do not refer to an existing physical resource name (e.g. a file, an internal module program or an executable program) but to a logical part (e.g. a command or a qualifier part) that has to be passed separately to the first part of the path that identifies an executable module or program managed by a web server; this is often used to select dynamic content (a document, etc.) or to tailor it as requested (see also: CGI and PATH_INFO, etc.).

Example:

URI: `"http://www.example.com/questions/3456/my-document"`

where: `"/questions"` is the first part of the *path* (an executable module or program) and `"/3456/my-document"` is the second part of the *path* named *pathinfo*, which is passed to the executable module or program named `"/questions"` to select the requested document.

An **http** or **https** URI containing a *pathinfo* part without a query part may also be referred to as a 'clean URL' whose last part may be a 'slug'.



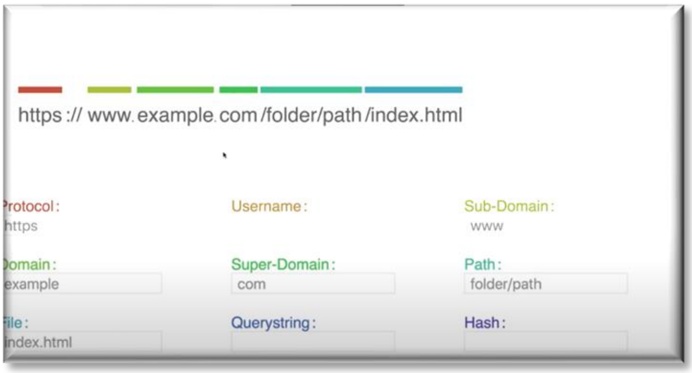
- An optional **query** component preceded by a question mark (`?`), containing a query string of non-hierarchical data. Its syntax is not well defined, but by convention is most often a sequence of attribute–value pairs separated by a delimiter.

Query delimiter	Example
Ampersand (`&`)	`key1=value1&key2=value2`
Semicolon (`;`)[d]	`key1=value1;key2=value2`

- An optional **fragment** component preceded by a hash (`#`). The fragment contains a fragment identifier providing direction to a secondary resource, such as a section heading in an article identified by the remainder of the URI. When the primary resource is an HTML document, the fragment is often an `id` attribute of a specific element, and web browsers will scroll this element into view.

A web browser will usually dereference a URL by performing an HTTP request to the specified host, by default on port number 80. URLs using the `https` scheme require that requests and responses be made over a secure connection to the website.

Some videos on URLs:



https :// www. example. com /folder/path /index.html

Protocol: https

Username:

Sub-Domain: www

Domain: example

Super-Domain: com

Path: folder/path

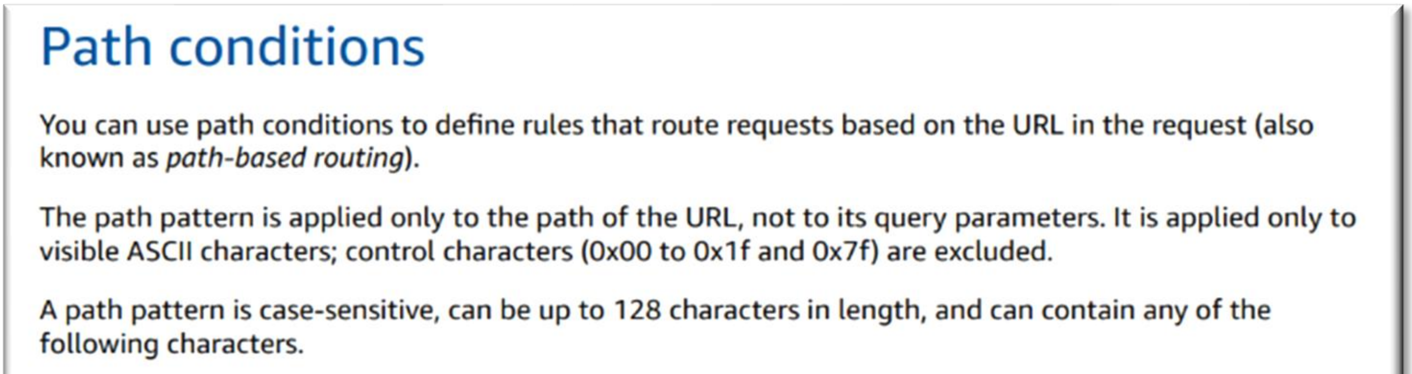
File: index.html

Querystring:

Hash:

Back to path-based routing

Now that we are clear about what a "path" is, we can understand routing based on the path. We are paying attention to one particular part of the URL. We are *not* necessarily routing based on the domain name (alone); *not* the name of the file (as such); *not* the query string. This is path-based routing.

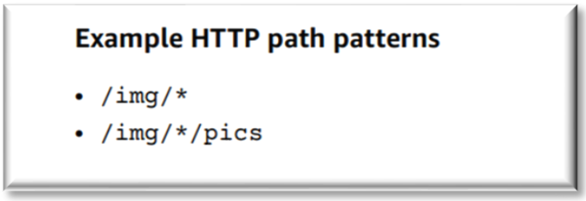


Path conditions

You can use path conditions to define rules that route requests based on the URL in the request (also known as *path-based routing*).

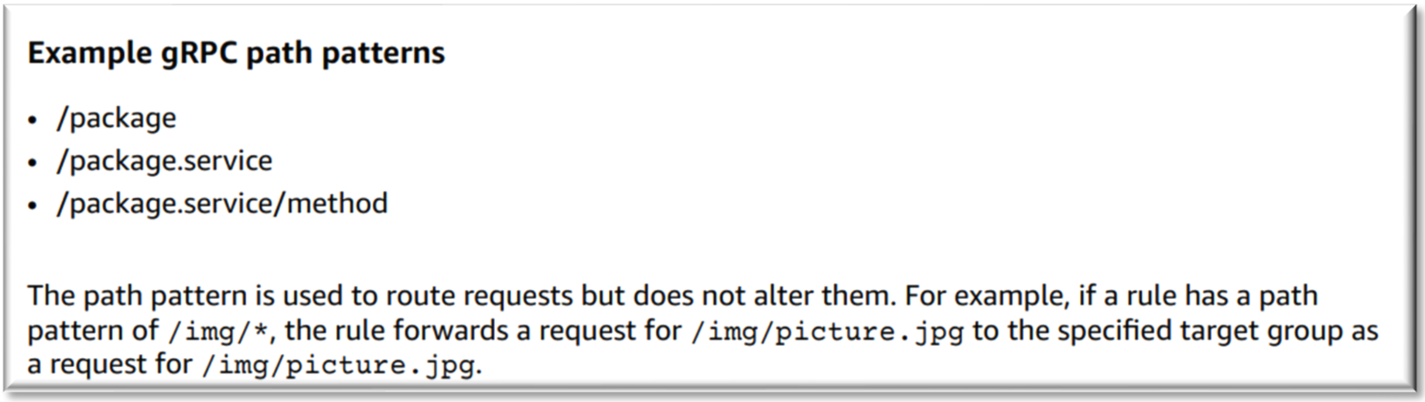
The path pattern is applied only to the path of the URL, not to its query parameters. It is applied only to visible ASCII characters; control characters (0x00 to 0x1f and 0x7f) are excluded.

A path pattern is case-sensitive, can be up to 128 characters in length, and can contain any of the following characters.



Example HTTP path patterns

- `/img/*`
- `/img/*/pics`

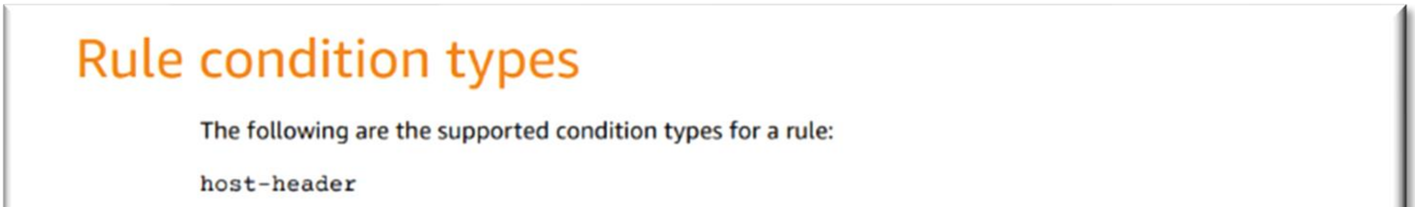


Example gRPC path patterns

- /package
- /package.service
- /package.service/method

The path pattern is used to route requests but does not alter them. For example, if a rule has a path pattern of `/img/*`, the rule forwards a request for `/img/picture.jpg` to the specified target group as a request for `/img/picture.jpg`.

How does this all work in the terminology of ELB? Well, remember that every ALB must have at least one listener. The listener operates according to *RULES*. Rules consist of. And so path-based routing is just one of the possible conditions in a rule:



Rule condition types

The following are the supported condition types for a rule:

`host-header`

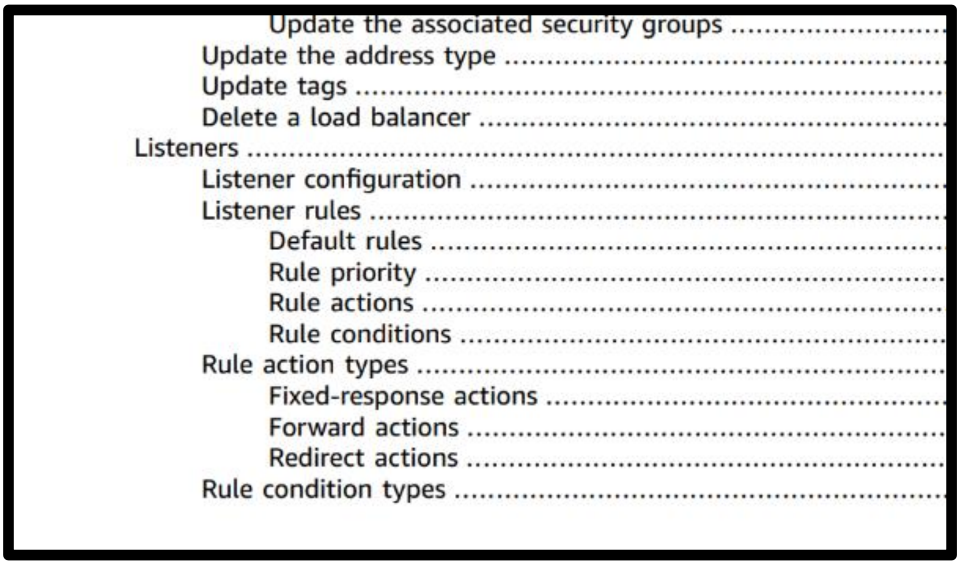
NOTE

Domain names, paths and URLs

We will have a detailed discussion of load balancers, and their different capabilities. To do this, make sure you have a **clear and crisp** idea of the following, and how they differ (See Appendix in the FOUNDATION book).

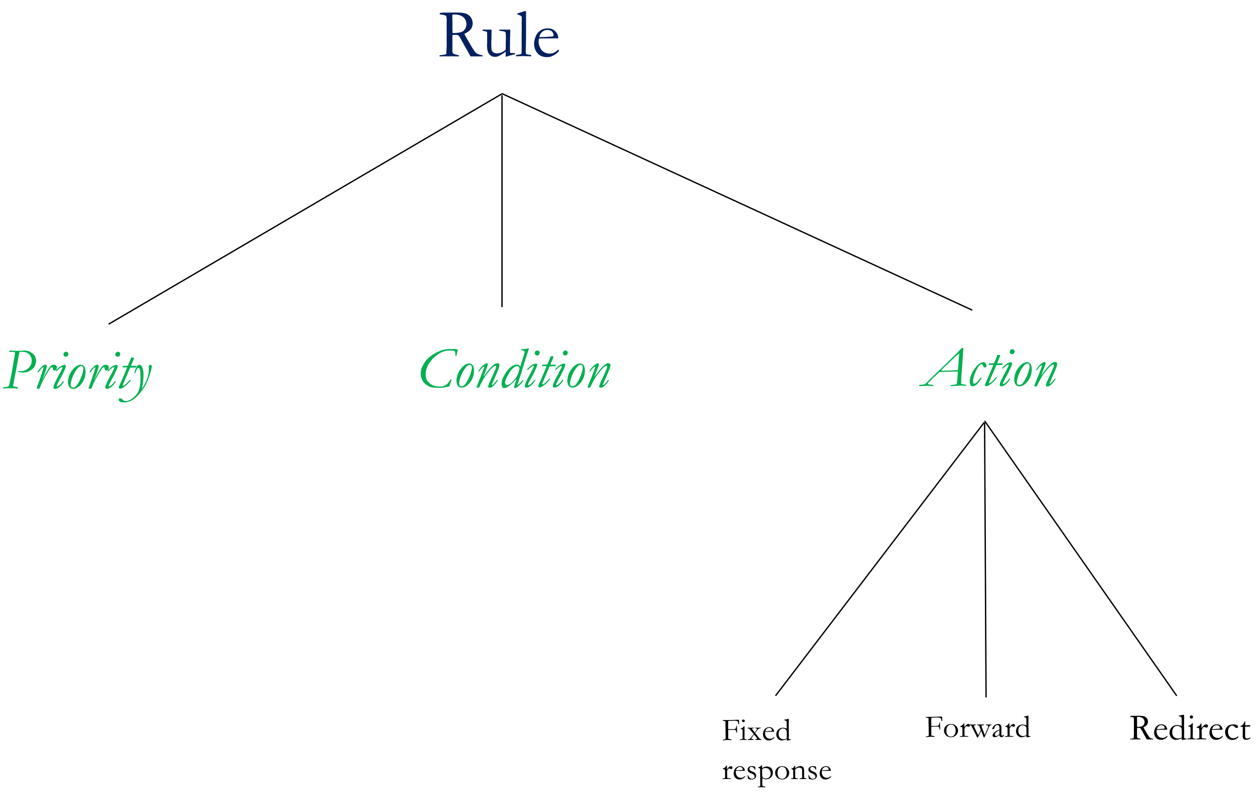
1. Domain name
2. URL
3. Endpoint (not really mentioned in this context)
4. Path
5. Query string parameter
6. Host

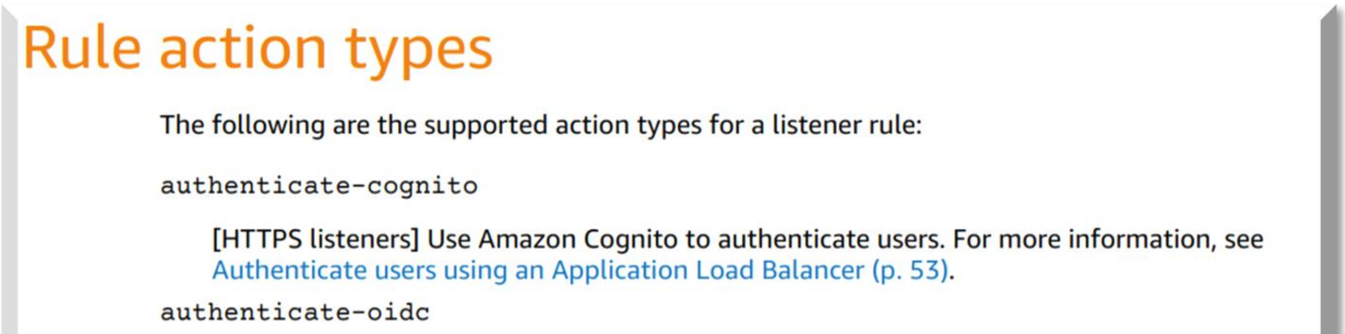
What is a fixed-response action?



- Update the associated security groups
- Update the address type
- Update tags
- Delete a load balancer
- Listeners
  - Listener configuration
  - Listener rules
    - Default rules
    - Rule priority
    - Rule actions
    - Rule conditions
  - Rule action types
    - Fixed-response actions
    - Forward actions
    - Redirect actions
  - Rule condition types

Screenshot from the contents page of the Application Load Balancer. A section is devoted to "rule action types". Recall that rules consist of a priority, conditions and actions.





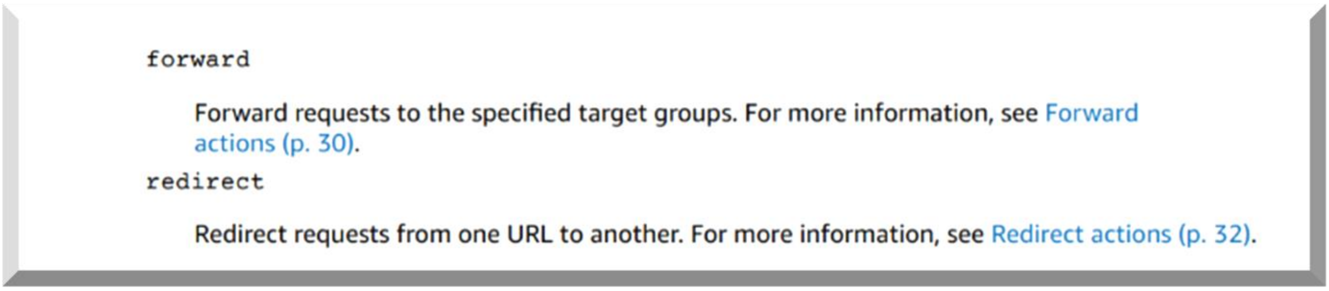
Rule action types

The following are the supported action types for a listener rule:

`authenticate-cognito`

[HTTPS listeners] Use Amazon Cognito to authenticate users. For more information, see Authenticate users using an Application Load Balancer (p. 53).

`authenticate-oidc`



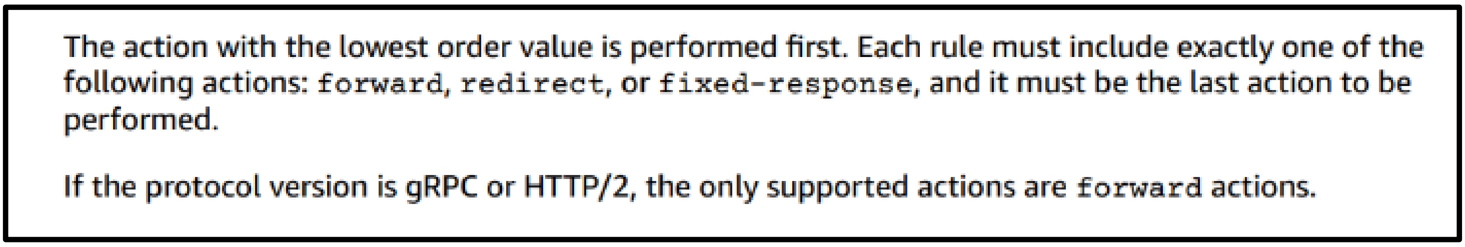
`forward`

Forward requests to the specified target groups. For more information, see Forward actions (p. 30).

`redirect`

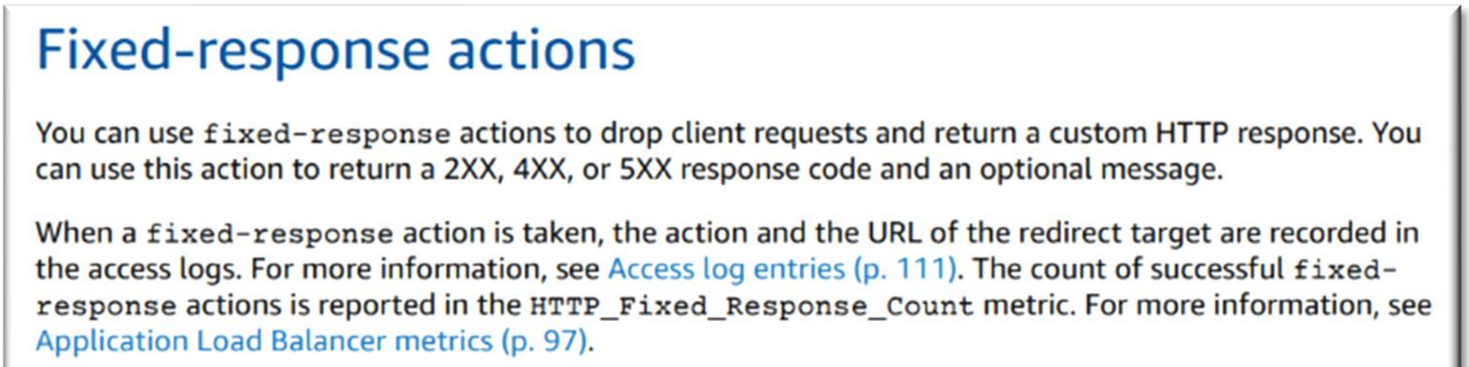
Redirect requests from one URL to another. For more information, see Redirect actions (p. 32).

Above, we're given five possibilities for the action type. It is made clear that if you are using the HTTP/2 protocol, you do not have the luxury of choosing any of the five possibilities:



The action with the lowest order value is performed first. Each rule must include exactly one of the following actions: `forward`, `redirect`, or `fixed-response`, and it must be the last action to be performed.

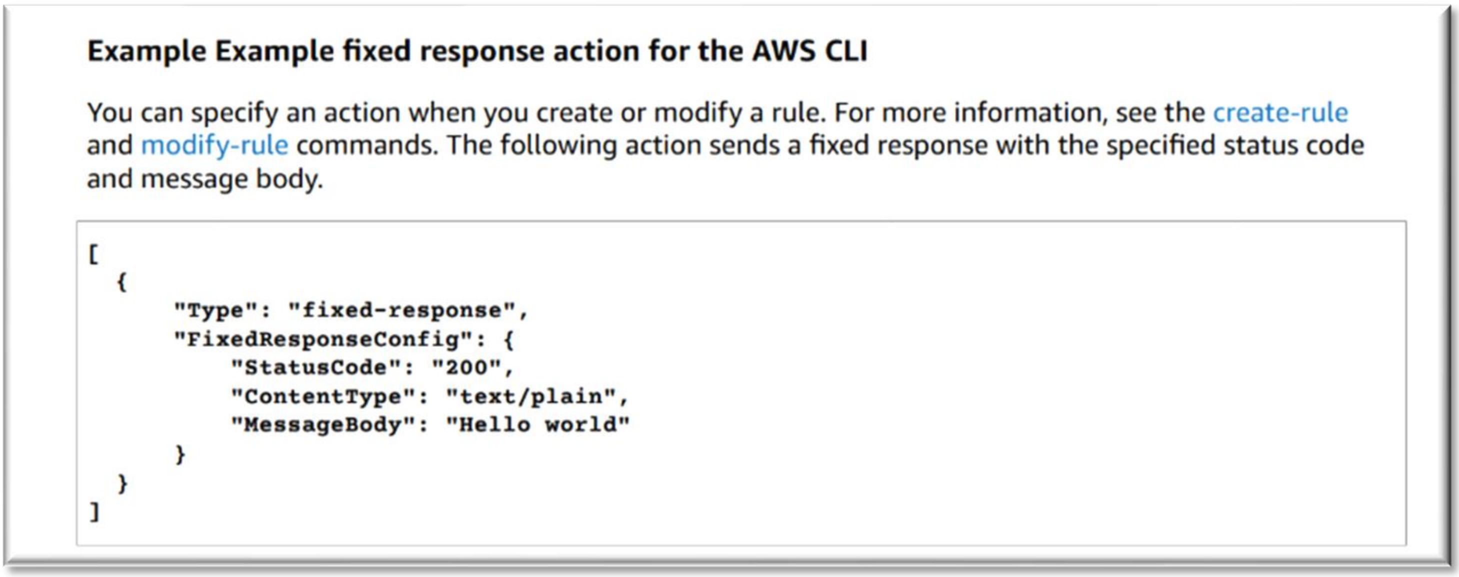
If the protocol version is gRPC or HTTP/2, the only supported actions are `forward` actions.



Fixed-response actions

You can use `fixed-response` actions to drop client requests and return a custom HTTP response. You can use this action to return a 2XX, 4XX, or 5XX response code and an optional message.

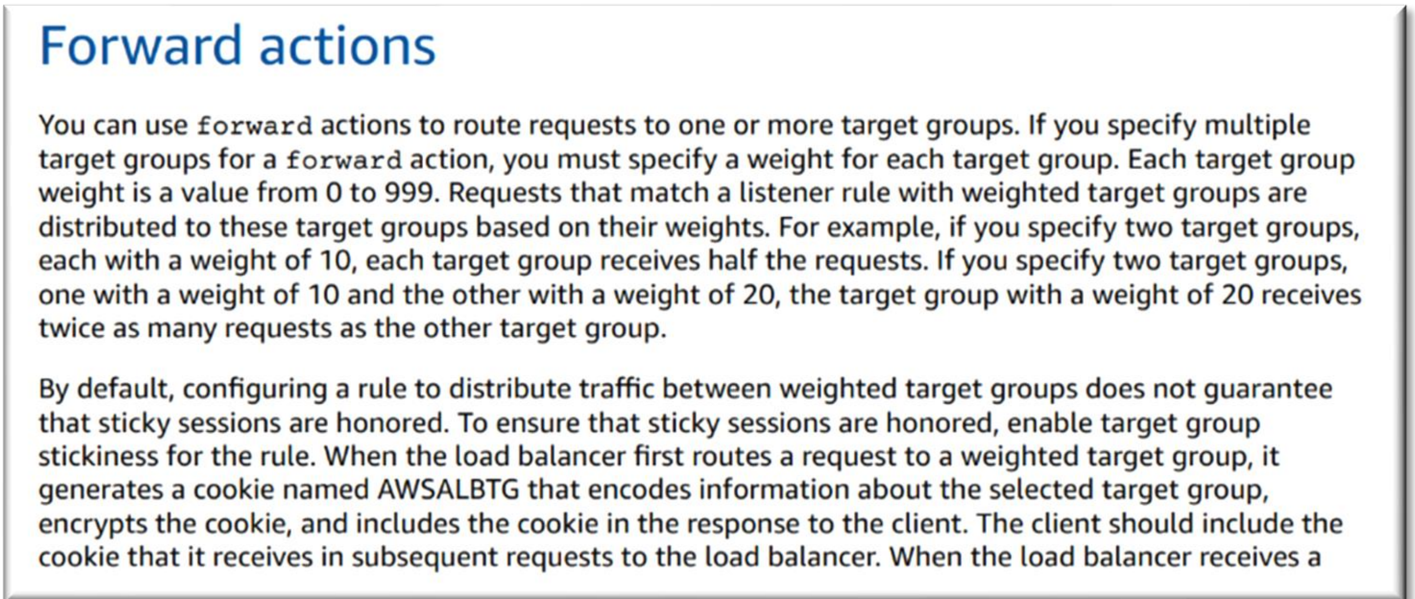
When a `fixed-response` action is taken, the action and the URL of the redirect target are recorded in the access logs. For more information, see Access log entries (p. 111). The count of successful `fixed-response` actions is reported in the `HTTP_Fixed_Response_Count` metric. For more information, see Application Load Balancer metrics (p. 97).



Example Example fixed response action for the AWS CLI

You can specify an action when you create or modify a rule. For more information, see the create-rule and modify-rule commands. The following action sends a fixed response with the specified status code and message body.

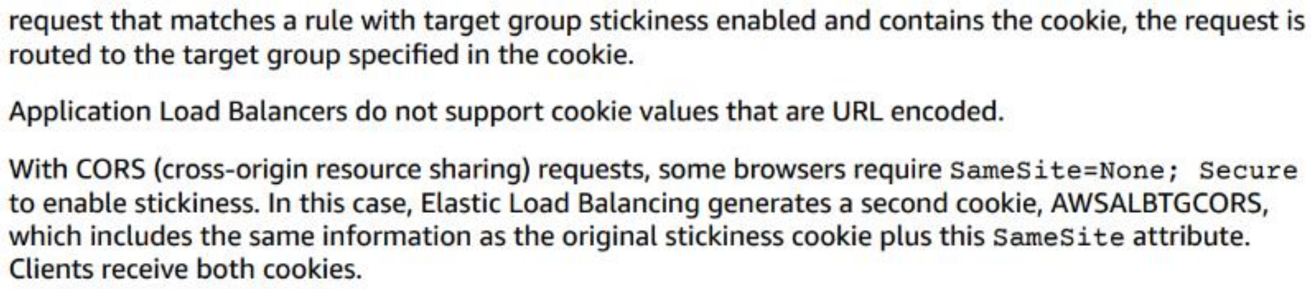
```
[
  {
      "Type": "fixed-response",
      "FixedResponseConfig": {
          "StatusCode": "200",
          "ContentType": "text/plain",
          "MessageBody": "Hello world"
      }
  }
]
```



Forward actions

You can use `forward` actions to route requests to one or more target groups. If you specify multiple target groups for a `forward` action, you must specify a weight for each target group. Each target group weight is a value from 0 to 999. Requests that match a listener rule with weighted target groups are distributed to these target groups based on their weights. For example, if you specify two target groups, each with a weight of 10, each target group receives half the requests. If you specify two target groups, one with a weight of 10 and the other with a weight of 20, the target group with a weight of 20 receives twice as many requests as the other target group.

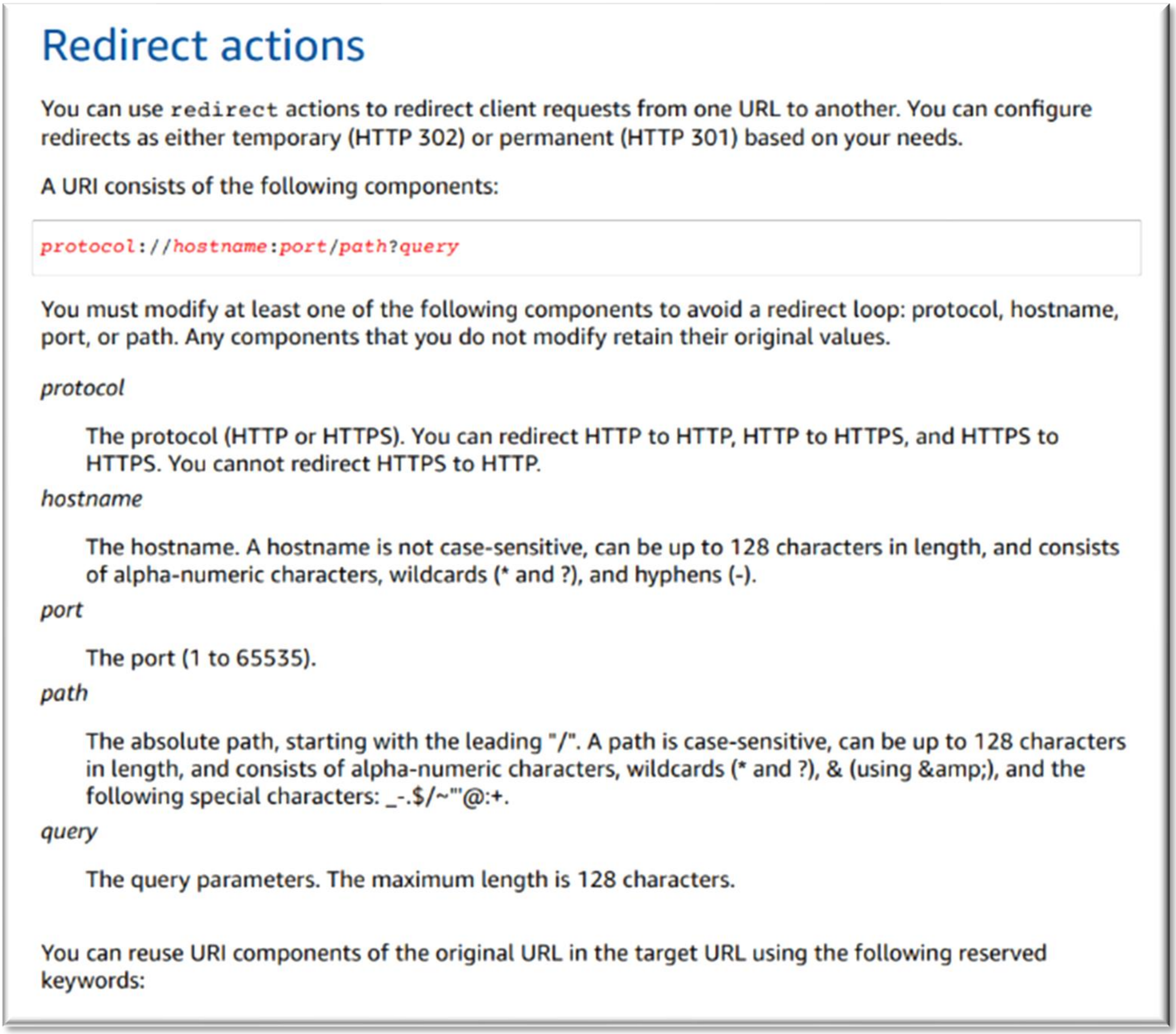
By default, configuring a rule to distribute traffic between weighted target groups does not guarantee that sticky sessions are honored. To ensure that sticky sessions are honored, enable target group stickiness for the rule. When the load balancer first routes a request to a weighted target group, it generates a cookie named AWSALBTG that encodes information about the selected target group, encrypts the cookie, and includes the cookie in the response to the client. The client should include the cookie that it receives in subsequent requests to the load balancer. When the load balancer receives a



request that matches a rule with target group stickiness enabled and contains the cookie, the request is routed to the target group specified in the cookie.

Application Load Balancers do not support cookie values that are URL encoded.

With CORS (cross-origin resource sharing) requests, some browsers require `SameSite=None; Secure` to enable stickiness. In this case, Elastic Load Balancing generates a second cookie, AWSALBTGCORS, which includes the same information as the original stickiness cookie plus this `SameSite` attribute. Clients receive both cookies.



Redirect actions

You can use `redirect` actions to redirect client requests from one URL to another. You can configure redirects as either temporary (HTTP 302) or permanent (HTTP 301) based on your needs.

A URI consists of the following components:

```
protocol://hostname:port/path?query
```

You must modify at least one of the following components to avoid a redirect loop: protocol, hostname, port, or path. Any components that you do not modify retain their original values.

protocol

The protocol (HTTP or HTTPS). You can redirect HTTP to HTTP, HTTP to HTTPS, and HTTPS to HTTPS. You cannot redirect HTTPS to HTTP.

hostname

The hostname. A hostname is not case-sensitive, can be up to 128 characters in length, and consists of alpha-numeric characters, wildcards (* and ?), and hyphens (-).

port

The port (1 to 65535).

path

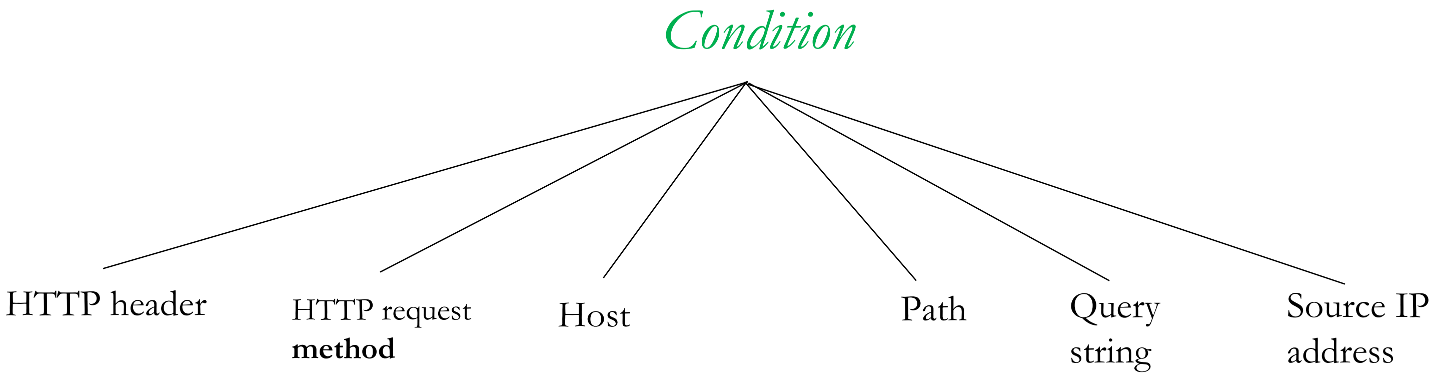
The absolute path, starting with the leading "/". A path is case-sensitive, can be up to 128 characters in length, and consists of alpha-numeric characters, wildcards (* and ?), & (using &amp;), and the following special characters: _-.$/~"'@:+.

query

The query parameters. The maximum length is 128 characters.

You can reuse URI components of the original URL in the target URL using the following reserved keywords:

In the above image, notice how AWS call a portion of the URL a hostname. Most people would call this portion the domain name.



Condition

HTTP header

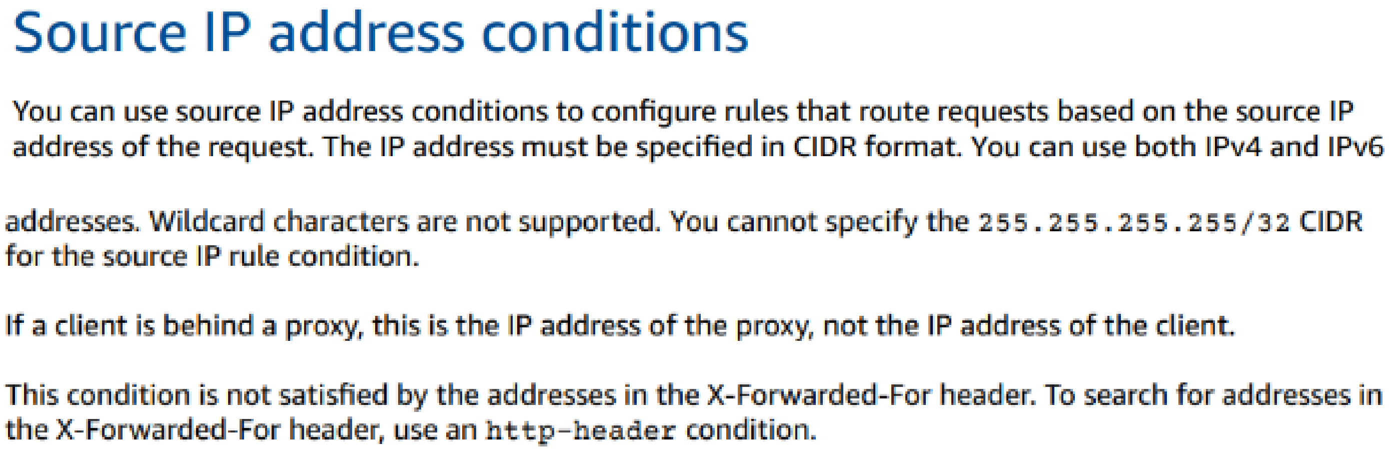
HTTP request **method**

Host

Path

Query string

Source IP address

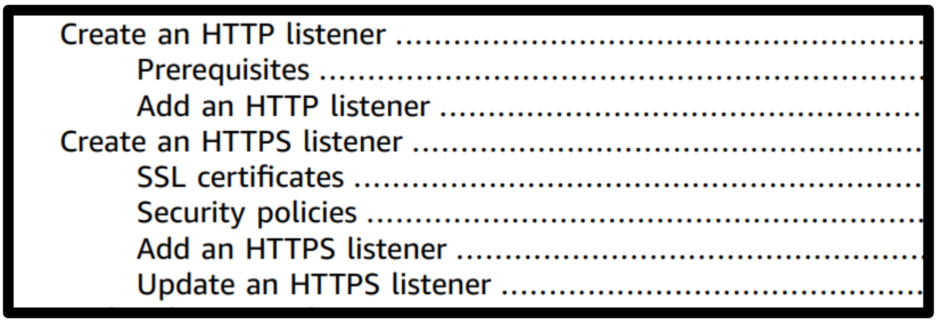


Source IP address conditions

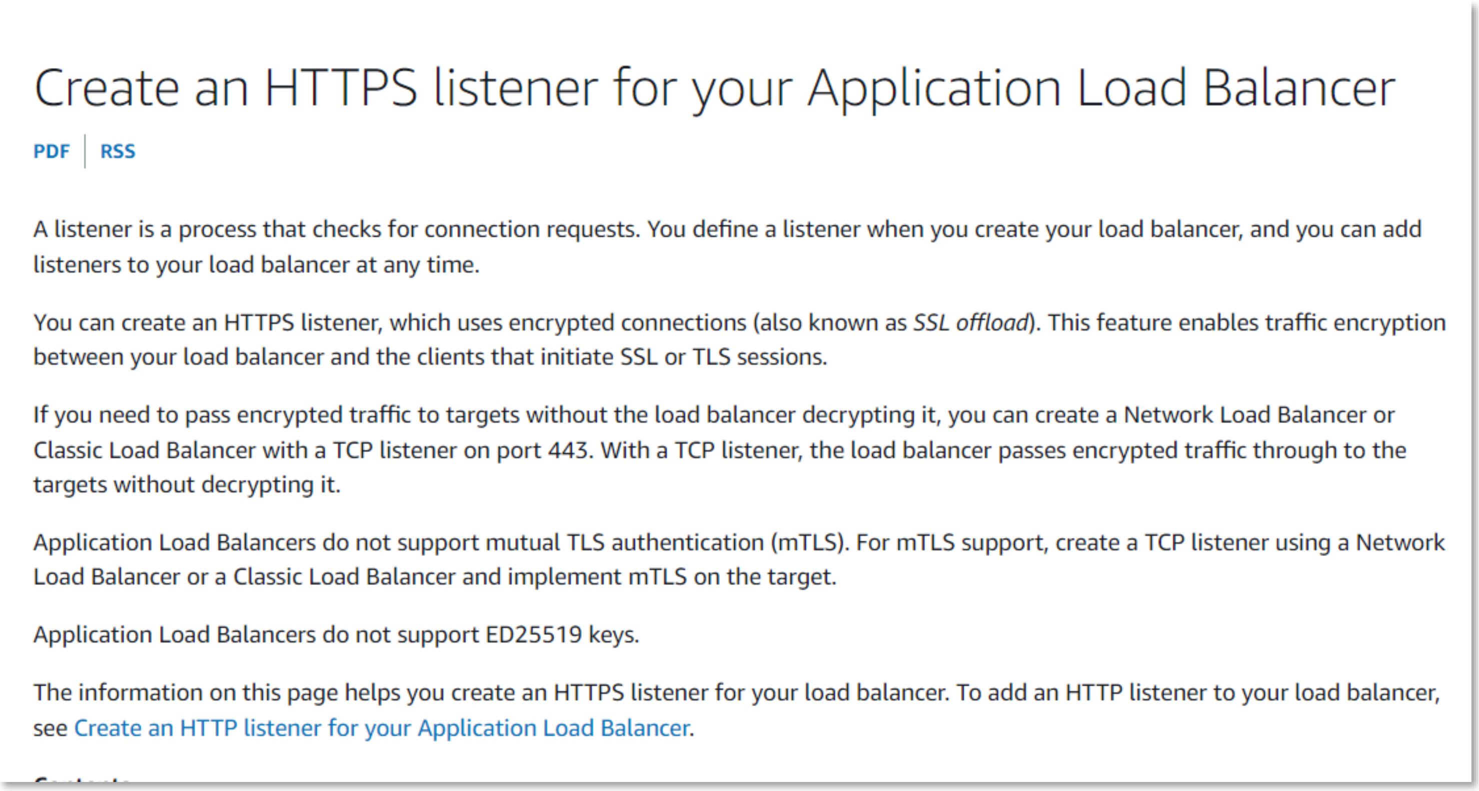
You can use source IP address conditions to configure rules that route requests based on the source IP address of the request. The IP address must be specified in CIDR format. You can use both IPv4 and IPv6 addresses. Wildcard characters are not supported. You cannot specify the `255.255.255.255/32` CIDR for the source IP rule condition.

If a client is behind a proxy, this is the IP address of the proxy, not the IP address of the client.

This condition is not satisfied by the addresses in the X-Forwarded-For header. To search for addresses in the X-Forwarded-For header, use an `http-header` condition.



Let’s talk about SSL certificates



Create an HTTPS listener for your Application Load Balancer

PDF | RSS

A listener is a process that checks for connection requests. You define a listener when you create your load balancer, and you can add listeners to your load balancer at any time.

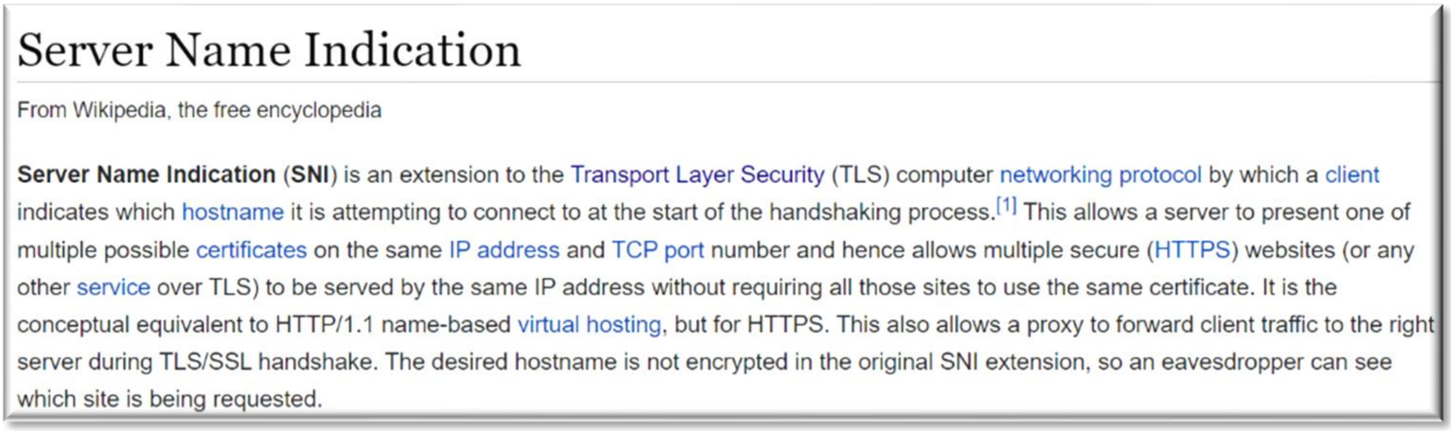
You can create an HTTPS listener, which uses encrypted connections (also known as *SSL offload*). This feature enables traffic encryption between your load balancer and the clients that initiate SSL or TLS sessions.

If you need to pass encrypted traffic to targets without the load balancer decrypting it, you can create a Network Load Balancer or Classic Load Balancer with a TCP listener on port 443. With a TCP listener, the load balancer passes encrypted traffic through to the targets without decrypting it.

Application Load Balancers do not support mutual TLS authentication (mTLS). For mTLS support, create a TCP listener using a Network Load Balancer or a Classic Load Balancer and implement mTLS on the target.

Application Load Balancers do not support ED25519 keys.

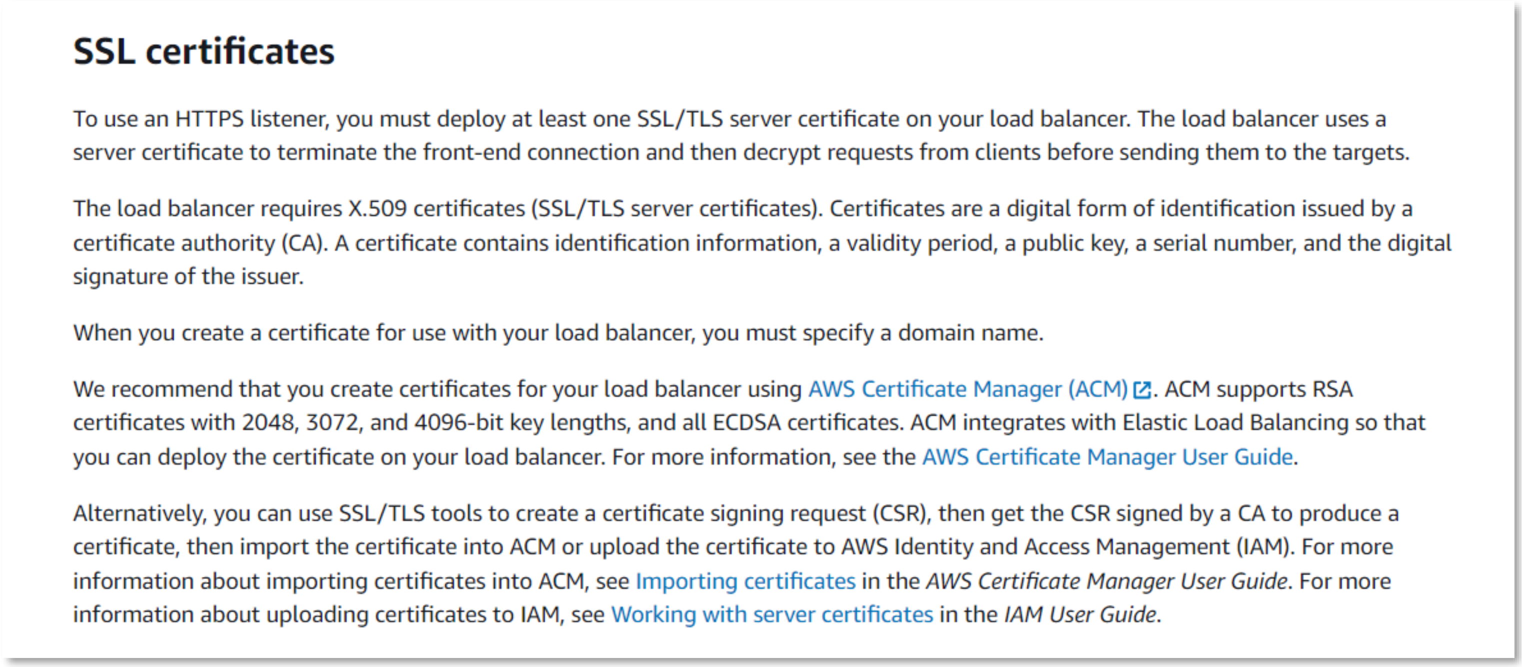
The information on this page helps you create an HTTPS listener for your load balancer. To add an HTTP listener to your load balancer, see Create an HTTP listener for your Application Load Balancer.



Server Name Indication

From Wikipedia, the free encyclopedia

Server Name Indication (**SNI**) is an extension to the Transport Layer Security (TLS) computer networking protocol by which a client indicates which hostname it is attempting to connect to at the start of the handshaking process.[1] This allows a server to present one of multiple possible certificates on the same IP address and TCP port number and hence allows multiple secure (HTTPS) websites (or any other service over TLS) to be served by the same IP address without requiring all those sites to use the same certificate. It is the conceptual equivalent to HTTP/1.1 name-based virtual hosting, but for HTTPS. This also allows a proxy to forward client traffic to the right server during TLS/SSL handshake. The desired hostname is not encrypted in the original SNI extension, so an eavesdropper can see which site is being requested.



SSL certificates

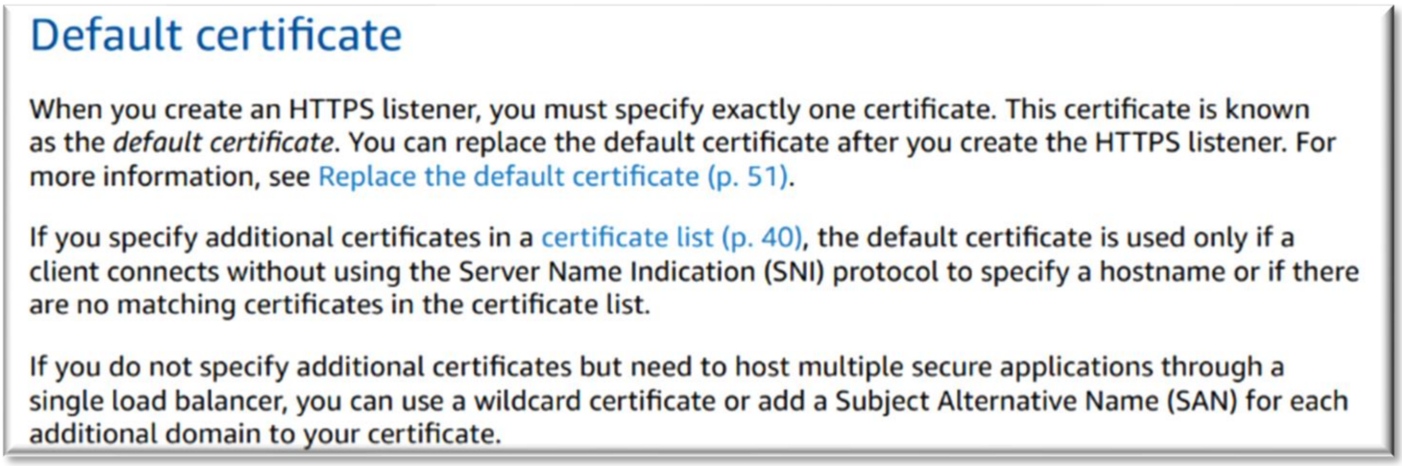
To use an HTTPS listener, you must deploy at least one SSL/TLS server certificate on your load balancer. The load balancer uses a server certificate to terminate the front-end connection and then decrypt requests from clients before sending them to the targets.

The load balancer requires X.509 certificates (SSL/TLS server certificates). Certificates are a digital form of identification issued by a certificate authority (CA). A certificate contains identification information, a validity period, a public key, a serial number, and the digital signature of the issuer.

When you create a certificate for use with your load balancer, you must specify a domain name.

We recommend that you create certificates for your load balancer using AWS Certificate Manager (ACM). ACM supports RSA certificates with 2048, 3072, and 4096-bit key lengths, and all ECDSA certificates. ACM integrates with Elastic Load Balancing so that you can deploy the certificate on your load balancer. For more information, see the AWS Certificate Manager User Guide.

Alternatively, you can use SSL/TLS tools to create a certificate signing request (CSR), then get the CSR signed by a CA to produce a certificate, then import the certificate into ACM or upload the certificate to AWS Identity and Access Management (IAM). For more information about importing certificates into ACM, see Importing certificates in the *AWS Certificate Manager User Guide*. For more information about uploading certificates to IAM, see Working with server certificates in the *IAM User Guide*.



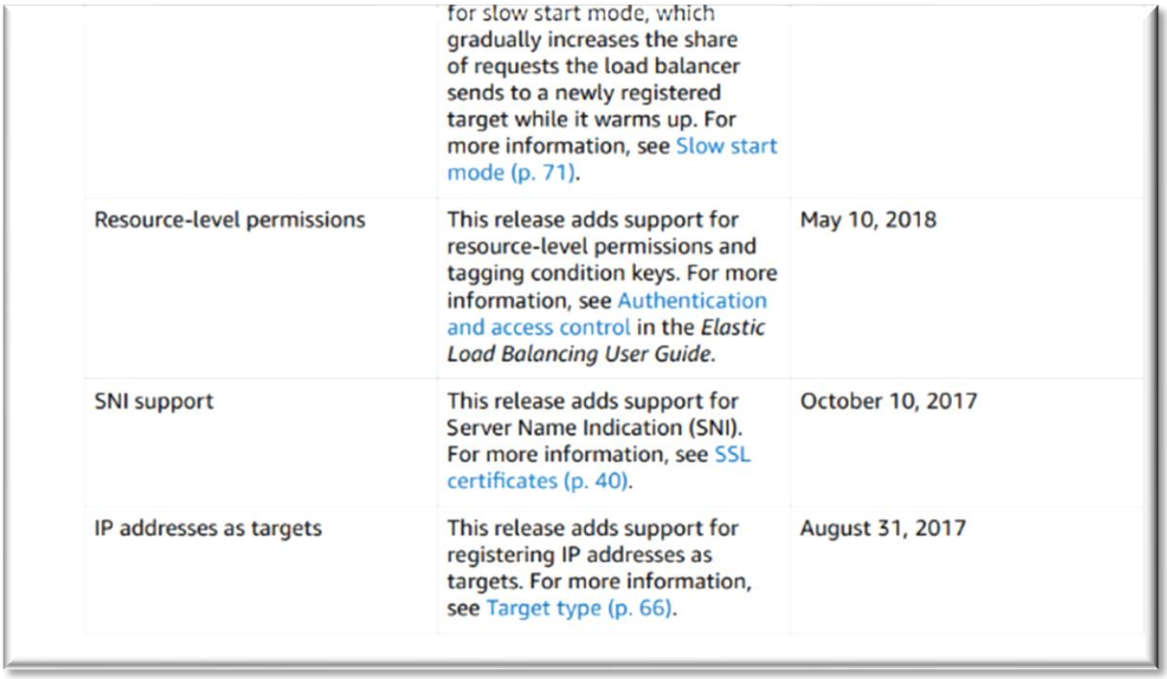
Default certificate

When you create an HTTPS listener, you must specify exactly one certificate. This certificate is known as the *default certificate*. You can replace the default certificate after you create the HTTPS listener. For more information, see Replace the default certificate (p. 51).

If you specify additional certificates in a certificate list (p. 40), the default certificate is used only if a client connects without using the Server Name Indication (SNI) protocol to specify a hostname or if there are no matching certificates in the certificate list.

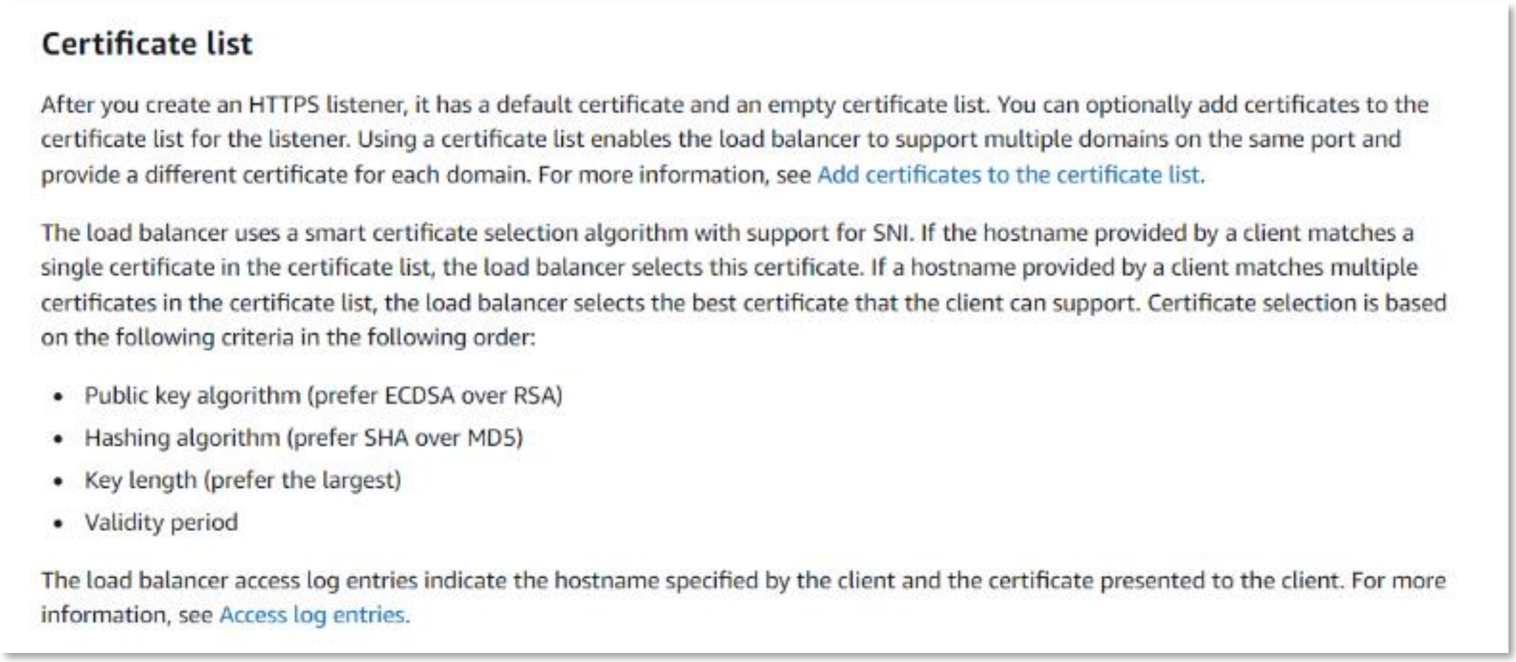
If you do not specify additional certificates but need to host multiple secure applications through a single load balancer, you can use a wildcard certificate or add a Subject Alternative Name (SAN) for each additional domain to your certificate.

Screenshot from the User Guide for the Application Load Balancer



	for slow start mode, which gradually increases the share of requests the load balancer sends to a newly registered target while it warms up. For more information, see Slow start mode (p. 71).	
Resource-level permissions	This release adds support for resource-level permissions and tagging condition keys. For more information, see Authentication and access control in the *Elastic Load Balancing User Guide*.	May 10, 2018
SNI support	This release adds support for Server Name Indication (SNI). For more information, see SSL certificates (p. 40).	October 10, 2017
IP addresses as targets	This release adds support for registering IP addresses as targets. For more information, see Target type (p. 66).	August 31, 2017

Screenshot from the bottom of the documentation for the ALB, showing that support for SNI (Server Name Indication) was added in 2017



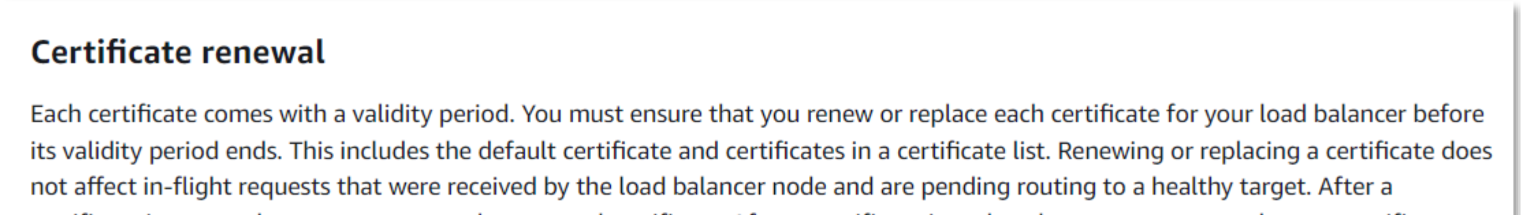
Certificate list

After you create an HTTPS listener, it has a default certificate and an empty certificate list. You can optionally add certificates to the certificate list for the listener. Using a certificate list enables the load balancer to support multiple domains on the same port and provide a different certificate for each domain. For more information, see Add certificates to the certificate list.

The load balancer uses a smart certificate selection algorithm with support for SNI. If the hostname provided by a client matches a single certificate in the certificate list, the load balancer selects this certificate. If a hostname provided by a client matches multiple certificates in the certificate list, the load balancer selects the best certificate that the client can support. Certificate selection is based on the following criteria in the following order:

- Public key algorithm (prefer ECDSA over RSA)
- Hashing algorithm (prefer SHA over MD5)
- Key length (prefer the largest)
- Validity period

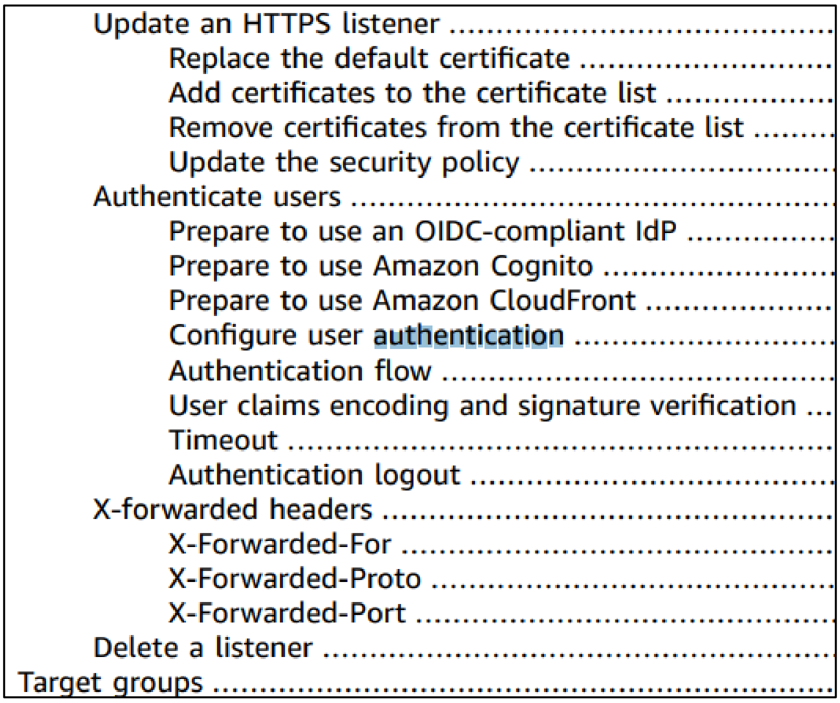
The load balancer access log entries indicate the hostname specified by the client and the certificate presented to the client. For more information, see Access log entries.



Certificate renewal

Each certificate comes with a validity period. You must ensure that you renew or replace each certificate for your load balancer before its validity period ends. This includes the default certificate and certificates in a certificate list. Renewing or replacing a certificate does not affect in-flight requests that were received by the load balancer node and are pending routing to a healthy target. After a

Let’s talk about authenticating users with the ALB

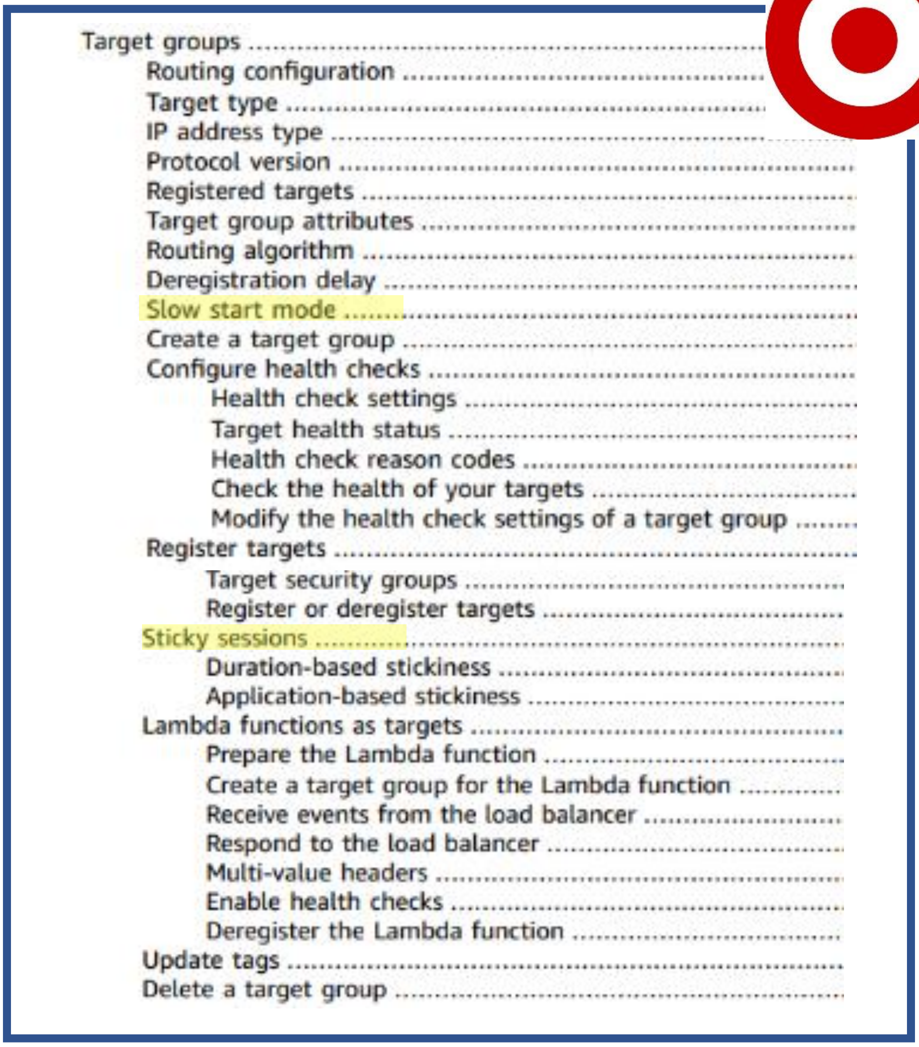


- Update an HTTPS listener
  - Replace the default certificate
  - Add certificates to the certificate list
  - Remove certificates from the certificate list
  - Update the security policy
- Authenticate users
  - Prepare to use an OIDC-compliant IdP
  - Prepare to use Amazon Cognito
  - Prepare to use Amazon CloudFront
  - Configure user authentication
  - Authentication flow
  - User claims encoding and signature verification
  - Timeout
  - Authentication logout
- X-forwarded headers
  - X-Forwarded-For
  - X-Forwarded-Proto
  - X-Forwarded-Port
- Delete a listener

Target groups

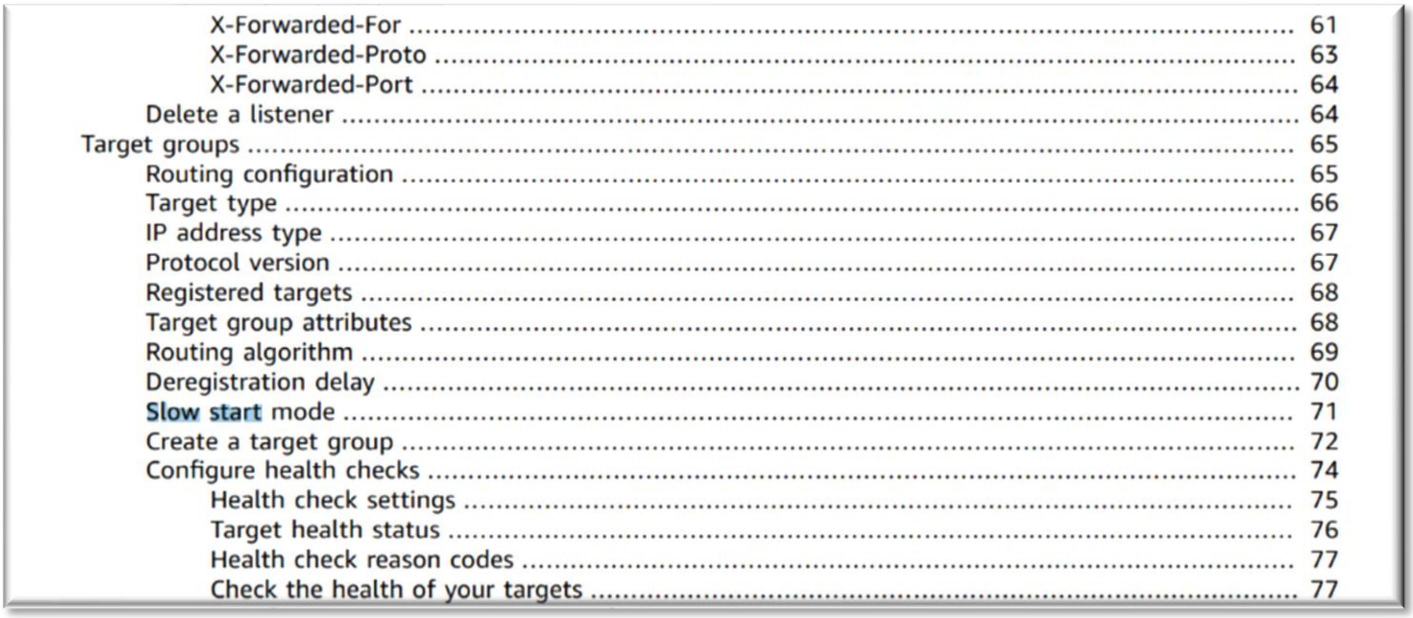
Screenshot from the documentation for the ALB. The section on authenticating users is within the fifth chapter, entitled “Listeners”.

Let’s talk about the target group…



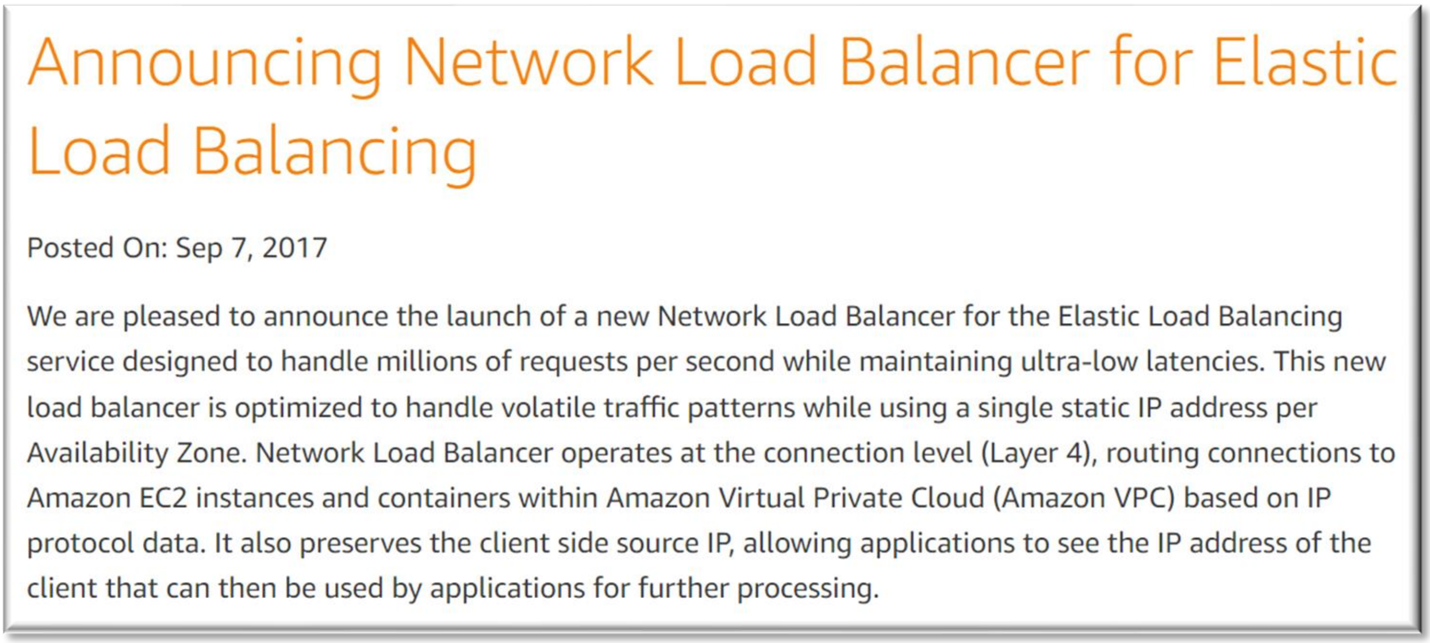
- Target groups
  - Routing configuration
  - Target type
  - IP address type
  - Protocol version
  - Registered targets
  - Target group attributes
  - Routing algorithm
  - Deregistration delay
  - Slow start mode
  - Create a target group
  - Configure health checks
    - Health check settings
    - Target health status
    - Health check reason codes
    - Check the health of your targets
    - Modify the health check settings of a target group
  - Register targets
    - Target security groups
    - Register or deregister targets
  - Sticky sessions
    - Duration-based stickiness
    - Application-based stickiness
  - Lambda functions as targets
    - Prepare the Lambda function
    - Create a target group for the Lambda function
    - Receive events from the load balancer
    - Respond to the load balancer
    - Multi-value headers
    - Enable health checks
    - Deregister the Lambda function
  - Update tags
  - Delete a target group

What on earth is “slow start mode”?



X-Forwarded-For	61
X-Forwarded-Proto	63
X-Forwarded-Port	64
Delete a listener	64
Target groups	65
Routing configuration	65
Target type	66
IP address type	67
Protocol version	67
Registered targets	68
Target group attributes	68
Routing algorithm	69
Deregistration delay	70
Slow start mode	71
Create a target group	72
Configure health checks	74
Health check settings	75
Target health status	76
Health check reason codes	77
Check the health of your targets	77

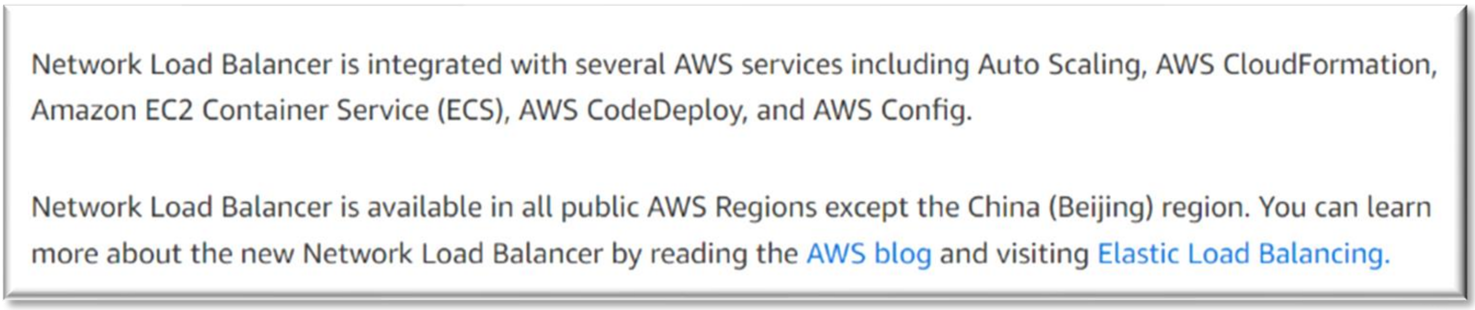
Screenshot from the documentation for the Application Load Balancer, showing (highlighted) slow start mode.



Announcing Network Load Balancer for Elastic Load Balancing

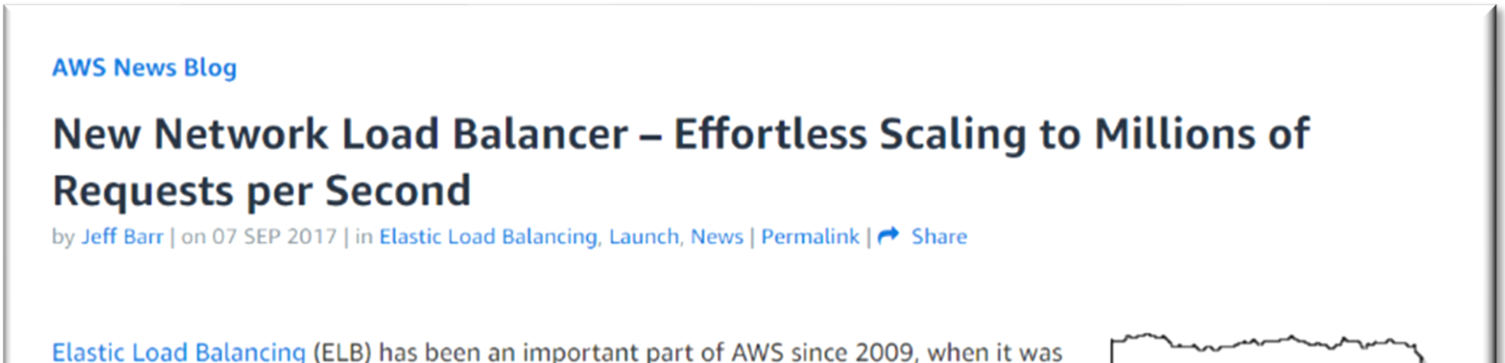
Posted On: Sep 7, 2017

We are pleased to announce the launch of a new Network Load Balancer for the Elastic Load Balancing service designed to handle millions of requests per second while maintaining ultra-low latencies. This new load balancer is optimized to handle volatile traffic patterns while using a single static IP address per Availability Zone. Network Load Balancer operates at the connection level (Layer 4), routing connections to Amazon EC2 instances and containers within Amazon Virtual Private Cloud (Amazon VPC) based on IP protocol data. It also preserves the client side source IP, allowing applications to see the IP address of the client that can then be used by applications for further processing.



Network Load Balancer is integrated with several AWS services including Auto Scaling, AWS CloudFormation, Amazon EC2 Container Service (ECS), AWS CodeDeploy, and AWS Config.

Network Load Balancer is available in all public AWS Regions except the China (Beijing) region. You can learn more about the new Network Load Balancer by reading the AWS blog and visiting Elastic Load Balancing.

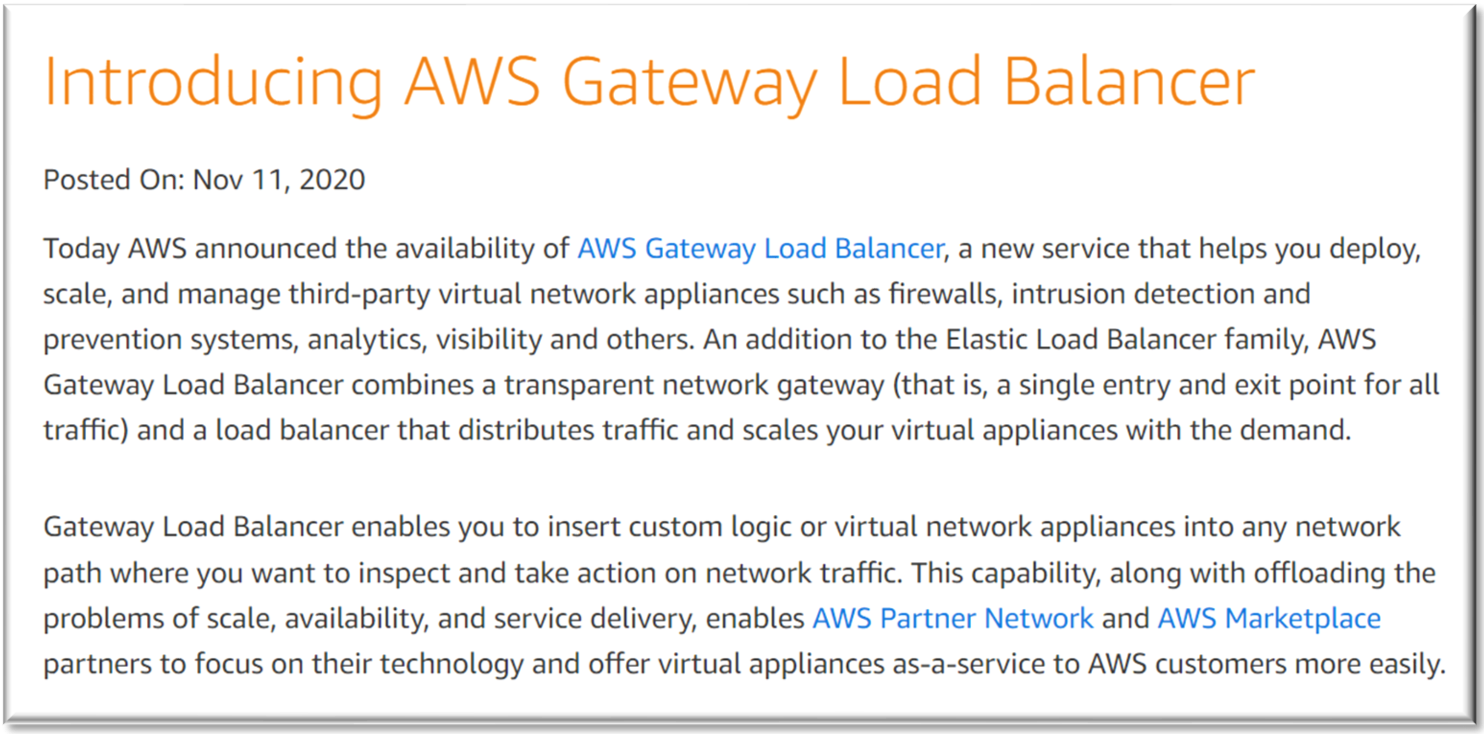


AWS News Blog

New Network Load Balancer – Effortless Scaling to Millions of Requests per Second

by Jeff Barr | on 07 SEP 2017 | in Elastic Load Balancing, Launch, News | Permalink | Share

Elastic Load Balancing (ELB) has been an important part of AWS since 2009, when it was



Introducing AWS Gateway Load Balancer

Posted On: Nov 11, 2020

Today AWS announced the availability of AWS Gateway Load Balancer, a new service that helps you deploy, scale, and manage third-party virtual network appliances such as firewalls, intrusion detection and prevention systems, analytics, visibility and others. An addition to the Elastic Load Balancer family, AWS Gateway Load Balancer combines a transparent network gateway (that is, a single entry and exit point for all traffic) and a load balancer that distributes traffic and scales your virtual appliances with the demand.

Gateway Load Balancer enables you to insert custom logic or virtual network appliances into any network path where you want to inspect and take action on network traffic. This capability, along with offloading the problems of scale, availability, and service delivery, enables AWS Partner Network and AWS Marketplace partners to focus on their technology and offer virtual appliances as-a-service to AWS customers more easily.

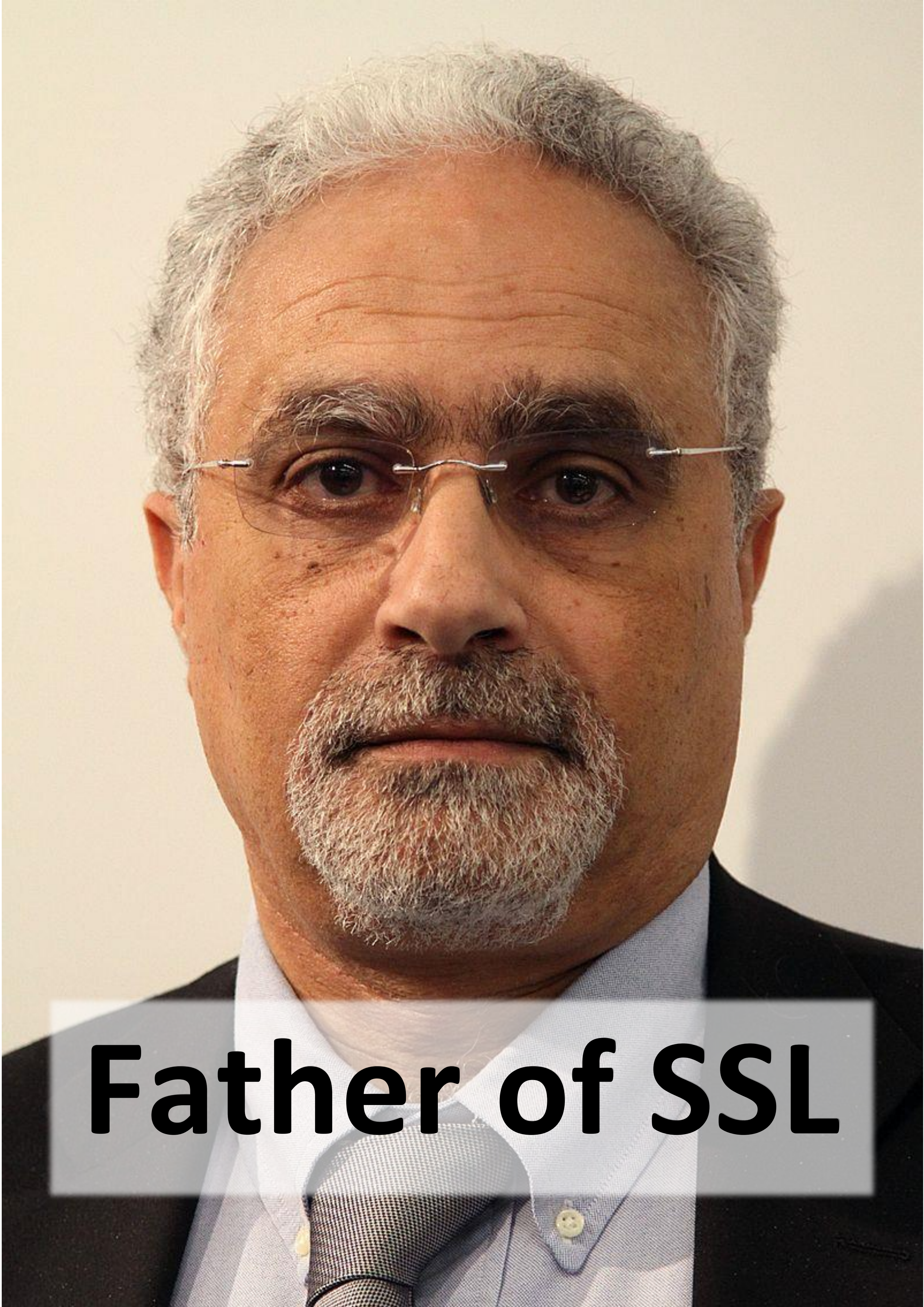
Michael Carter

Ten facts about SSL

1. SSL stands for “Secure Sockets Layers”. A number of SSL specifications were set out

in the 1990s, by a company called Netscape.

2. Taher Elgamal has been described as the “father of SSL”. He was the chief scientist as Netscape Communications for three years (1995-1998).



Father of SSL

3. SLL version 1.0 was never released

It is in fact true that version 1.0 of SSL was never released because it had serious security flaws. Instead, version 2.0 of SSL was released in February of 1995.

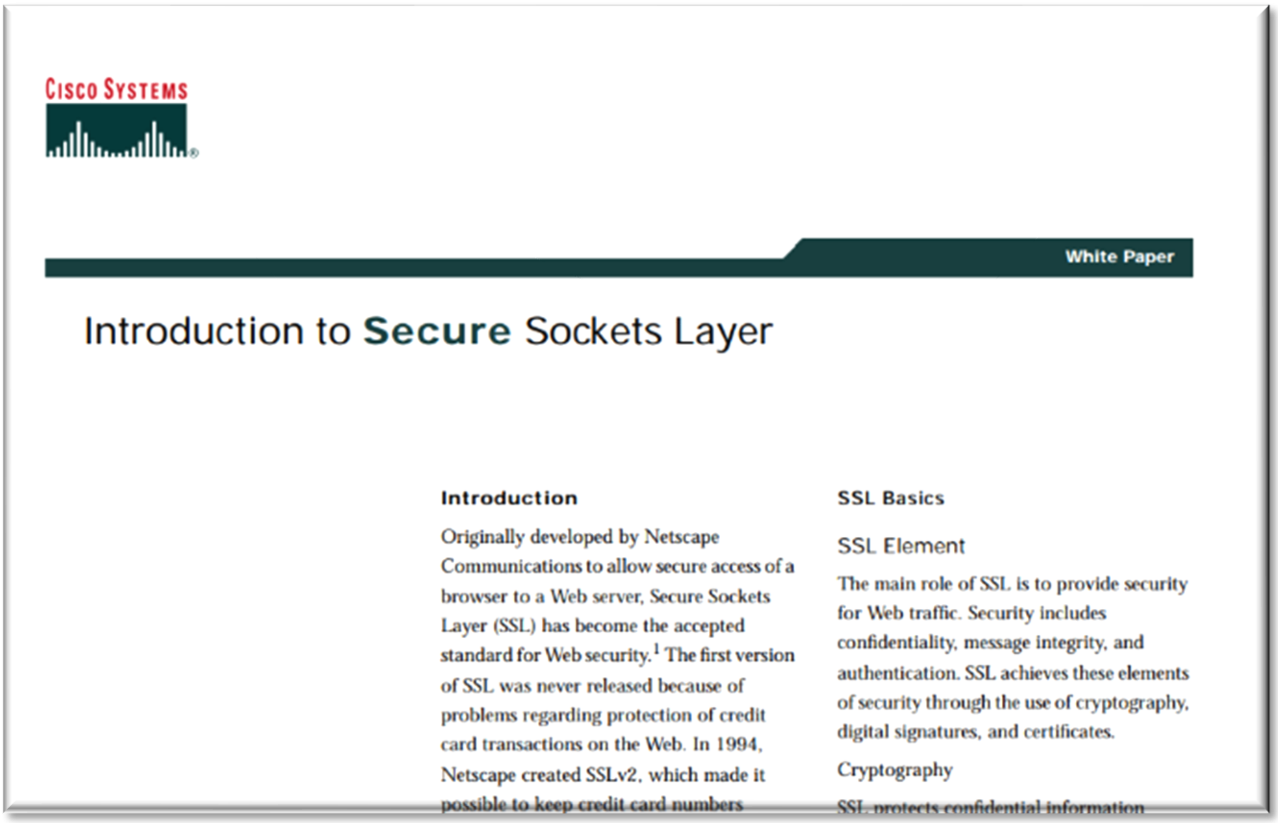
4. SSL 2.0 has been deprecated

In 2011, the decision was made to deprecate SSL 2.0. This is set out in RFC 6176.

Further reading on SSL

The complete guide to SSL Certificates

Brave Little Tank



Cisco Systems

White Paper

Introduction to Secure Sockets Layer

Introduction

Originally developed by Netscape Communications to allow secure access of a browser to a Web server, Secure Sockets Layer (SSL) has become the accepted standard for Web security.[1] The first version of SSL was never released because of problems regarding protection of credit card transactions on the Web. In 1994, Netscape created SSLv2, which made it possible to keep credit card numbers

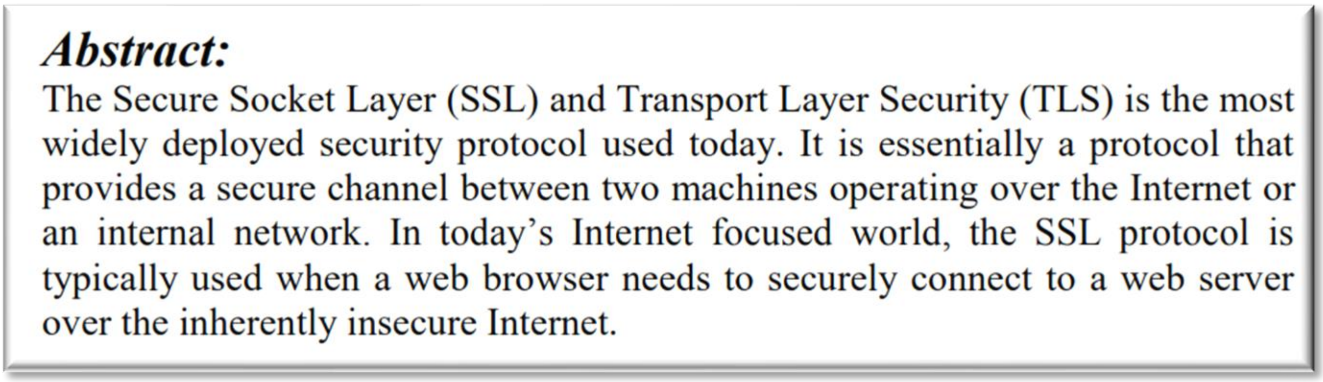
SSL Basics

SSL Element

The main role of SSL is to provide security for Web traffic. Security includes confidentiality, message integrity, and authentication. SSL achieves these elements of security through the use of cryptography, digital signatures, and certificates.

Cryptography

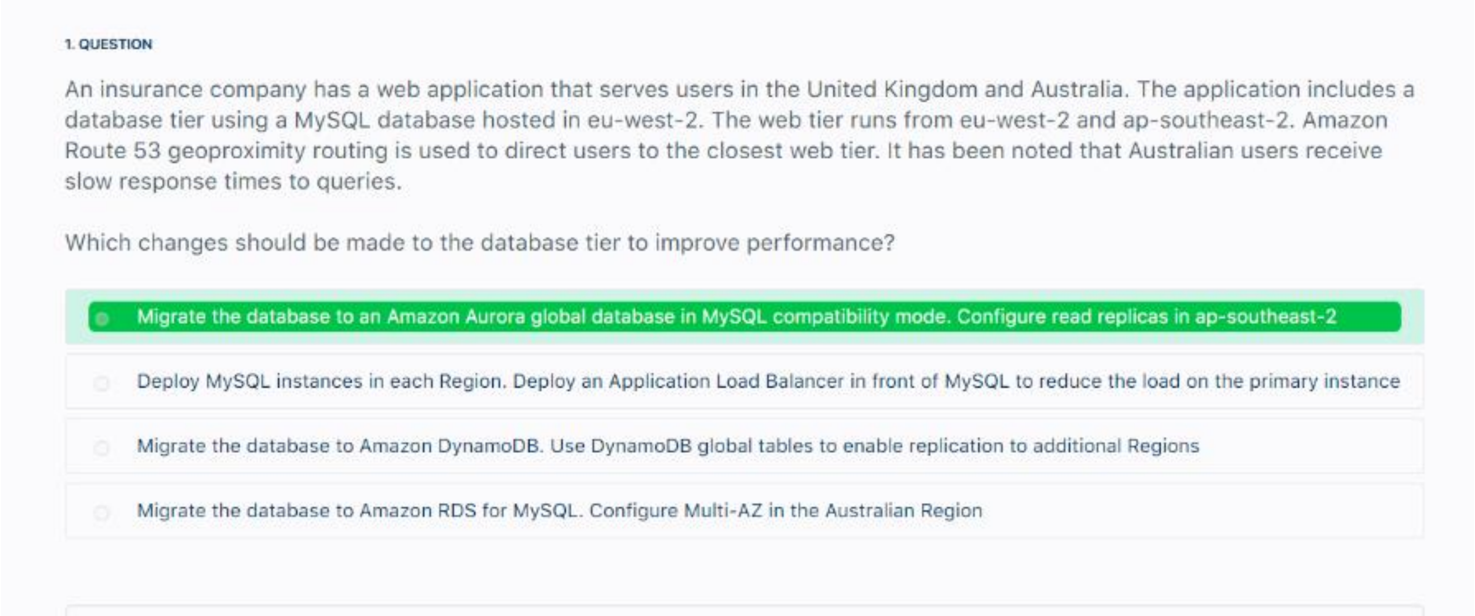
SSL protects confidential information



Abstract:

The Secure Socket Layer (SSL) and Transport Layer Security (TLS) is the most widely deployed security protocol used today. It is essentially a protocol that provides a secure channel between two machines operating over the Internet or an internal network. In today's Internet focused world, the SSL protocol is typically used when a web browser needs to securely connect to a web server over the inherently insecure Internet.

DOI:10.13140/RG.2.1.2671.3044



1. QUESTION

An insurance company has a web application that serves users in the United Kingdom and Australia. The application includes a database tier using a MySQL database hosted in eu-west-2. The web tier runs from eu-west-2 and ap-southeast-2. Amazon Route 53 geoproximity routing is used to direct users to the closest web tier. It has been noted that Australian users receive slow response times to queries.

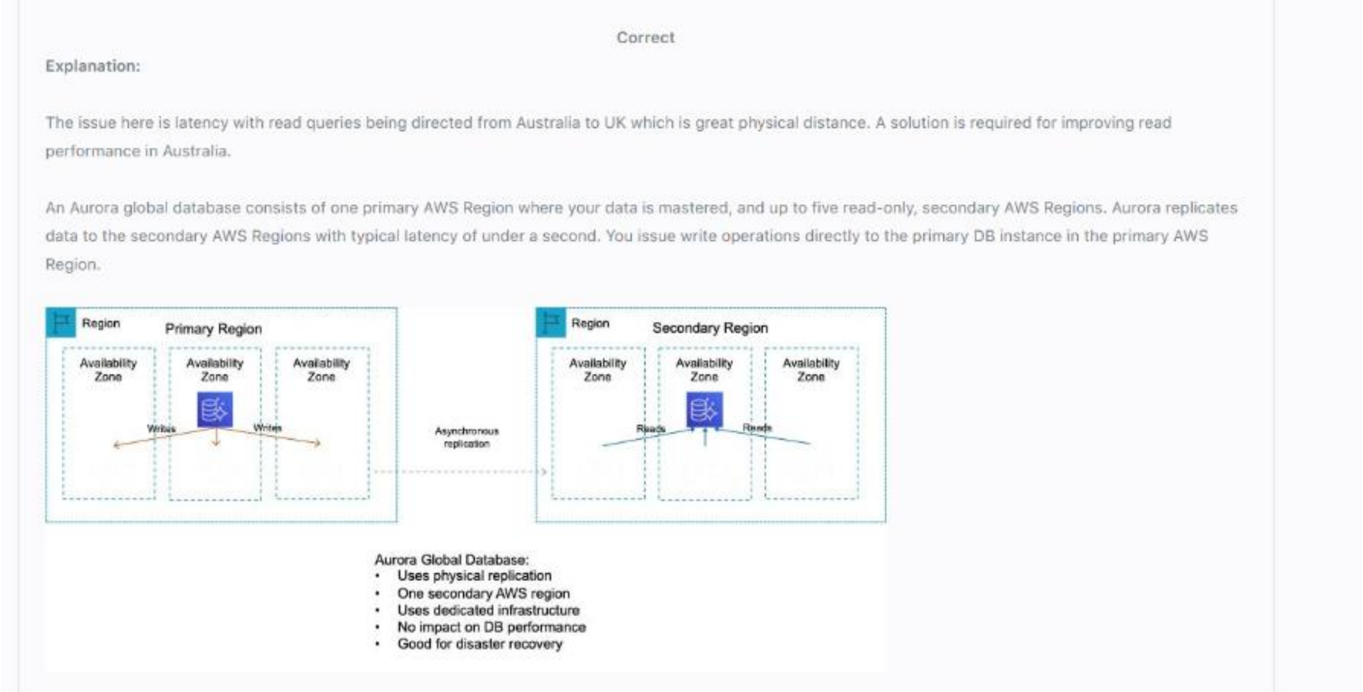
Which changes should be made to the database tier to improve performance?

● Migrate the database to an Amazon Aurora global database in MySQL compatibility mode. Configure read replicas in ap-southeast-2

○ Deploy MySQL instances in each Region. Deploy an Application Load Balancer in front of MySQL to reduce the load on the primary instance

○ Migrate the database to Amazon DynamoDB. Use DynamoDB global tables to enable replication to additional Regions

○ Migrate the database to Amazon RDS for MySQL. Configure Multi-AZ in the Australian Region



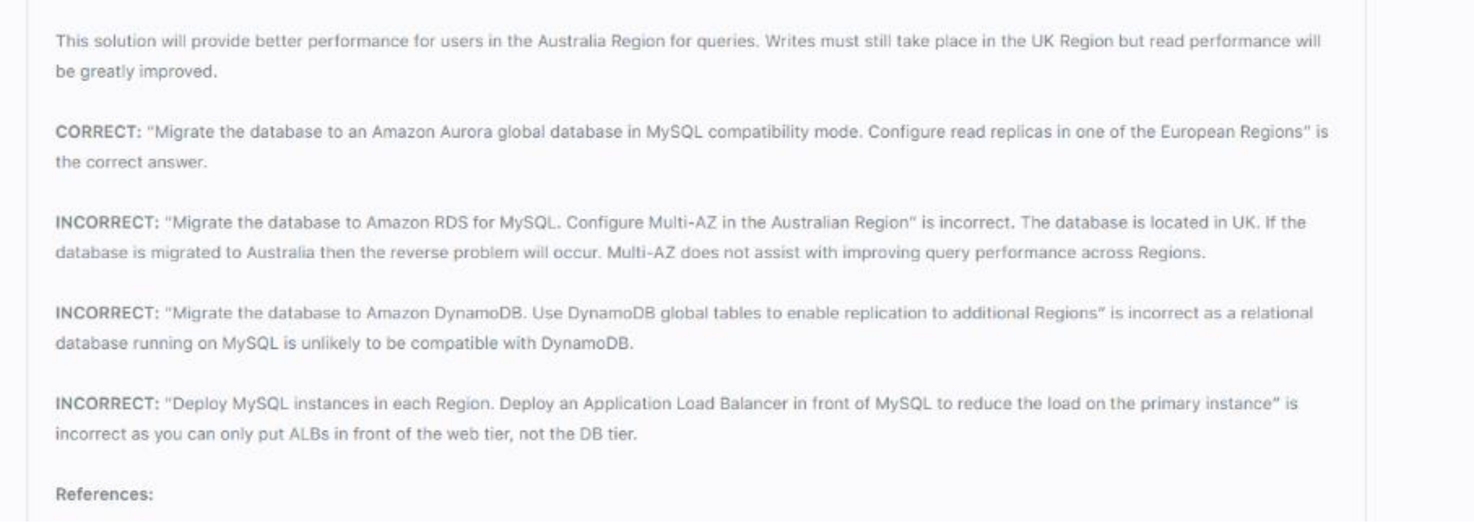
Correct

Explanation:

The issue here is latency with read queries being directed from Australia to UK which is great physical distance. A solution is required for improving read performance in Australia.

An Aurora global database consists of one primary AWS Region where your data is mastered, and up to five read-only, secondary AWS Regions. Aurora replicates data to the secondary AWS Regions with typical latency of under a second. You issue write operations directly to the primary DB instance in the primary AWS Region.

Aurora Global Database:
- Uses physical replication
- One secondary AWS region
- Uses dedicated infrastructure
- No impact on DB performance
- Good for disaster recovery



This solution will provide better performance for users in the Australia Region for queries. Writes must still take place in the UK Region but read performance will be greatly improved.

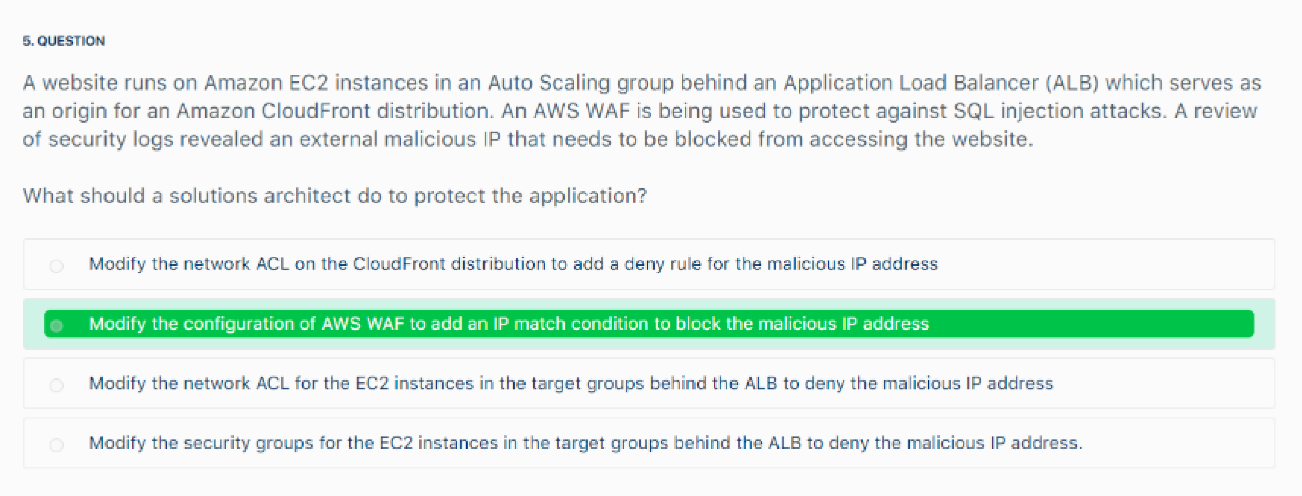
CORRECT: "Migrate the database to an Amazon Aurora global database in MySQL compatibility mode. Configure read replicas in one of the European Regions" is the correct answer.

INCORRECT: "Migrate the database to Amazon RDS for MySQL. Configure Multi-AZ in the Australian Region" is incorrect. The database is located in UK. If the database is migrated to Australia then the reverse problem will occur. Multi-AZ does not assist with improving query performance across Regions.

INCORRECT: "Migrate the database to Amazon DynamoDB. Use DynamoDB global tables to enable replication to additional Regions" is incorrect as a relational database running on MySQL is unlikely to be compatible with DynamoDB.

INCORRECT: "Deploy MySQL instances in each Region. Deploy an Application Load Balancer in front of MySQL to reduce the load on the primary instance" is incorrect as you can only put ALBs in front of the web tier, not the DB tier.

References:



5. QUESTION

A website runs on Amazon EC2 instances in an Auto Scaling group behind an Application Load Balancer (ALB) which serves as an origin for an Amazon CloudFront distribution. An AWS WAF is being used to protect against SQL injection attacks. A review of security logs revealed an external malicious IP that needs to be blocked from accessing the website.

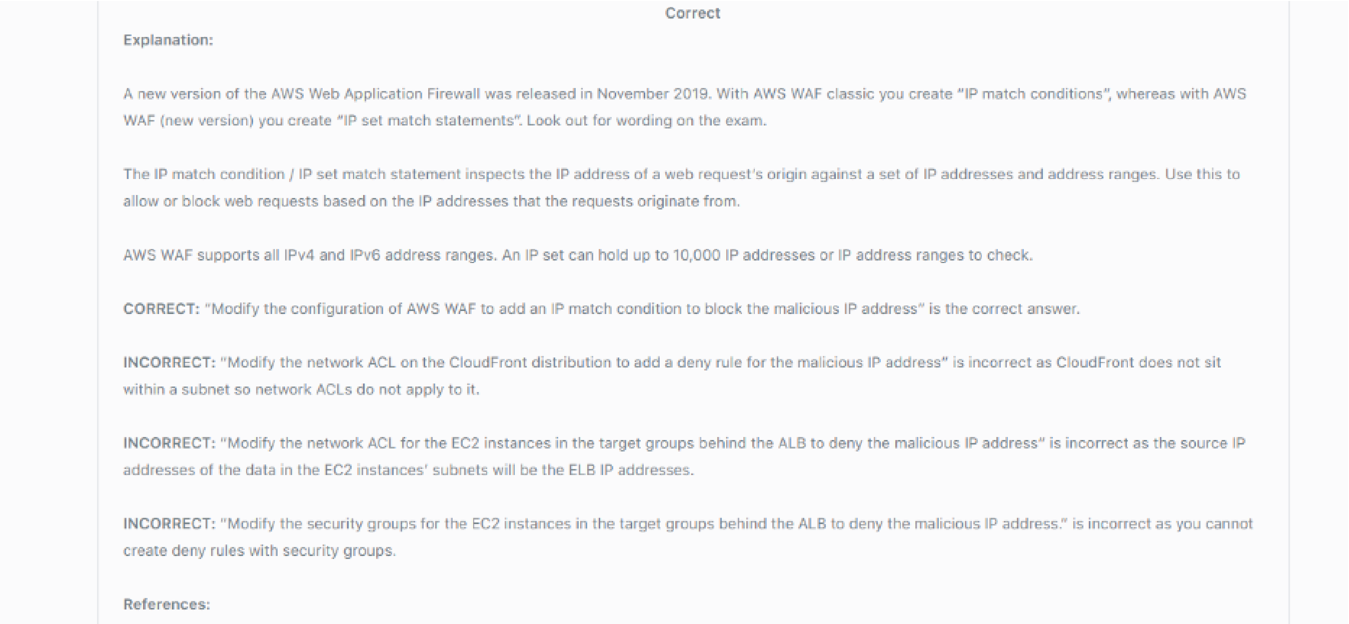
What should a solutions architect do to protect the application?

○ Modify the network ACL on the CloudFront distribution to add a deny rule for the malicious IP address

● Modify the configuration of AWS WAF to add an IP match condition to block the malicious IP address

○ Modify the network ACL for the EC2 instances in the target groups behind the ALB to deny the malicious IP address

○ Modify the security groups for the EC2 instances in the target groups behind the ALB to deny the malicious IP address.



Correct

Explanation:

A new version of the AWS Web Application Firewall was released in November 2019. With AWS WAF classic you create "IP match conditions", whereas with AWS WAF (new version) you create "IP set match statements". Look out for wording on the exam.

The IP match condition / IP set match statement inspects the IP address of a web request's origin against a set of IP addresses and address ranges. Use this to allow or block web requests based on the IP addresses that the requests originate from.

AWS WAF supports all IPv4 and IPv6 address ranges. An IP set can hold up to 10,000 IP addresses or IP address ranges to check.

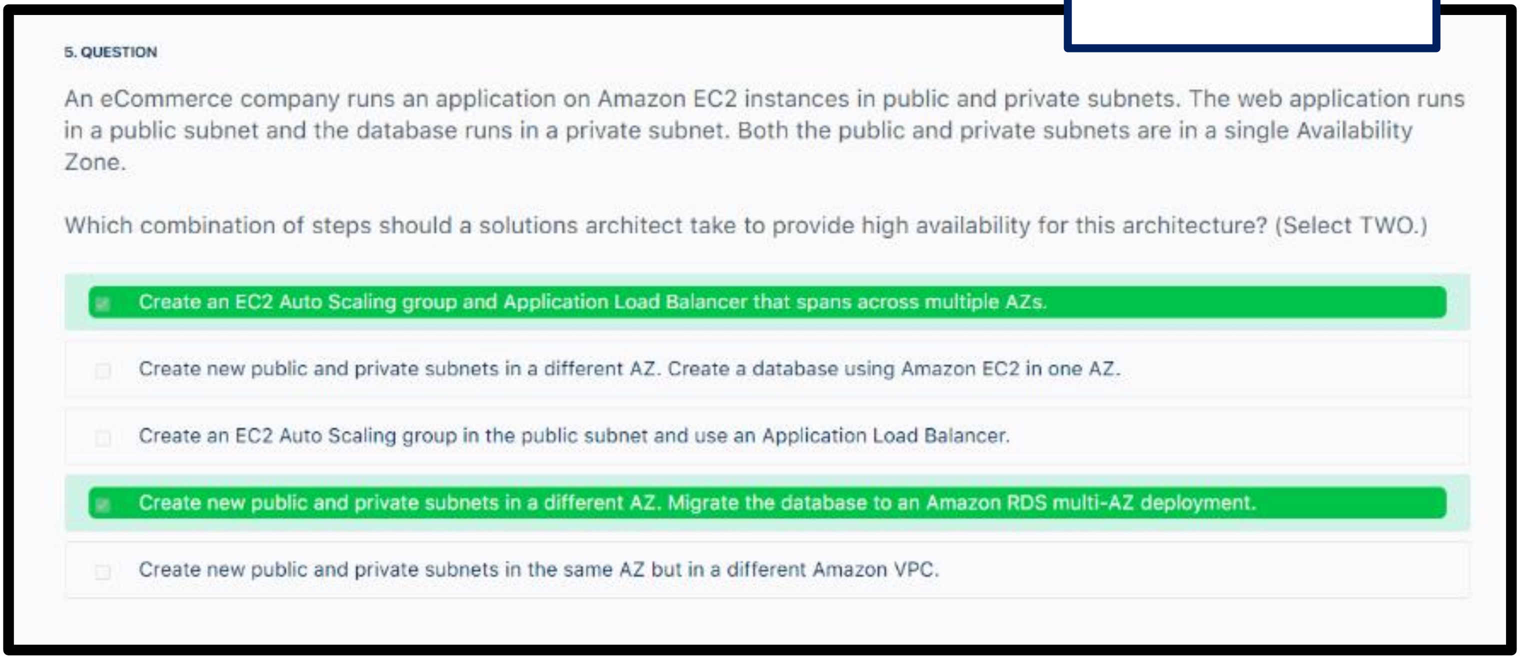
CORRECT: "Modify the configuration of AWS WAF to add an IP match condition to block the malicious IP address" is the correct answer.

INCORRECT: "Modify the network ACL on the CloudFront distribution to add a deny rule for the malicious IP address" is incorrect as CloudFront does not sit within a subnet so network ACLs do not apply to it.

INCORRECT: "Modify the network ACL for the EC2 instances in the target groups behind the ALB to deny the malicious IP address" is incorrect as the source IP addresses of the data in the EC2 instances' subnets will be the ELB IP addresses.

INCORRECT: "Modify the security groups for the EC2 instances in the target groups behind the ALB to deny the malicious IP address." is incorrect as you cannot create deny rules with security groups.

References:



5. QUESTION

An eCommerce company runs an application on Amazon EC2 instances in public and private subnets. The web application runs in a public subnet and the database runs in a private subnet. Both the public and private subnets are in a single Availability Zone.

Which combination of steps should a solutions architect take to provide high availability for this architecture? (Select TWO.)

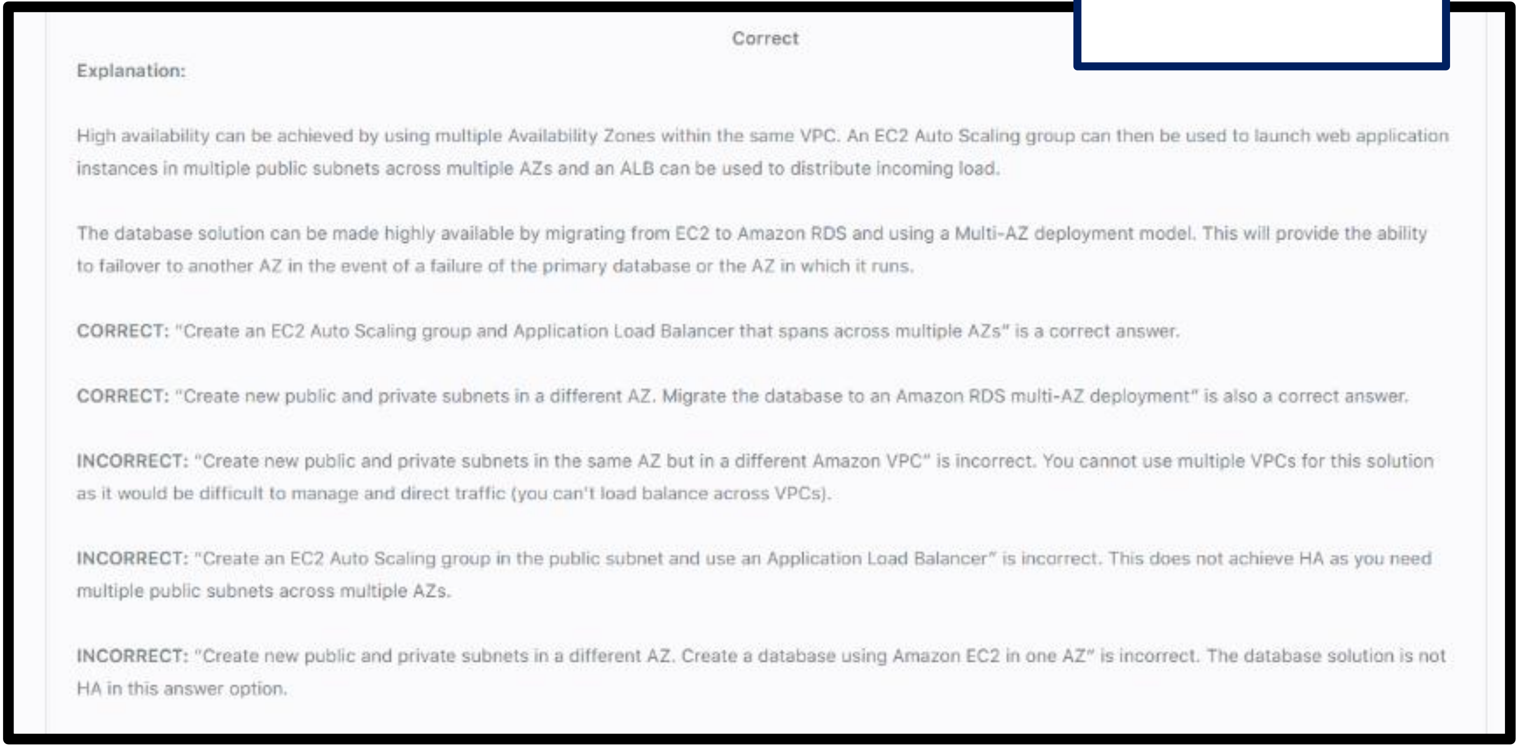
☑ Create an EC2 Auto Scaling group and Application Load Balancer that spans across multiple AZs.

☐ Create new public and private subnets in a different AZ. Create a database using Amazon EC2 in one AZ.

☐ Create an EC2 Auto Scaling group in the public subnet and use an Application Load Balancer.

☑ Create new public and private subnets in a different AZ. Migrate the database to an Amazon RDS multi-AZ deployment.

☐ Create new public and private subnets in the same AZ but in a different Amazon VPC.



Correct

Explanation:

High availability can be achieved by using multiple Availability Zones within the same VPC. An EC2 Auto Scaling group can then be used to launch web application instances in multiple public subnets across multiple AZs and an ALB can be used to distribute incoming load.

The database solution can be made highly available by migrating from EC2 to Amazon RDS and using a Multi-AZ deployment model. This will provide the ability to failover to another AZ in the event of a failure of the primary database or the AZ in which it runs.

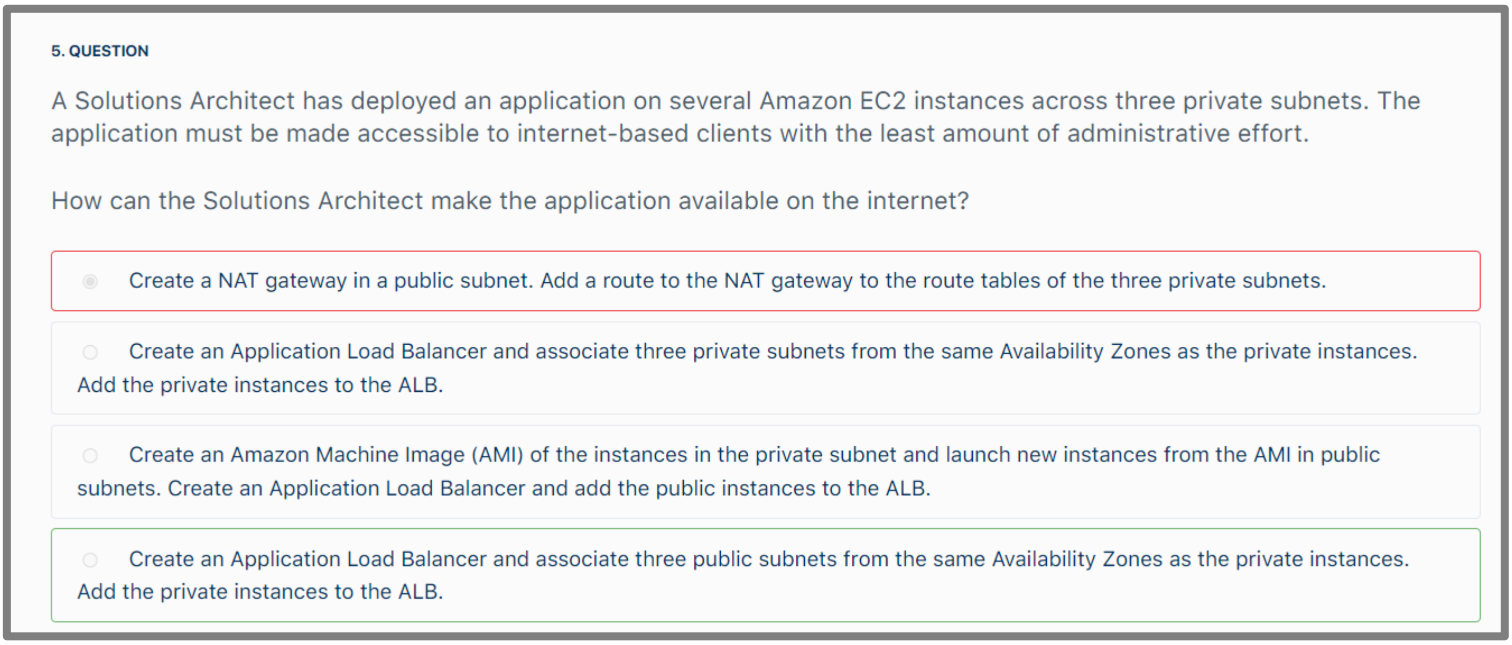
CORRECT: "Create an EC2 Auto Scaling group and Application Load Balancer that spans across multiple AZs" is a correct answer.

CORRECT: "Create new public and private subnets in a different AZ. Migrate the database to an Amazon RDS multi-AZ deployment" is also a correct answer.

INCORRECT: "Create new public and private subnets in the same AZ but in a different Amazon VPC" is incorrect. You cannot use multiple VPCs for this solution as it would be difficult to manage and direct traffic (you can't load balance across VPCs).

INCORRECT: "Create an EC2 Auto Scaling group in the public subnet and use an Application Load Balancer" is incorrect. This does not achieve HA as you need multiple public subnets across multiple AZs.

INCORRECT: "Create new public and private subnets in a different AZ. Create a database using Amazon EC2 in one AZ" is incorrect. The database solution is not HA in this answer option.



5. QUESTION

A Solutions Architect has deployed an application on several Amazon EC2 instances across three private subnets. The application must be made accessible to internet-based clients with the least amount of administrative effort.

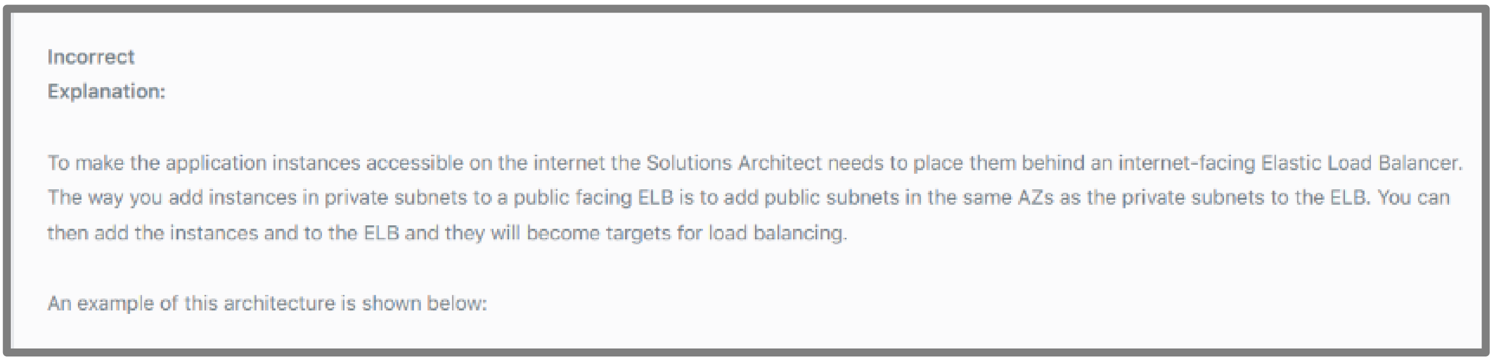
How can the Solutions Architect make the application available on the internet?

● Create a NAT gateway in a public subnet. Add a route to the NAT gateway to the route tables of the three private subnets.

○ Create an Application Load Balancer and associate three private subnets from the same Availability Zones as the private instances. Add the private instances to the ALB.

○ Create an Amazon Machine Image (AMI) of the instances in the private subnet and launch new instances from the AMI in public subnets. Create an Application Load Balancer and add the public instances to the ALB.

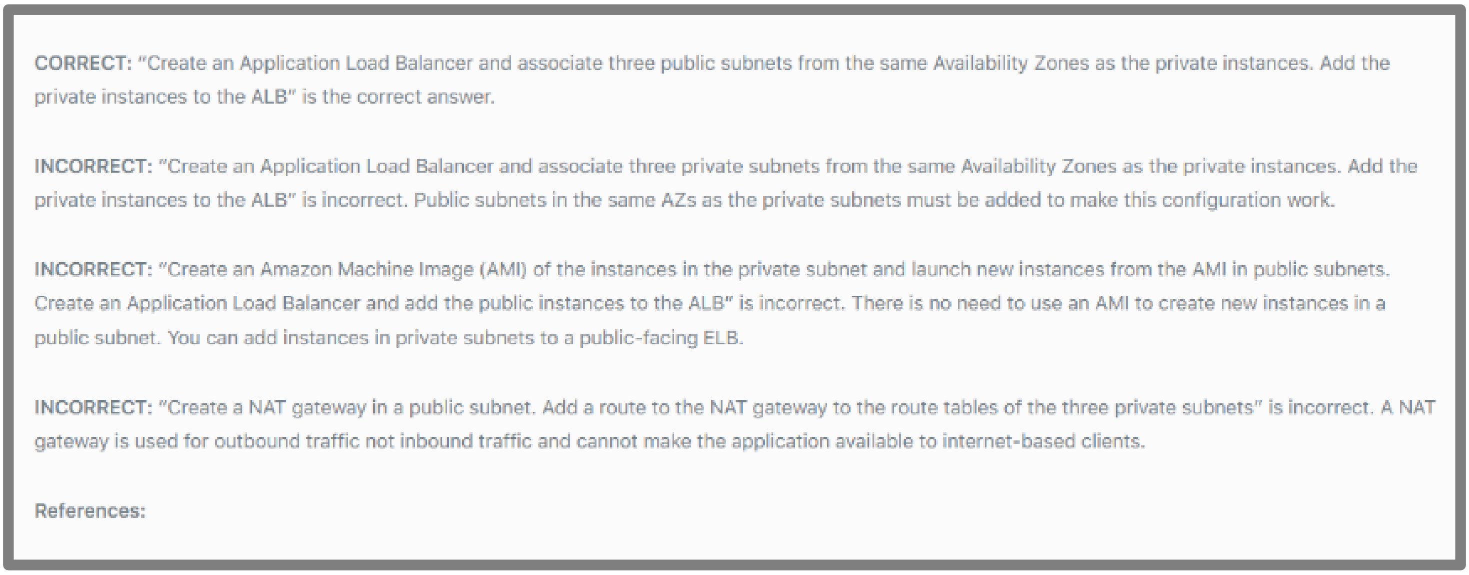
○ Create an Application Load Balancer and associate three public subnets from the same Availability Zones as the private instances. Add the private instances to the ALB.



Incorrect
Explanation:

To make the application instances accessible on the internet the Solutions Architect needs to place them behind an internet-facing Elastic Load Balancer. The way you add instances in private subnets to a public facing ELB is to add public subnets in the same AZs as the private subnets to the ELB. You can then add the instances and to the ELB and they will become targets for load balancing.

An example of this architecture is shown below:



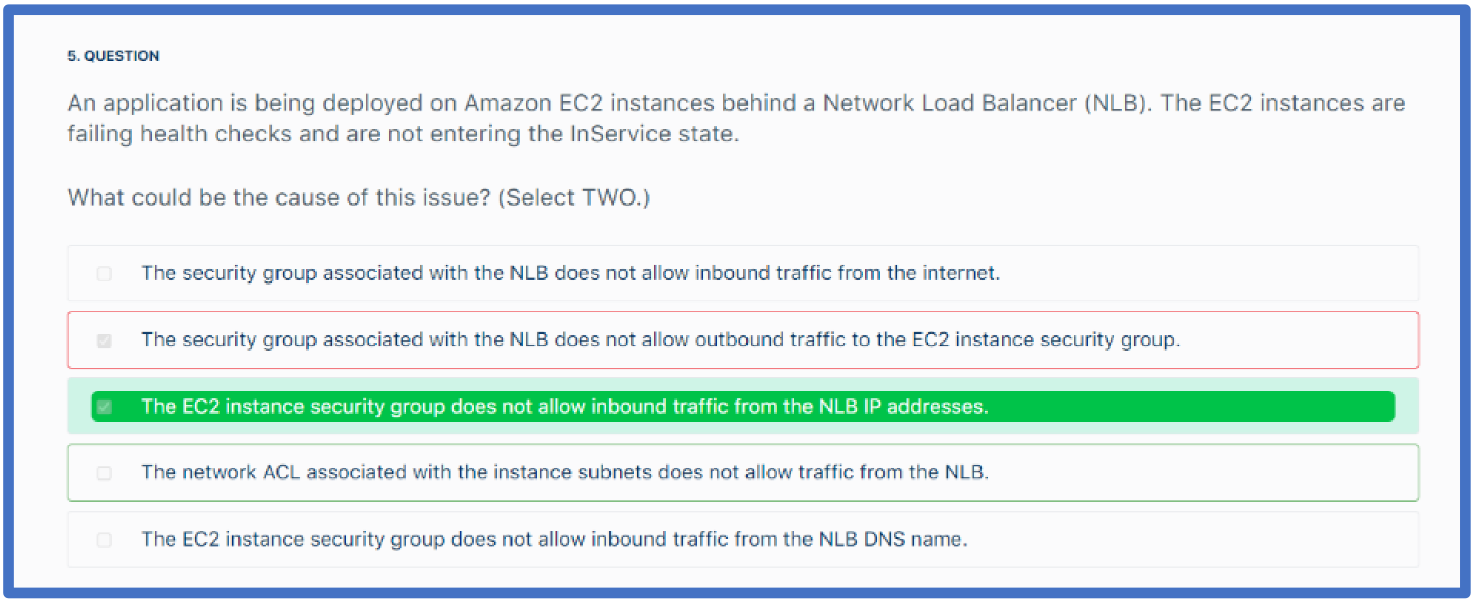
CORRECT: "Create an Application Load Balancer and associate three public subnets from the same Availability Zones as the private instances. Add the private instances to the ALB" is the correct answer.

INCORRECT: "Create an Application Load Balancer and associate three private subnets from the same Availability Zones as the private instances. Add the private instances to the ALB" is incorrect. Public subnets in the same AZs as the private subnets must be added to make this configuration work.

INCORRECT: "Create an Amazon Machine Image (AMI) of the instances in the private subnet and launch new instances from the AMI in public subnets. Create an Application Load Balancer and add the public instances to the ALB" is incorrect. There is no need to use an AMI to create new instances in a public subnet. You can add instances in private subnets to a public-facing ELB.

INCORRECT: "Create a NAT gateway in a public subnet. Add a route to the NAT gateway to the route tables of the three private subnets" is incorrect. A NAT gateway is used for outbound traffic not inbound traffic and cannot make the application available to internet-based clients.

References:



5. QUESTION

An application is being deployed on Amazon EC2 instances behind a Network Load Balancer (NLB). The EC2 instances are failing health checks and are not entering the InService state.

What could be the cause of this issue? (Select TWO.)

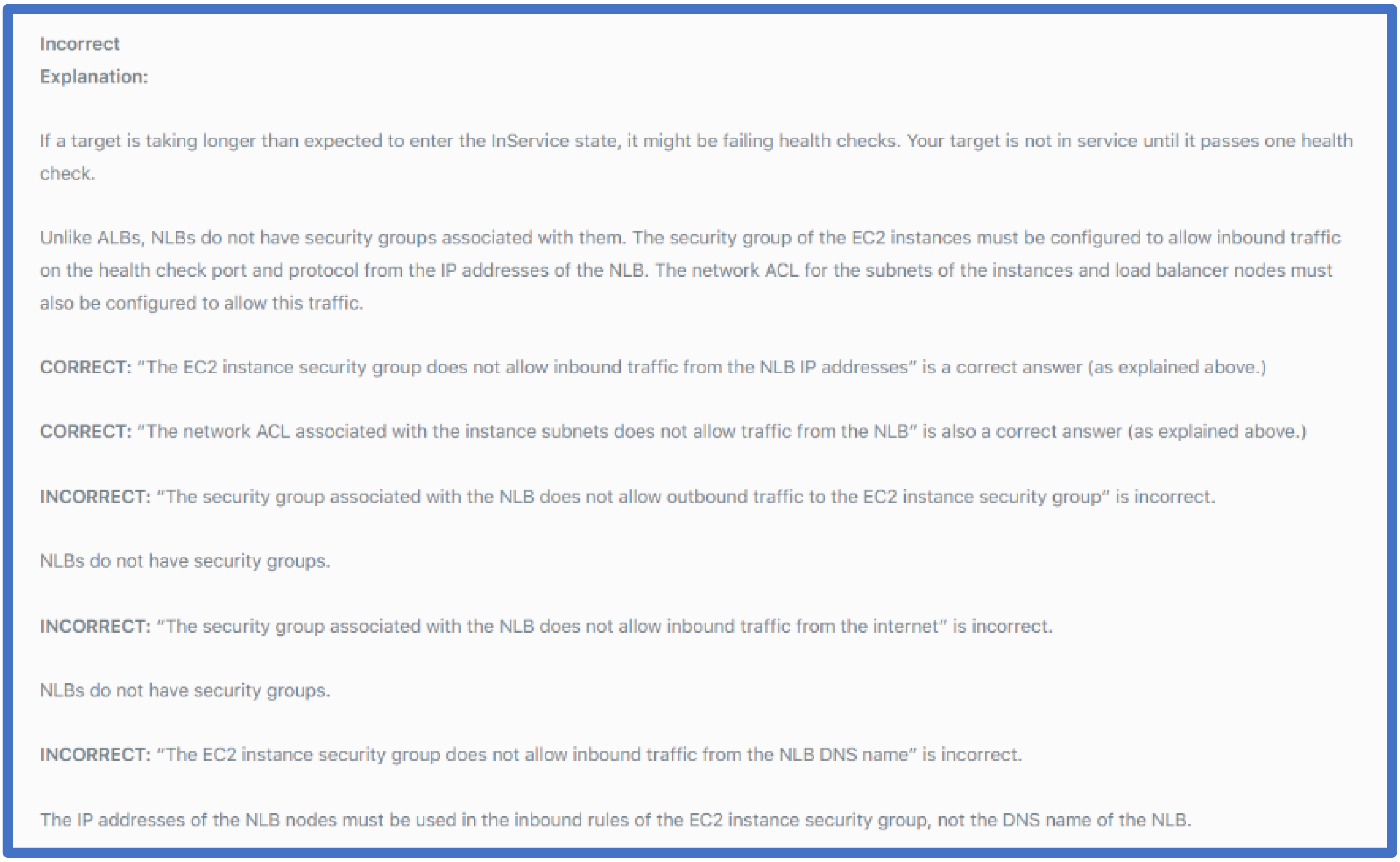
☐ The security group associated with the NLB does not allow inbound traffic from the internet.

☑ The security group associated with the NLB does not allow outbound traffic to the EC2 instance security group.

☑ The EC2 instance security group does not allow inbound traffic from the NLB IP addresses.

☐ The network ACL associated with the instance subnets does not allow traffic from the NLB.

☐ The EC2 instance security group does not allow inbound traffic from the NLB DNS name.



Incorrect

Explanation:

If a target is taking longer than expected to enter the InService state, it might be failing health checks. Your target is not in service until it passes one health check.

Unlike ALBs, NLBs do not have security groups associated with them. The security group of the EC2 instances must be configured to allow inbound traffic on the health check port and protocol from the IP addresses of the NLB. The network ACL for the subnets of the instances and load balancer nodes must also be configured to allow this traffic.

CORRECT: "The EC2 instance security group does not allow inbound traffic from the NLB IP addresses" is a correct answer (as explained above.)

CORRECT: "The network ACL associated with the instance subnets does not allow traffic from the NLB" is also a correct answer (as explained above.)

INCORRECT: "The security group associated with the NLB does not allow outbound traffic to the EC2 instance security group" is incorrect.

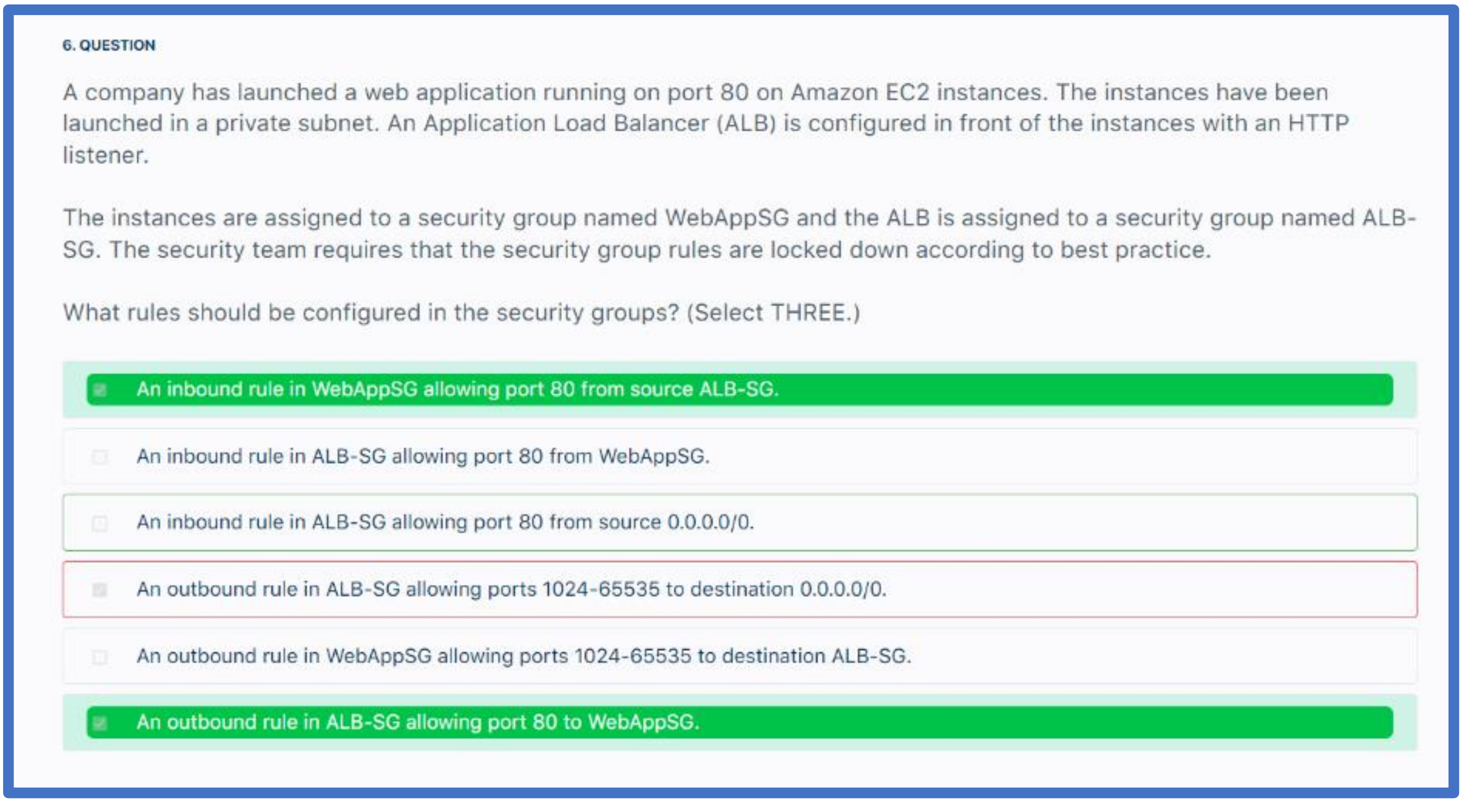
NLBs do not have security groups.

INCORRECT: "The security group associated with the NLB does not allow inbound traffic from the internet" is incorrect.

NLBs do not have security groups.

INCORRECT: "The EC2 instance security group does not allow inbound traffic from the NLB DNS name" is incorrect.

The IP addresses of the NLB nodes must be used in the inbound rules of the EC2 instance security group, not the DNS name of the NLB.



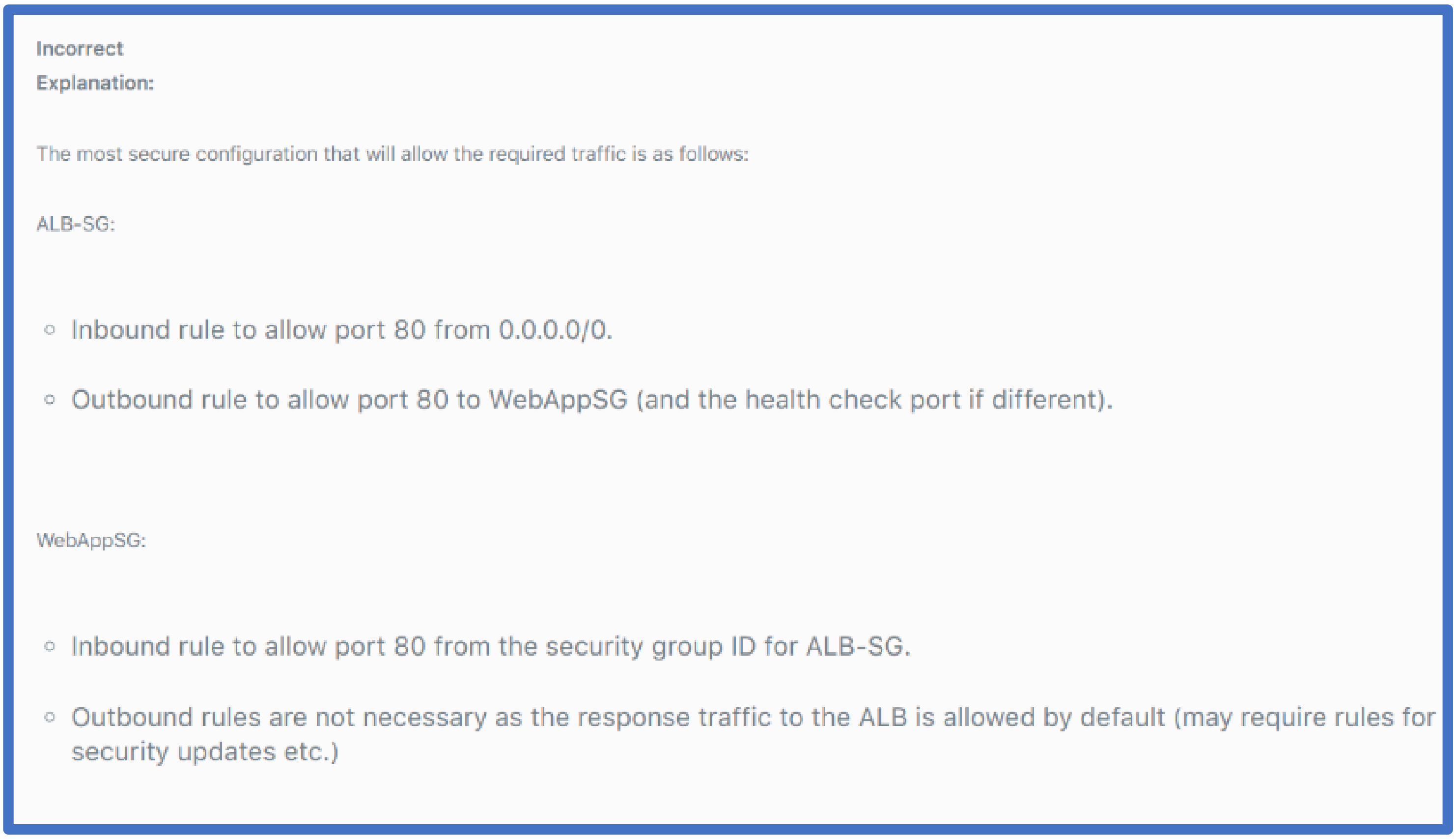
6. QUESTION

A company has launched a web application running on port 80 on Amazon EC2 instances. The instances have been launched in a private subnet. An Application Load Balancer (ALB) is configured in front of the instances with an HTTP listener.

The instances are assigned to a security group named WebAppSG and the ALB is assigned to a security group named ALB-SG. The security team requires that the security group rules are locked down according to best practice.

What rules should be configured in the security groups? (Select THREE.)

- ☑ An inbound rule in WebAppSG allowing port 80 from source ALB-SG.
- □ An inbound rule in ALB-SG allowing port 80 from WebAppSG.
- □ An inbound rule in ALB-SG allowing port 80 from source 0.0.0.0/0.
- ☑ An outbound rule in ALB-SG allowing ports 1024-65535 to destination 0.0.0.0/0.
- □ An outbound rule in WebAppSG allowing ports 1024-65535 to destination ALB-SG.
- ☑ An outbound rule in ALB-SG allowing port 80 to WebAppSG.



Incorrect

Explanation:

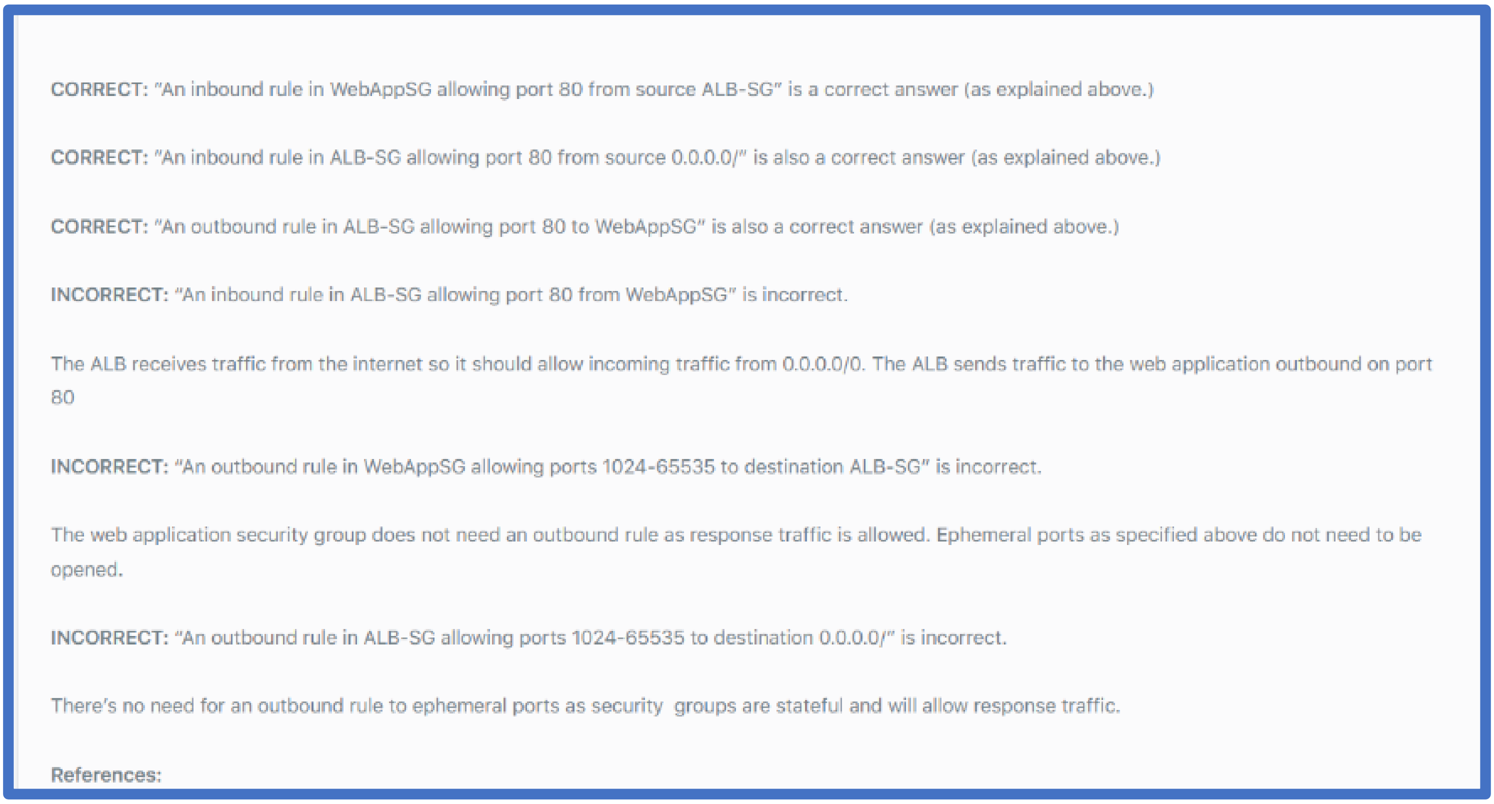
The most secure configuration that will allow the required traffic is as follows:

ALB-SG:

- Inbound rule to allow port 80 from 0.0.0.0/0.
- Outbound rule to allow port 80 to WebAppSG (and the health check port if different).

WebAppSG:

- Inbound rule to allow port 80 from the security group ID for ALB-SG.
- Outbound rules are not necessary as the response traffic to the ALB is allowed by default (may require rules for security updates etc.)



CORRECT: "An inbound rule in WebAppSG allowing port 80 from source ALB-SG" is a correct answer (as explained above.)

CORRECT: "An inbound rule in ALB-SG allowing port 80 from source 0.0.0.0/" is also a correct answer (as explained above.)

CORRECT: "An outbound rule in ALB-SG allowing port 80 to WebAppSG" is also a correct answer (as explained above.)

INCORRECT: "An inbound rule in ALB-SG allowing port 80 from WebAppSG" is incorrect.

The ALB receives traffic from the internet so it should allow incoming traffic from 0.0.0.0/0. The ALB sends traffic to the web application outbound on port 80

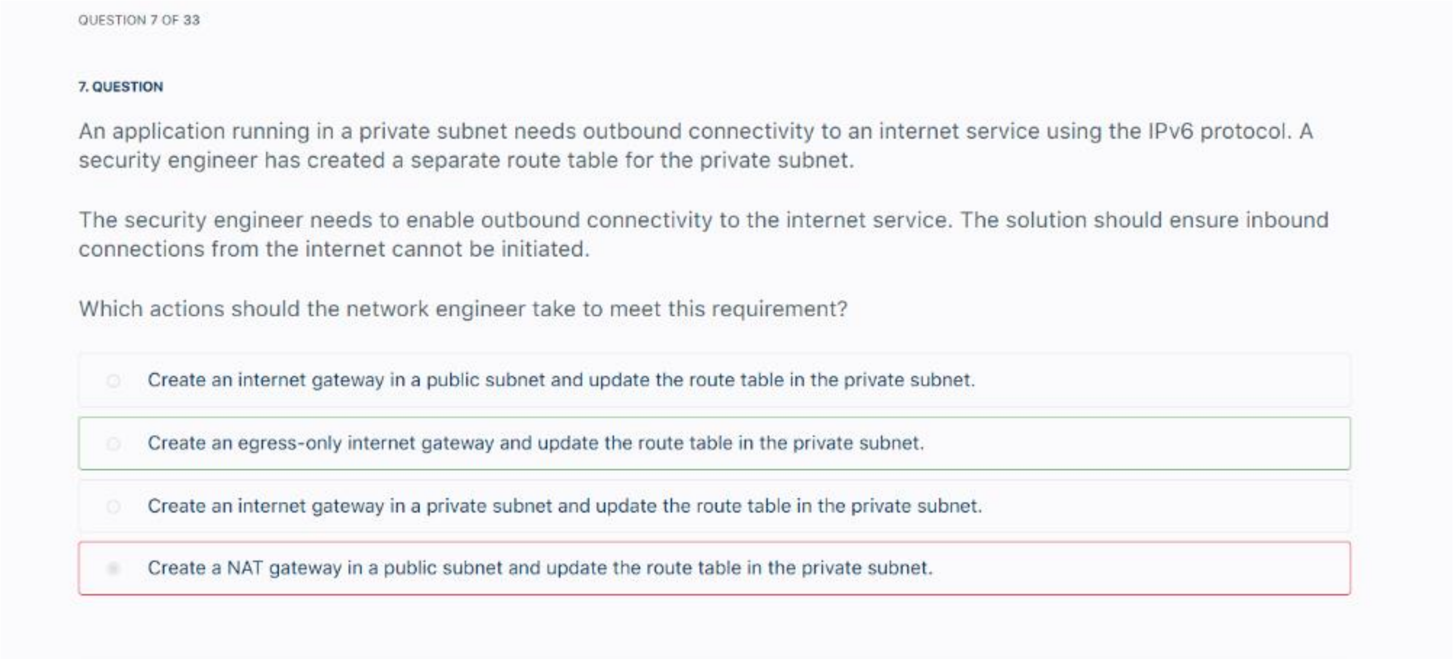
INCORRECT: "An outbound rule in WebAppSG allowing ports 1024-65535 to destination ALB-SG" is incorrect.

The web application security group does not need an outbound rule as response traffic is allowed. Ephemeral ports as specified above do not need to be opened.

INCORRECT: "An outbound rule in ALB-SG allowing ports 1024-65535 to destination 0.0.0.0/" is incorrect.

There's no need for an outbound rule to ephemeral ports as security groups are stateful and will allow response traffic.

References:



QUESTION 7 OF 33

7. QUESTION

An application running in a private subnet needs outbound connectivity to an internet service using the IPv6 protocol. A security engineer has created a separate route table for the private subnet.

The security engineer needs to enable outbound connectivity to the internet service. The solution should ensure inbound connections from the internet cannot be initiated.

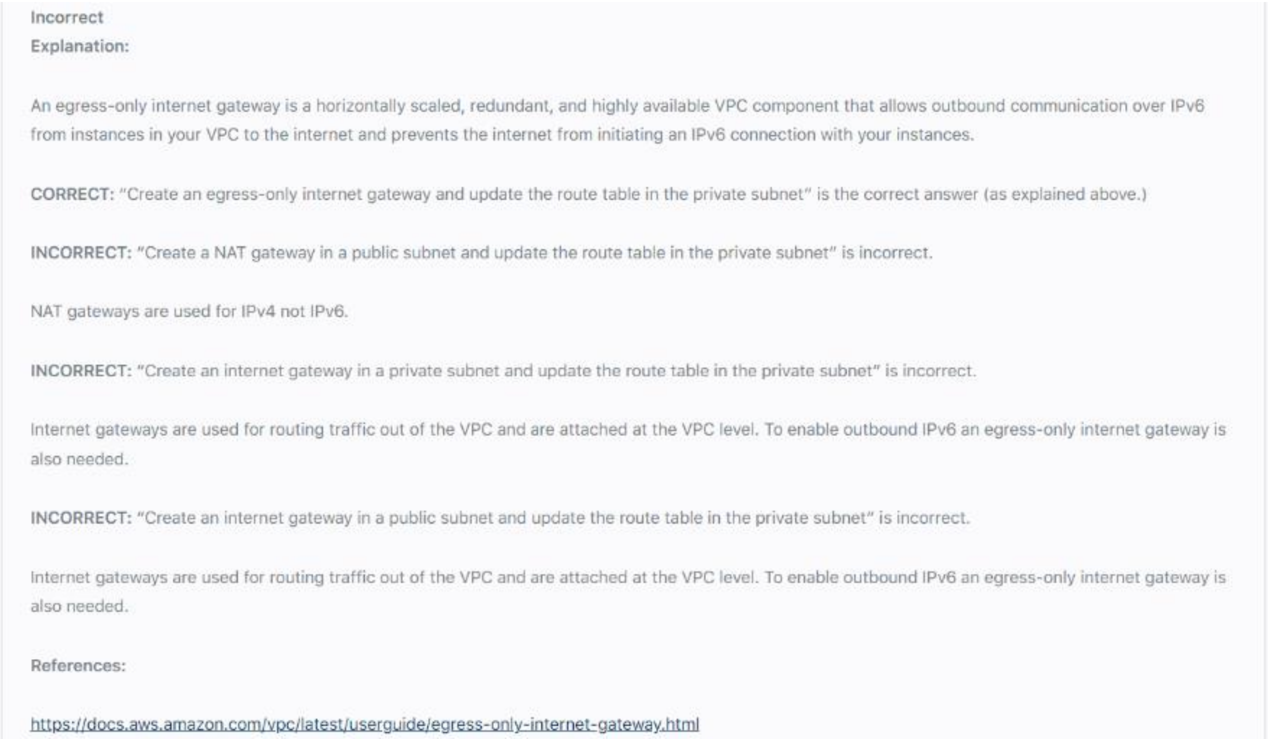
Which actions should the network engineer take to meet this requirement?

○ Create an internet gateway in a public subnet and update the route table in the private subnet.

○ Create an egress-only internet gateway and update the route table in the private subnet.

○ Create an internet gateway in a private subnet and update the route table in the private subnet.

● Create a NAT gateway in a public subnet and update the route table in the private subnet.



Incorrect

Explanation:

An egress-only internet gateway is a horizontally scaled, redundant, and highly available VPC component that allows outbound communication over IPv6 from instances in your VPC to the internet and prevents the internet from initiating an IPv6 connection with your instances.

CORRECT: "Create an egress-only internet gateway and update the route table in the private subnet" is the correct answer (as explained above.)

INCORRECT: "Create a NAT gateway in a public subnet and update the route table in the private subnet" is incorrect.

NAT gateways are used for IPv4 not IPv6.

INCORRECT: "Create an internet gateway in a private subnet and update the route table in the private subnet" is incorrect.

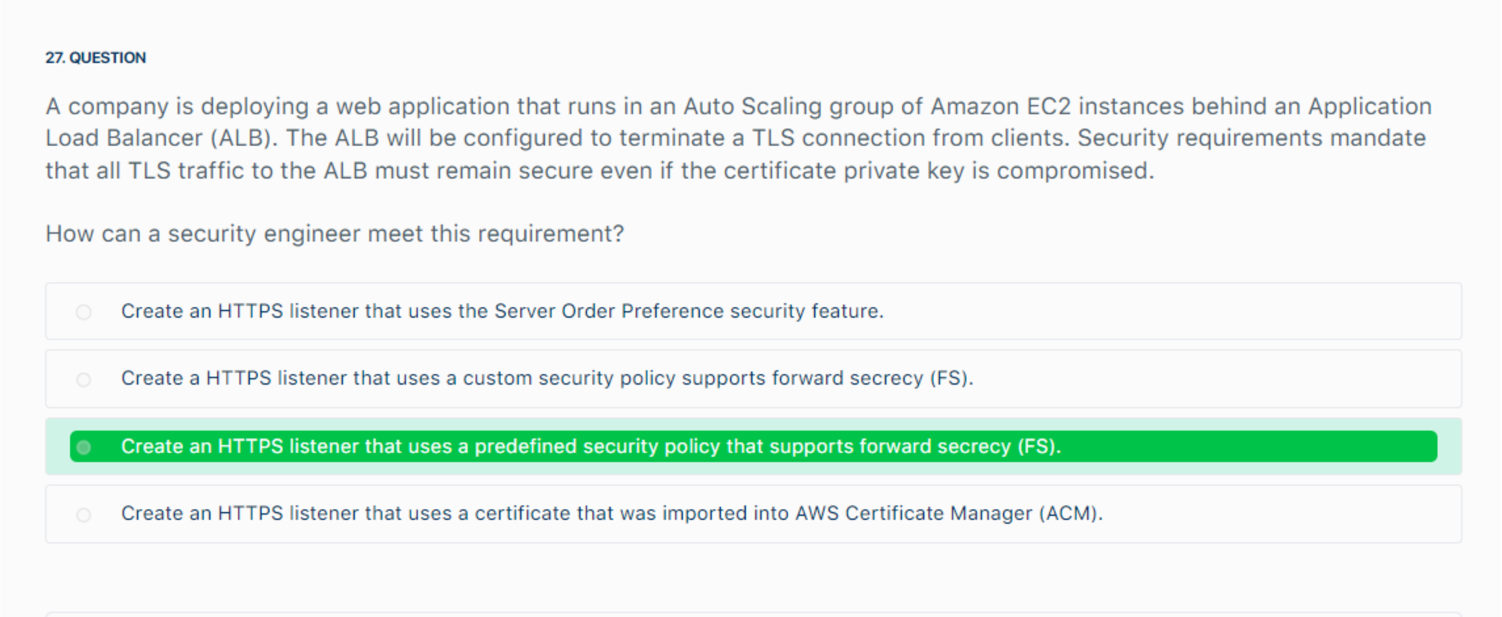
Internet gateways are used for routing traffic out of the VPC and are attached at the VPC level. To enable outbound IPv6 an egress-only internet gateway is also needed.

INCORRECT: "Create an internet gateway in a public subnet and update the route table in the private subnet" is incorrect.

Internet gateways are used for routing traffic out of the VPC and are attached at the VPC level. To enable outbound IPv6 an egress-only internet gateway is also needed.

References:

https://docs.aws.amazon.com/vpc/latest/userguide/egress-only-internet-gateway.html



27. QUESTION

A company is deploying a web application that runs in an Auto Scaling group of Amazon EC2 instances behind an Application Load Balancer (ALB). The ALB will be configured to terminate a TLS connection from clients. Security requirements mandate that all TLS traffic to the ALB must remain secure even if the certificate private key is compromised.

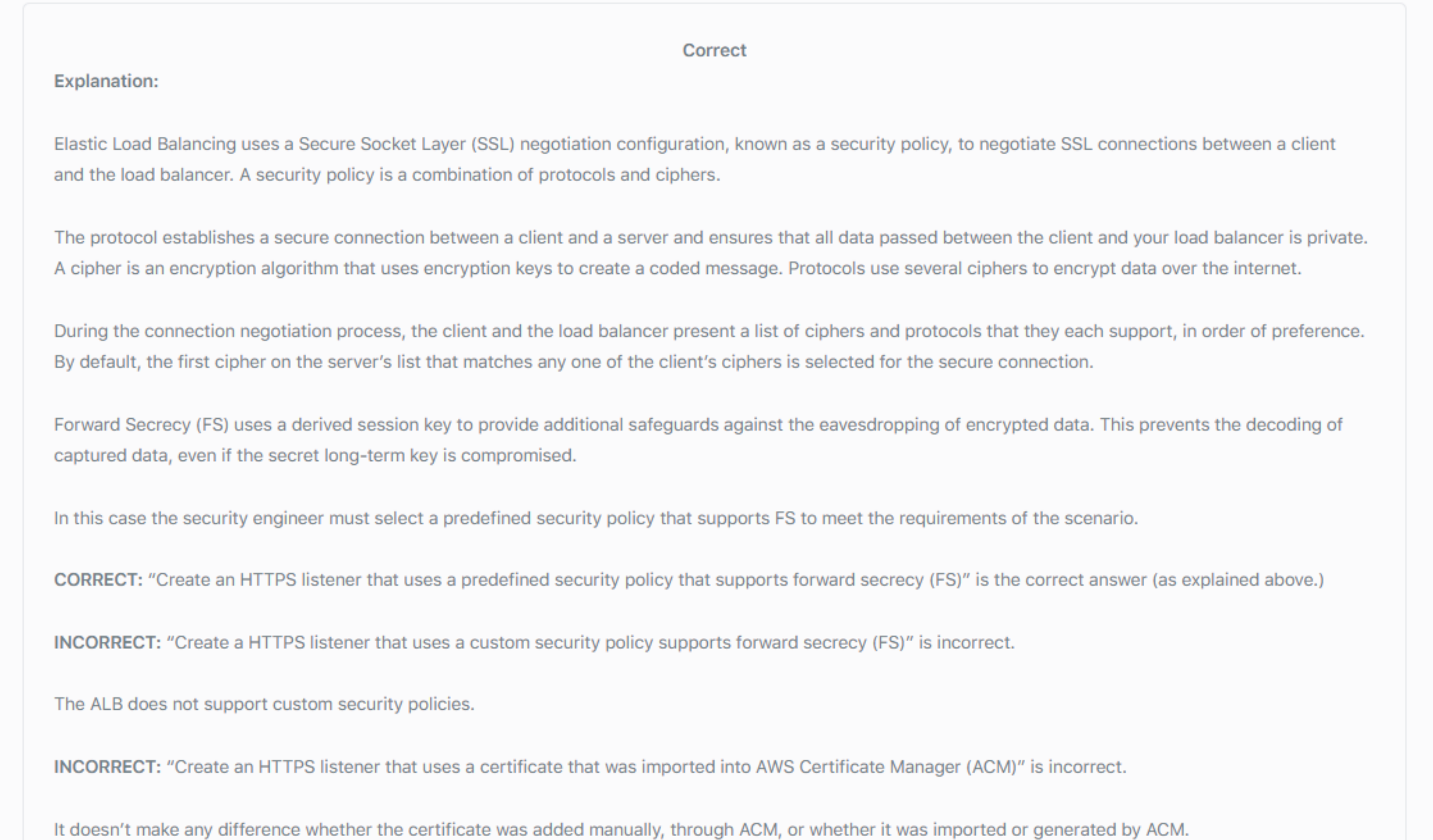
How can a security engineer meet this requirement?

○ Create an HTTPS listener that uses the Server Order Preference security feature.

○ Create a HTTPS listener that uses a custom security policy supports forward secrecy (FS).

● Create an HTTPS listener that uses a predefined security policy that supports forward secrecy (FS).

○ Create an HTTPS listener that uses a certificate that was imported into AWS Certificate Manager (ACM).



Correct

Explanation:

Elastic Load Balancing uses a Secure Socket Layer (SSL) negotiation configuration, known as a security policy, to negotiate SSL connections between a client and the load balancer. A security policy is a combination of protocols and ciphers.

The protocol establishes a secure connection between a client and a server and ensures that all data passed between the client and your load balancer is private. A cipher is an encryption algorithm that uses encryption keys to create a coded message. Protocols use several ciphers to encrypt data over the internet.

During the connection negotiation process, the client and the load balancer present a list of ciphers and protocols that they each support, in order of preference. By default, the first cipher on the server's list that matches any one of the client's ciphers is selected for the secure connection.

Forward Secrecy (FS) uses a derived session key to provide additional safeguards against the eavesdropping of encrypted data. This prevents the decoding of captured data, even if the secret long-term key is compromised.

In this case the security engineer must select a predefined security policy that supports FS to meet the requirements of the scenario.

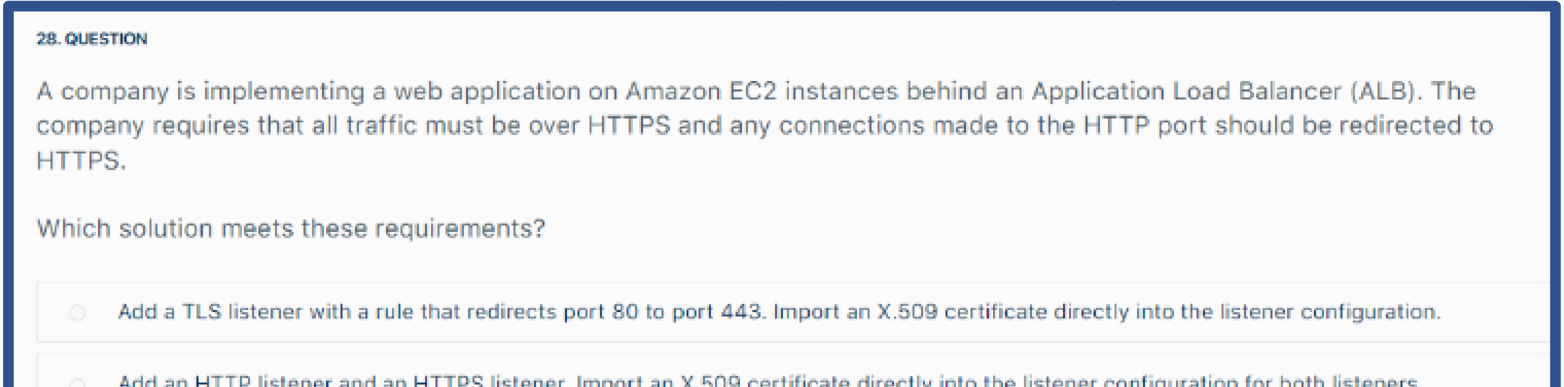
CORRECT: "Create an HTTPS listener that uses a predefined security policy that supports forward secrecy (FS)" is the correct answer (as explained above.)

INCORRECT: "Create a HTTPS listener that uses a custom security policy supports forward secrecy (FS)" is incorrect.

The ALB does not support custom security policies.

INCORRECT: "Create an HTTPS listener that uses a certificate that was imported into AWS Certificate Manager (ACM)" is incorrect.

It doesn't make any difference whether the certificate was added manually, through ACM, or whether it was imported or generated by ACM.



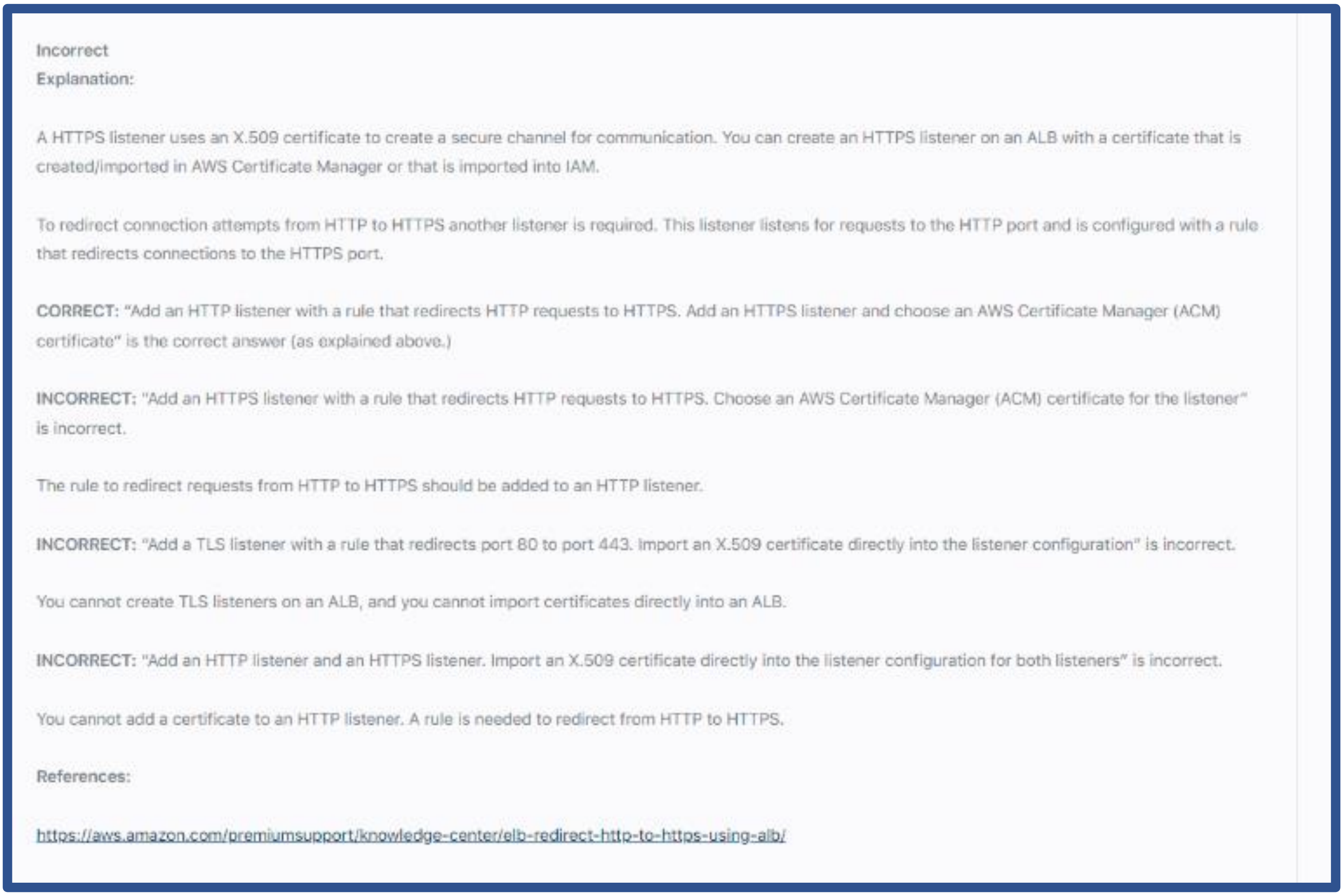
28. QUESTION

A company is implementing a web application on Amazon EC2 instances behind an Application Load Balancer (ALB). The company requires that all traffic must be over HTTPS and any connections made to the HTTP port should be redirected to HTTPS.

Which solution meets these requirements?

○ Add a TLS listener with a rule that redirects port 80 to port 443. Import an X.509 certificate directly into the listener configuration.

○ Add an HTTP listener and an HTTPS listener. Import an X.509 certificate directly into the listener configuration for both listeners.



Incorrect

Explanation:

A HTTPS listener uses an X.509 certificate to create a secure channel for communication. You can create an HTTPS listener on an ALB with a certificate that is created/imported in AWS Certificate Manager or that is imported into IAM.

To redirect connection attempts from HTTP to HTTPS another listener is required. This listener listens for requests to the HTTP port and is configured with a rule that redirects connections to the HTTPS port.

CORRECT: "Add an HTTP listener with a rule that redirects HTTP requests to HTTPS. Add an HTTPS listener and choose an AWS Certificate Manager (ACM) certificate" is the correct answer (as explained above.)

INCORRECT: "Add an HTTPS listener with a rule that redirects HTTP requests to HTTPS. Choose an AWS Certificate Manager (ACM) certificate for the listener" is incorrect.

The rule to redirect requests from HTTP to HTTPS should be added to an HTTP listener.

INCORRECT: "Add a TLS listener with a rule that redirects port 80 to port 443. Import an X.509 certificate directly into the listener configuration" is incorrect.

You cannot create TLS listeners on an ALB, and you cannot import certificates directly into an ALB.

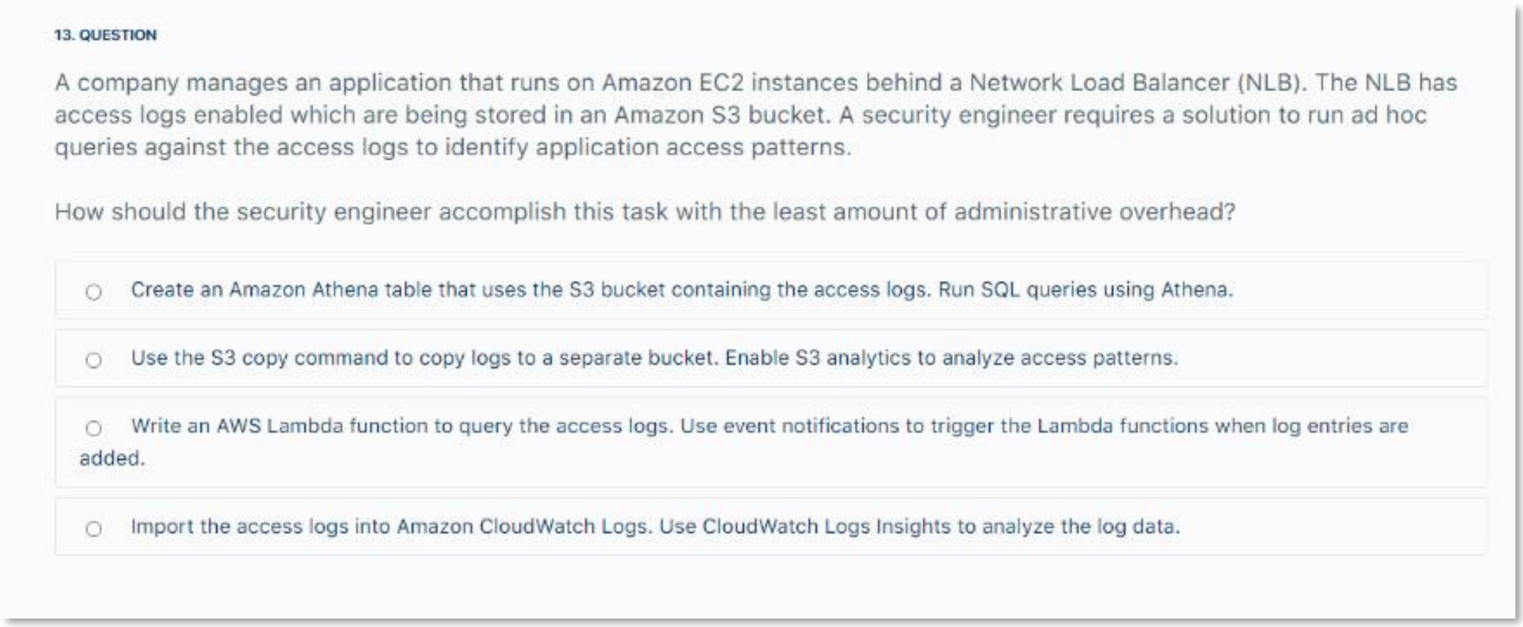
INCORRECT: "Add an HTTP listener and an HTTPS listener. Import an X.509 certificate directly into the listener configuration for both listeners" is incorrect.

You cannot add a certificate to an HTTP listener. A rule is needed to redirect from HTTP to HTTPS.

References:

https://aws.amazon.com/premiumsupport/knowledge-center/elb-redirect-http-to-https-using-alb/

Surely, we should add an HTTPS listener if we want to only use HTTPS. Given that we will be asking any requests that come in to transform into HTTPS, if will be okay if we only have an HTTPS listener.



13. QUESTION

A company manages an application that runs on Amazon EC2 instances behind a Network Load Balancer (NLB). The NLB has access logs enabled which are being stored in an Amazon S3 bucket. A security engineer requires a solution to run ad hoc queries against the access logs to identify application access patterns.

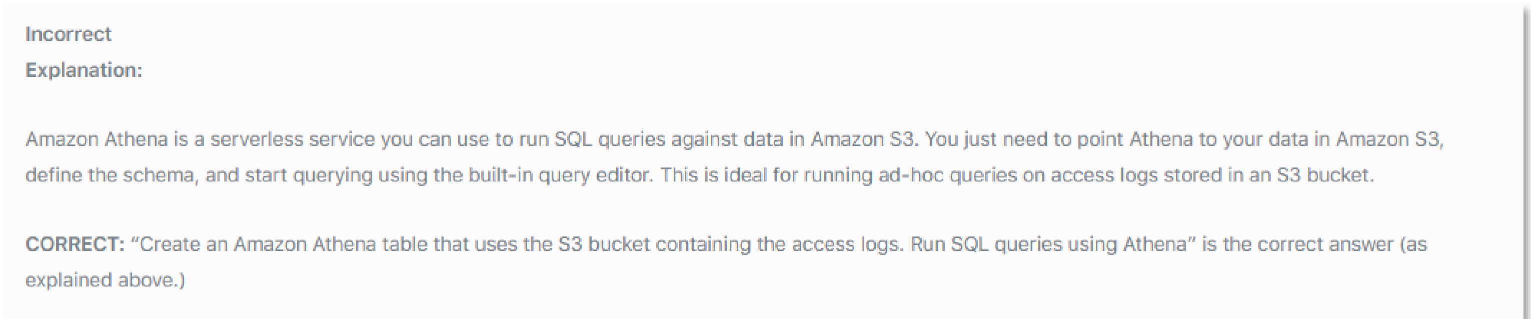
How should the security engineer accomplish this task with the least amount of administrative overhead?

○ Create an Amazon Athena table that uses the S3 bucket containing the access logs. Run SQL queries using Athena.

○ Use the S3 copy command to copy logs to a separate bucket. Enable S3 analytics to analyze access patterns.

○ Write an AWS Lambda function to query the access logs. Use event notifications to trigger the Lambda functions when log entries are added.

○ Import the access logs into Amazon CloudWatch Logs. Use CloudWatch Logs Insights to analyze the log data.

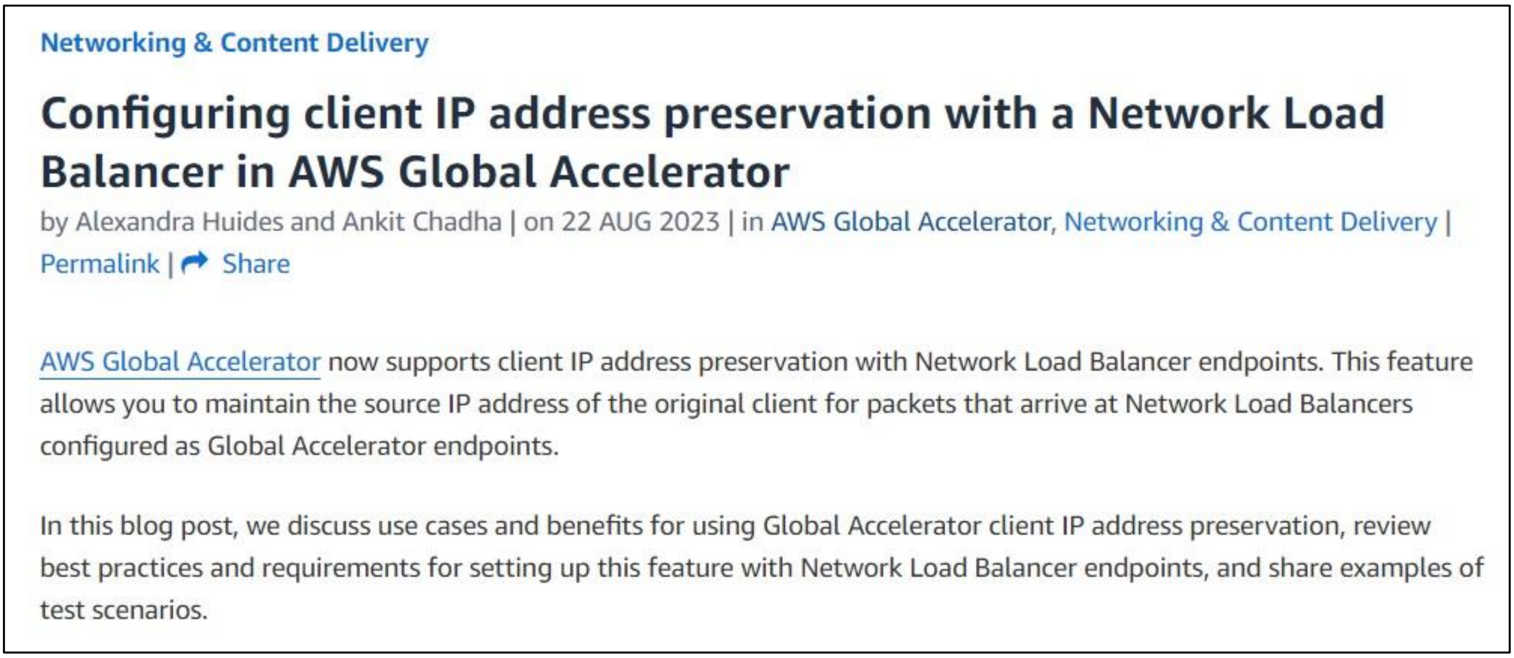


Incorrect

Explanation:

Amazon Athena is a serverless service you can use to run SQL queries against data in Amazon S3. You just need to point Athena to your data in Amazon S3, define the schema, and start querying using the built-in query editor. This is ideal for running ad-hoc queries on access logs stored in an S3 bucket.

CORRECT: "Create an Amazon Athena table that uses the S3 bucket containing the access logs. Run SQL queries using Athena" is the correct answer (as explained above.)



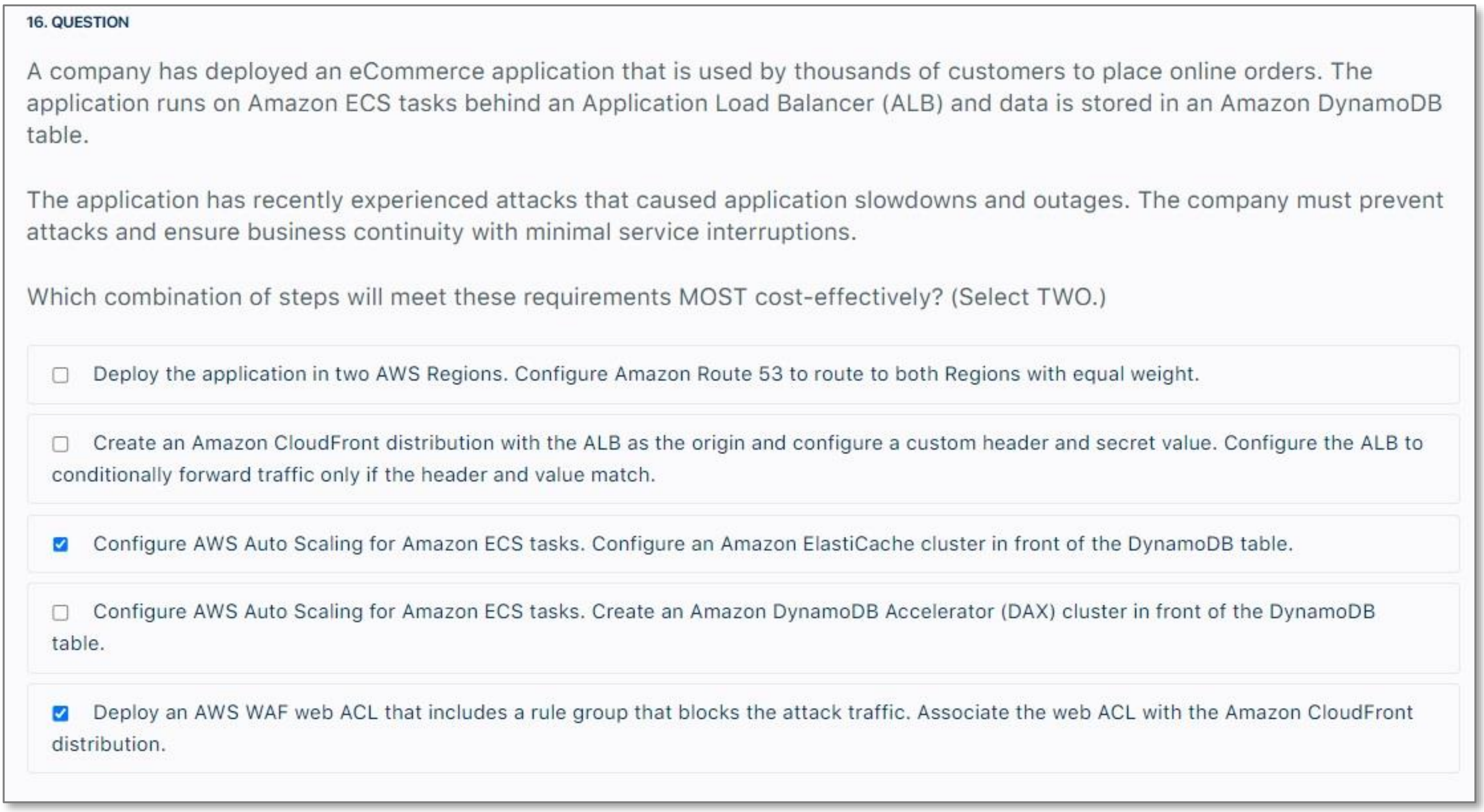
Networking & Content Delivery

Configuring client IP address preservation with a Network Load Balancer in AWS Global Accelerator

by Alexandra Huides and Ankit Chadha | on 22 AUG 2023 | in AWS Global Accelerator, Networking & Content Delivery | Permalink | Share

AWS Global Accelerator now supports client IP address preservation with Network Load Balancer endpoints. This feature allows you to maintain the source IP address of the original client for packets that arrive at Network Load Balancers configured as Global Accelerator endpoints.

In this blog post, we discuss use cases and benefits for using Global Accelerator client IP address preservation, review best practices and requirements for setting up this feature with Network Load Balancer endpoints, and share examples of test scenarios.



16. QUESTION

A company has deployed an eCommerce application that is used by thousands of customers to place online orders. The application runs on Amazon ECS tasks behind an Application Load Balancer (ALB) and data is stored in an Amazon DynamoDB table.

The application has recently experienced attacks that caused application slowdowns and outages. The company must prevent attacks and ensure business continuity with minimal service interruptions.

Which combination of steps will meet these requirements MOST cost-effectively? (Select TWO.)

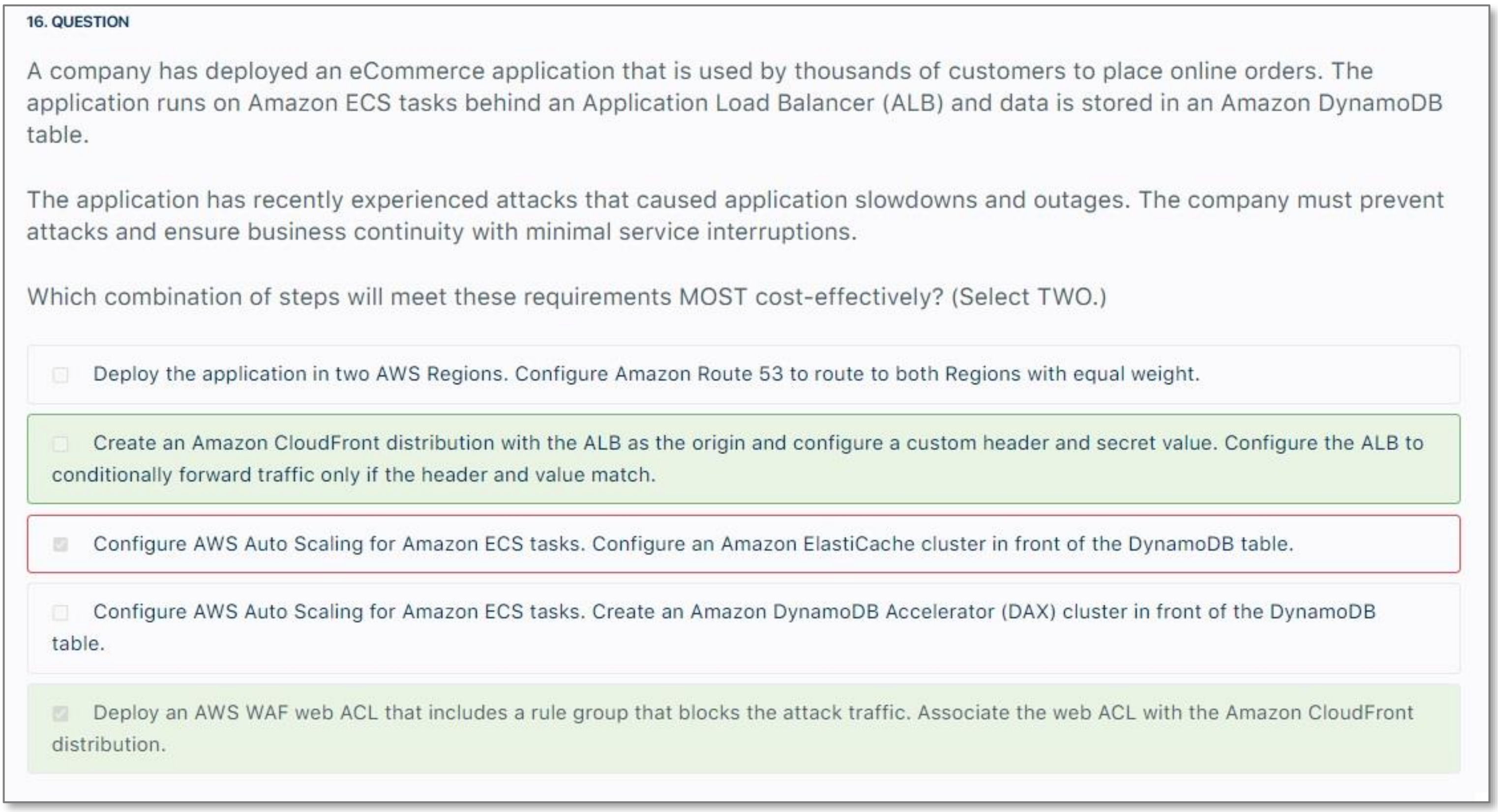
☐ Deploy the application in two AWS Regions. Configure Amazon Route 53 to route to both Regions with equal weight.

☐ Create an Amazon CloudFront distribution with the ALB as the origin and configure a custom header and secret value. Configure the ALB to conditionally forward traffic only if the header and value match.

☑ Configure AWS Auto Scaling for Amazon ECS tasks. Configure an Amazon ElastiCache cluster in front of the DynamoDB table.

☐ Configure AWS Auto Scaling for Amazon ECS tasks. Create an Amazon DynamoDB Accelerator (DAX) cluster in front of the DynamoDB table.

☑ Deploy an AWS WAF web ACL that includes a rule group that blocks the attack traffic. Associate the web ACL with the Amazon CloudFront distribution.



16. QUESTION

A company has deployed an eCommerce application that is used by thousands of customers to place online orders. The application runs on Amazon ECS tasks behind an Application Load Balancer (ALB) and data is stored in an Amazon DynamoDB table.

The application has recently experienced attacks that caused application slowdowns and outages. The company must prevent attacks and ensure business continuity with minimal service interruptions.

Which combination of steps will meet these requirements MOST cost-effectively? (Select TWO.)

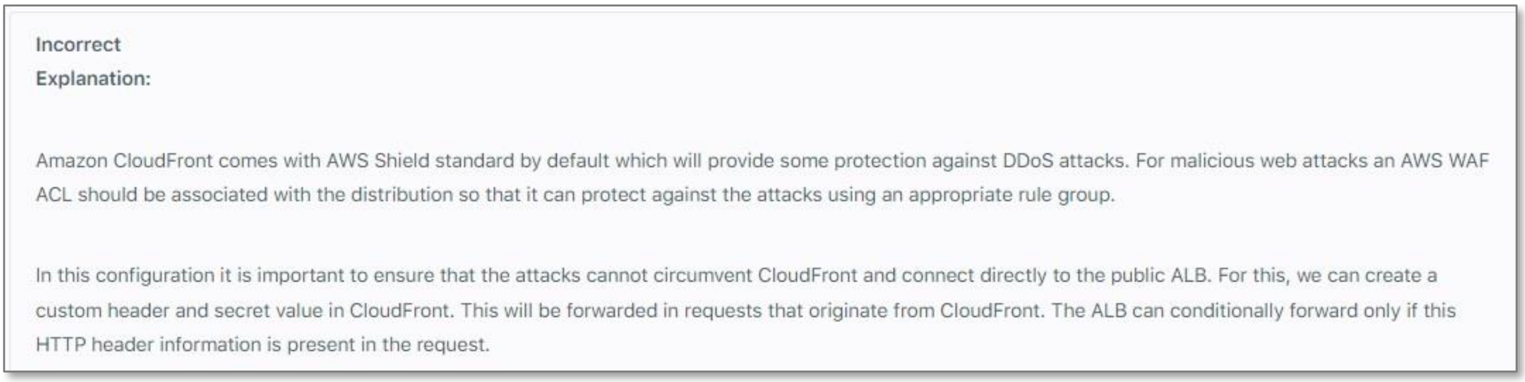
□ Deploy the application in two AWS Regions. Configure Amazon Route 53 to route to both Regions with equal weight.

□ Create an Amazon CloudFront distribution with the ALB as the origin and configure a custom header and secret value. Configure the ALB to conditionally forward traffic only if the header and value match.

☑ Configure AWS Auto Scaling for Amazon ECS tasks. Configure an Amazon ElastiCache cluster in front of the DynamoDB table.

□ Configure AWS Auto Scaling for Amazon ECS tasks. Create an Amazon DynamoDB Accelerator (DAX) cluster in front of the DynamoDB table.

☑ Deploy an AWS WAF web ACL that includes a rule group that blocks the attack traffic. Associate the web ACL with the Amazon CloudFront distribution.

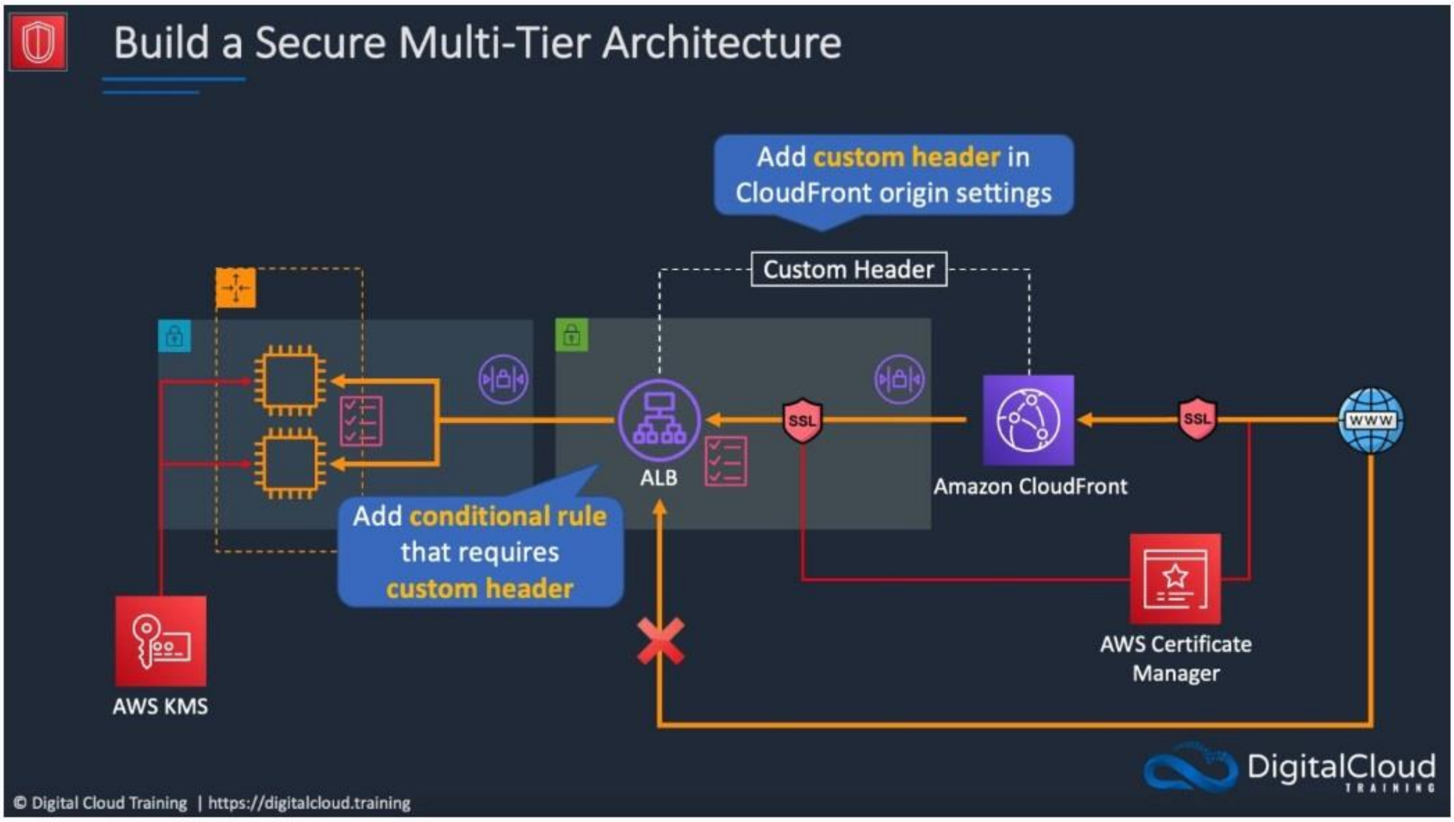


Incorrect

Explanation:

Amazon CloudFront comes with AWS Shield standard by default which will provide some protection against DDoS attacks. For malicious web attacks an AWS WAF ACL should be associated with the distribution so that it can protect against the attacks using an appropriate rule group.

In this configuration it is important to ensure that the attacks cannot circumvent CloudFront and connect directly to the public ALB. For this, we can create a custom header and secret value in CloudFront. This will be forwarded in requests that originate from CloudFront. The ALB can conditionally forward only if this HTTP header information is present in the request.

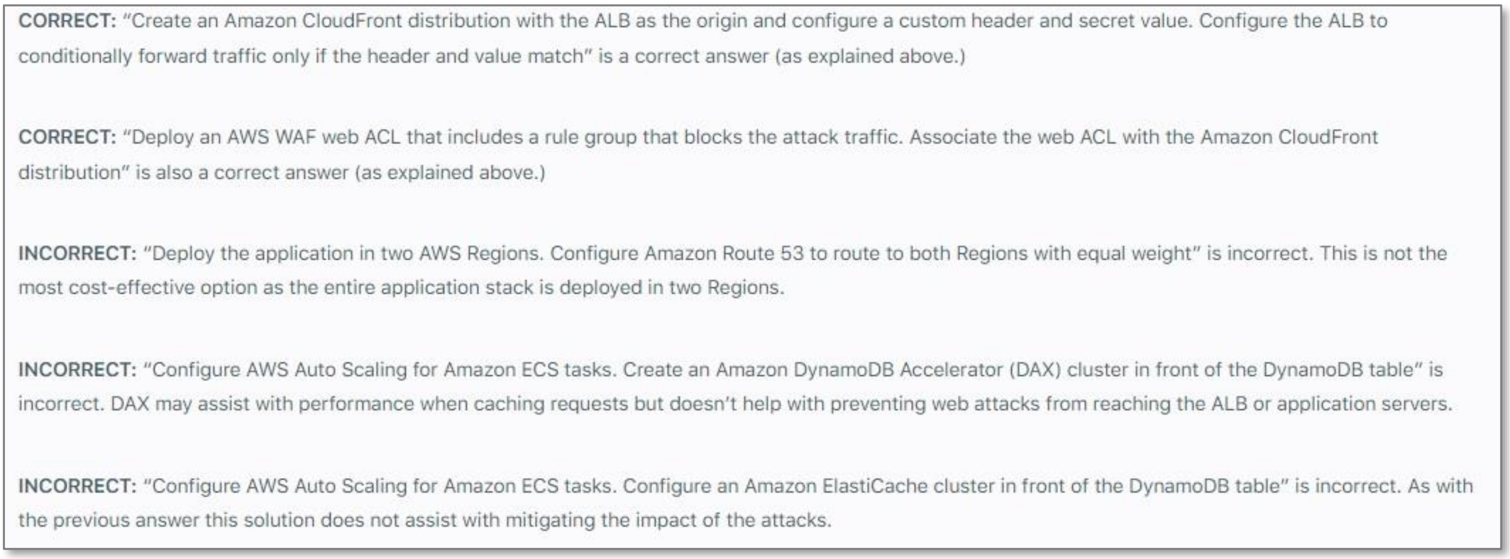


Build a Secure Multi-Tier Architecture

Add custom header in CloudFront origin settings

Add conditional rule that requires custom header

© Digital Cloud Training | https://digitalcloud.training



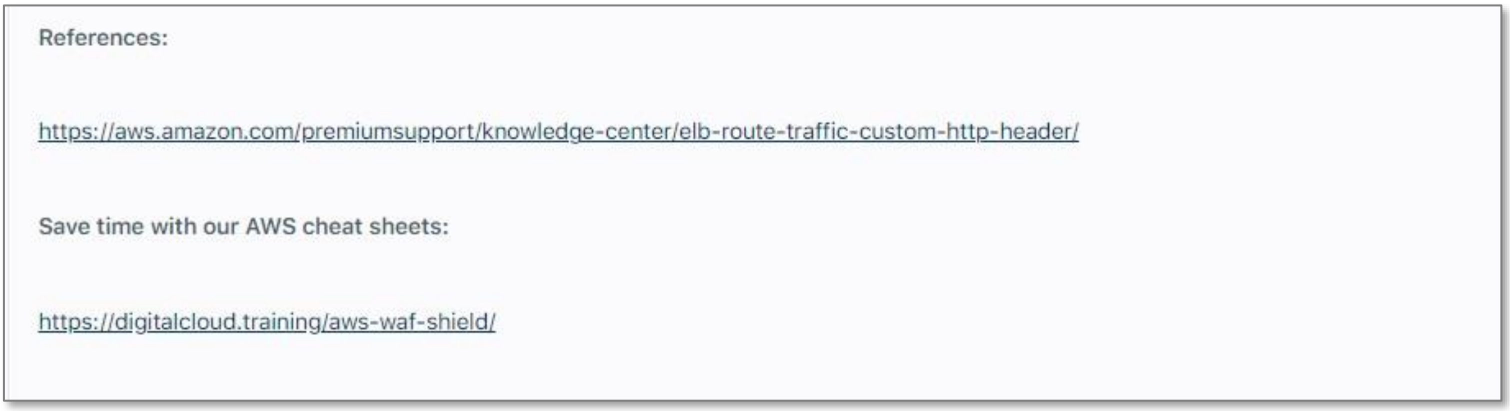
CORRECT: "Create an Amazon CloudFront distribution with the ALB as the origin and configure a custom header and secret value. Configure the ALB to conditionally forward traffic only if the header and value match" is a correct answer (as explained above.)

CORRECT: "Deploy an AWS WAF web ACL that includes a rule group that blocks the attack traffic. Associate the web ACL with the Amazon CloudFront distribution" is also a correct answer (as explained above.)

INCORRECT: "Deploy the application in two AWS Regions. Configure Amazon Route 53 to route to both Regions with equal weight" is incorrect. This is not the most cost-effective option as the entire application stack is deployed in two Regions.

INCORRECT: "Configure AWS Auto Scaling for Amazon ECS tasks. Create an Amazon DynamoDB Accelerator (DAX) cluster in front of the DynamoDB table" is incorrect. DAX may assist with performance when caching requests but doesn't help with preventing web attacks from reaching the ALB or application servers.

INCORRECT: "Configure AWS Auto Scaling for Amazon ECS tasks. Configure an Amazon ElastiCache cluster in front of the DynamoDB table" is incorrect. As with the previous answer this solution does not assist with mitigating the impact of the attacks.



References:

https://aws.amazon.com/premiumsupport/knowledge-center/elb-route-traffic-custom-http-header/

Save time with our AWS cheat sheets:

https://digitalcloud.training/aws-waf-shield/

TPN

21. ***Phenomenon1*** – the tendency of X to Y.
22. ***Phen2*** – the tendency of X to Y.
23. ***Phen3*** – the tendency of X to Y.
24. ***Phen4*** – the tendency of X to Y.
25. ***Phen5*** – the tendency of X to Y.
26. ***Phen6*** – the tendency of X to Y.
27. ***Phen7*** – the tendency of X to Y.
28. ***Phen8*** – the tendency of X to Y.
29. ***Phen9*** – the tendency of X to Y.
30. ***Phen10*** – the tendency of X to Y.

Glossary

Term1

Description of what term means here.

Term2

Description of what term means here.

Term3

Description of what term means here.

Bibliography

I. Official

[Barr 2009]

New Features for Amazon EC2: Elastic Load Balancing, Auto Scaling, and Amazon CloudWatch. AWS News Blog. 18th May 2009. Available at: https://aws.amazon.com/blogs/aws/new-aws-load-balancing-automatic-scaling-and-cloud-monitoring-services/

[Vogels 2009]

Automating the management of Amazon EC2 using Amazon CloudWatch, Auto Scaling and Elastic Load Balancing. 18th May 2009. *All Things Distributed* [Blog]. Available at: https://www.allthingsdistributed.com/2009/05/amazon_cloudwatch.html

[Peck 2017]

Peck, Nathan (2017). Using the New Network Load Balancer with Amazon ECS. 23rd October 2017. YouTube Channel: AWS Online Tech Talks. Available at: <https://www.youtube.com/watch?v=ekxSiLYwHfo&ab_channel=AWSOnlineTechTalks>

[Wood 2016]

Wood, Matt (2016). AWS Application Load Balancer. 1st Sept 2016 [date video published to YouTube]. YouTube Channel: Amazon Web Services. Available at:

<https://www.youtube.com/watch?v=LBow4273Ym8&ab_channel=AmazonWebServices>

[Gnaneshwari 2020]

Gnaneshwari (2020). How do I achieve path-based routing on an Application Load Balancer?. 7th Jan 2020. YouTube Channel: Amazon Web Services. Available at: <https://www.youtube.com/watch?v=KK5YwtLTNYw&ab_channel=AmazonWebServices>.

[Brown 2016]

Brown, David (2016). Elastic Load Balancing: Deep Dive and Best Practices. Reinvent conference 2016 (NET403). 1st December 2016. YouTube Channel: Amazon Web Services. Available at: https://www.youtube.com/watch?v=qy7zNaDTYGQ&ab_channel=AmazonWebServices

[Vecchioli 2016]

Vecchioli, Mariano and Ben Doyle (2016). Deep Dive on Elastic Load Balancing. AWS Summit Series: London. July 2016. YouTube Channel: Amazon Web Services. Available at: https://www.youtube.com/watch?v=HinwLb2lpLQ&ab_channel=AmazonWebServices

[MacCárthaigh 2014]

MacCárthaigh, Colm (2014). SSL with Amazon Web Services. 12th November 2014. Reinvent conference (SEC316). Available at: <https://www.youtube.com/watch?v=8AODa_AazY4&ab_channel=AmazonWebServices>

[Dalbhanjan 2016]

Dalbhanjan, Peter and Jeff Storey and Kit Ewbank (2016). From Amazon EC2 to Amazon ECS. 12th January 2016. Reinvent conference (NET203). Available at: <https://www.youtube.com/watch?v=uFs_EwJr-yc&ab_channel=AmazonWebServices>

[Suryadevara 2018]

Suryadevara, Pratibha and Will Rose (2018). Elastic Load Balancing: Deep Dive and Best Practices. November 2018. Reinvent conference (NET404R). Available at:

<https://www.youtube.com/watch?v=VIgAT7vjol8&ab_channel=AmazonWebServices>

[Suryadevara 2017]

Suryadevara, Pratibha and Narayan Subramaniam and Bryan McKenney (2017). Deep Dive into the New Network Load Balancer. 28th November 2017. Reinvent conference (NET304). Available at: <https://www.youtube.com/watch?v=z0FBGIT1Ub4&ab_channel=AmazonWebServices>

[Barr 2017a]

Barr, Jeff (2017). Introduction to Elastic Network Load Balancer. 8th September 2017. YouTube video ("AWS Launch"). Channel: Amazon Web Services. Available at: <https://www.youtube.com/watch?v=YXyMDNcxHkc&ab_channel=AmazonWebServices>

[Subramaniam 2019]

Subramaniam, Narayam and Priyank Goyal and Varun Lodaya (2019). Smart Tips on Application Load Balancer. YouTube channel: AWS Online Tech Talks. 23rd April 2019. Available at: <https://www.youtube.com/watch?v=Sdvp4qQ1rFA&t=2s&ab_channel=AWSOnlineTechTalks>

[Pessis 2017]

Pessis, David and Narayan Subramaniam (2017). Elastic Load Balancing Deep Dive. YouTube channel: AWS Online Tech Talks. December 2019. Available at: https://www.youtube.com/watch?v=bjehc2K5gyA&ab_channel=AWSOnlineTechTalks

[Wenzel 2021]

Wenzel, James (2021). How to choose the right load balancer for your AWS workloads. Reinvent 2021 (NET202). Available at: <https://www.youtube.com/watch?v=p0YZBF03r5A&ab_channel=AWSEvents>

[Zobrist 2021]

Zobrist, Jon (2021). Elastic Load Balancing: A Year of Innovations (NET402). Available at: https://www.youtube.com/watch?v=cntxaahxtfM&ab_channel=AWSEvents

[Barr 2017b]

New Network Load Balancer: Effortless Scaling to Millions of Requests per Second. *AWS News Blog*. 7th Sept 2017. Available at: https://aws.amazon.com/blogs/aws/new-network-load-balancer-effortless-scaling-to-millions-of-requests-per-second/#:~:text=Elastic%20Load%20Balancing%20(ELB)%20has,Auto%20Scaling%20and%20Amazon%20CloudWatch.

[AWS 2023]

Network Load Balancer Now Supports Security Groups. [Announcement]. Aug 10th 2023. Available at: <https://aws.amazon.com/about-aws/whats-new/2023/08/network-load-balancer-supports-security-groups/?ck_subscriber_id=1560524742>

[Huides 2023]

Huides, Alexandra and Ankit Chadha (2023). Configuring client IP address preservation with a Network Load Balancer in AWS Global Accelerator. *Networking and Content Delivery* [Blog]. Aug 22nd 2023. Available at: <https://aws.amazon.com/blogs/networking-and-content-delivery/configuring-client-ip-address-preservation-with-a-network-load-balancer-in-aws-global-accelerator/?ck_subscriber_id=1560524742>

II. Unofficial

https://www.youtube.com/watch?v=WLu41jAYjYk&ab_channel=JavaHomeCloud

https://www.youtube.com/watch?v=cAEtjMI1KcQ&ab_channel=ThornTechnologies

[Surname1]

McMahon, Nathan (2020). The 3 Generations of Load Balancing, According to a Lifer. 16th April 2020. VMWare Blogs. Available at: <https://blogs.vmware.com/load-balancing/2020/04/16/the-3-generations-of-load-balancing-according-to-a-lifer/>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Swidler 2010]

Swidler, Shlomo (2010). Elastic Load Balancing with Sticky Sessions. Shlomoswidler.com. Available at: https://shlomoswidler.com/tag/sticky-sessions/

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

IV. General

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

http://euro.ecom.cmu.edu/resources/elibrary/epay/SSL.pdf

https://en.wikipedia.org/wiki/HTTP/2

https://en.wikipedia.org/wiki/URL#Syntax

https://bravelittletank.com/wp-content/uploads/2018/02/the-complete-guide-to-ssl-certificates-ebook.pdf

VPC

There is a rough distinction maintained, between EC2 and VPC. The focus of the EC2 domain is matters related to compute. For example, information about instance types and EBS volumes. The focus of the VPC domain is more on networking. We cover packet-filterers such as the NACL here, as well as sub-networks and Classless Interdomain Routing block ("CIDR blocks"). To explain the below image, the icon for the EC2 service was originally an orange rectangle, so I've used this colour. Blue was somewhat arbitrary.

Two domains to master

You need to be aware of this distinction for two reasons. First, AWS currently maintain two documents: the EC2 user guide and the VPC user guide. You need to remember to search both as you solve problems as a professional. Second, each domain contains a substantial amount of material. If you forget that there are these TWO domains, you can make quite a large misconception about the quantity you have learned.

You will find that many topics can quite easily be put in either the VPC domain or the EC2 domain. I will call this property *OVERLAP*. For example, AWS currently mention High Performance Computing (HPC) in the EC2 user guide only. However, Piper and Clinton (2021, Sybex) put this topic in their *VPC* chapter. Here is a second example. Security groups (SGs) ought to be discussed in EC2, since they protect EC2 *instances*. Yet SGs are often discussed in the VPC domain, perhaps so that they can be contrasted with NACLs (which act at the sub-network level). AWS currently have a section devoted to SGs in the *VPC* user guide. The point is that the two domains experience OVERLAP.

But just because many topics get transferred between the two domains (EC2 and VPC)—this does not change the fact that the quantity we are dealing with is *two* domains, not one! Don't let the property of *OVERLAP* tempt you into thinking that because you've mastered one domain, you've mastered both. In the COPMUTE (EC2) and NETWORK (VPC) domains, there's a *lot* to learn, and it will take time for all the nitty issues to come out of the woodwork. My point is this:

> leaking into one another as they may be, please have *two* looming domains writ large in your mind, not one.

EC2

Matters relating to COMPUTE

VPC

Matters relating to NETWORKING

	Module One – Basics and background	
	Module Two	
	1	Subnets
	2	ENIs (Elastic Network Interfaces)
	3	Route Tables
		What is the default route?
	4	NACLs
	Module Three – Connect your VPC	
	1	Internet Gateways
	2	Egress-only Internet Gateway
	3	NAT Devices: • NAT Instances • NAT Gateways
	Module Four – Monitoring	
	1	
	2	
	3	
	Module Five - Security	
	1	What is a security group (SG)?
	2	
	3	Traffic mirroring
	4	Hybrid Cloud Networking

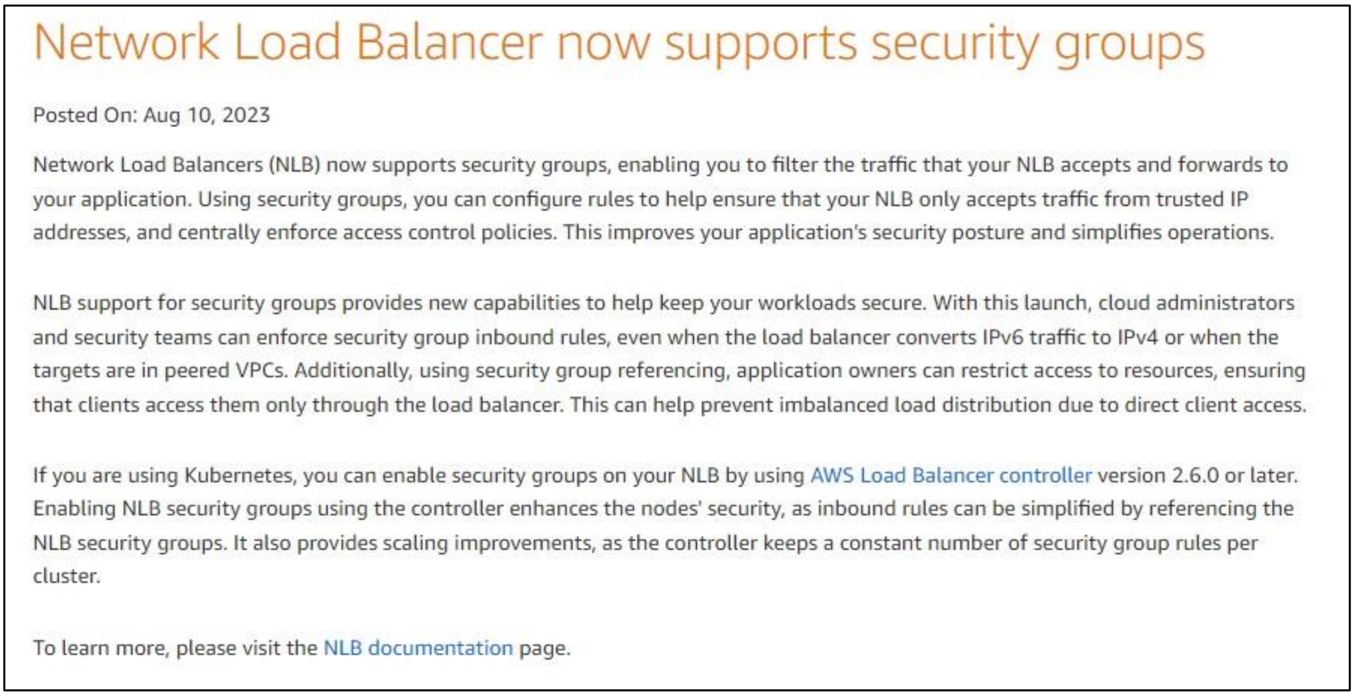
Module 1 – Basics and background

The preface “virtual private” is used for quite a few things.

Virtual private cloud (VPC)

Virtual private gateway

Virtual private network (VPN)



Network Load Balancer now supports security groups

Posted On: Aug 10, 2023

Network Load Balancers (NLB) now supports security groups, enabling you to filter the traffic that your NLB accepts and forwards to your application. Using security groups, you can configure rules to help ensure that your NLB only accepts traffic from trusted IP addresses, and centrally enforce access control policies. This improves your application's security posture and simplifies operations.

NLB support for security groups provides new capabilities to help keep your workloads secure. With this launch, cloud administrators and security teams can enforce security group inbound rules, even when the load balancer converts IPv6 traffic to IPv4 or when the targets are in peered VPCs. Additionally, using security group referencing, application owners can restrict access to resources, ensuring that clients access them only through the load balancer. This can help prevent imbalanced load distribution due to direct client access.

If you are using Kubernetes, you can enable security groups on your NLB by using AWS Load Balancer controller version 2.6.0 or later. Enabling NLB security groups using the controller enhances the nodes' security, as inbound rules can be simplified by referencing the NLB security groups. It also provides scaling improvements, as the controller keeps a constant number of security group rules per cluster.

To learn more, please visit the NLB documentation page.

TPN

31. ***Phenomenon1*** – the tendency of X to Y.
32. ***Phen2*** – the tendency of X to Y.
33. ***Phen3*** – the tendency of X to Y.
34. ***Phen4*** – the tendency of X to Y.
35. ***Phen5*** – the tendency of X to Y.
36. ***Phen6*** – the tendency of X to Y.
37. ***Phen7*** – the tendency of X to Y.
38. ***Phen8*** – the tendency of X to Y.
39. ***Phen9*** – the tendency of X to Y.
40. ***Phen10*** – the tendency of X to Y.

Glossary

Term1

Description of what term means here.

Term2

Description of what term means here.

Term3

Description of what term means here.

Bibliography

I. Official

https://medium.com/awesome-cloud/aws-vpc-difference-between-internet-gateway-and-nat-gateway-c9177e710af6#:~:text=Internet%20Gateway%20(IGW)%20allows%20instances,IPs%20to%20access%20the%20internet.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[AWS 2023]

Network Load Balancer Now Supports Security Groups. [Announcement]. Aug 10th 2023. Available at: <https://aws.amazon.com/about-aws/whats-new/2023/08/network-load-balancer-supports-security-groups/?ck_subscriber_id=1560524742>

II. Unofficial

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

IV. General

[Surname1]

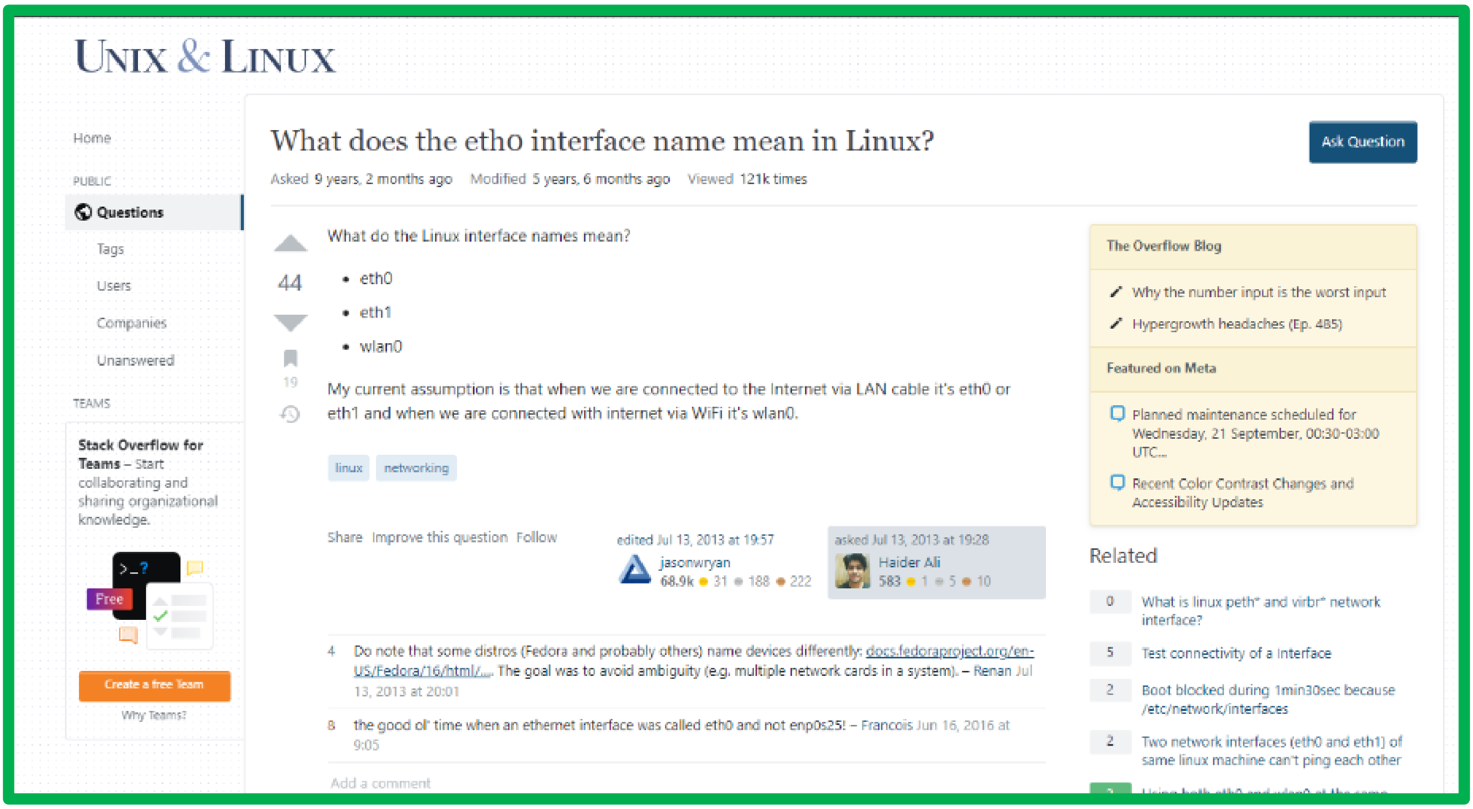
Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Sharwood 2024]

Sharwood, Simon (2024). Starlink offers ‘unusually hostile environment’ to TCP. *The Register*. May 22nd 2024. Available at:
<https://www.theregister.com/2024/05/22/starlink_tcp_performance_evaluation/>

Module 2

Basics and background



What does the eth0 interface name mean in Linux?

Asked 9 years, 2 months ago Modified 5 years, 6 months ago Viewed 121k times

What do the Linux interface names mean?

- eth0
- eth1
- wlan0

My current assumption is that when we are connected to the Internet via LAN cable it's eth0 or eth1 and when we are connected with internet via WiFi it's wlan0.

edited Jul 13, 2013 at 19:57
jasonwryan
68.9k ● 31 ● 188 ● 222

asked Jul 13, 2013 at 19:28
Haider Ali
583 ● 1 ● 5 ● 10

4 Do note that some distros (Fedora and probably others) name devices differently: docs.fedoraproject.org/en-US/Fedora/16/html/.... The goal was to avoid ambiguity (e.g. multiple network cards in a system). – Renan Jul 13, 2013 at 20:01

8 the good ol' time when an ethernet interface was called eth0 and not enp0s25! – Francois Jun 16, 2016 at 9:05

The Overflow Blog

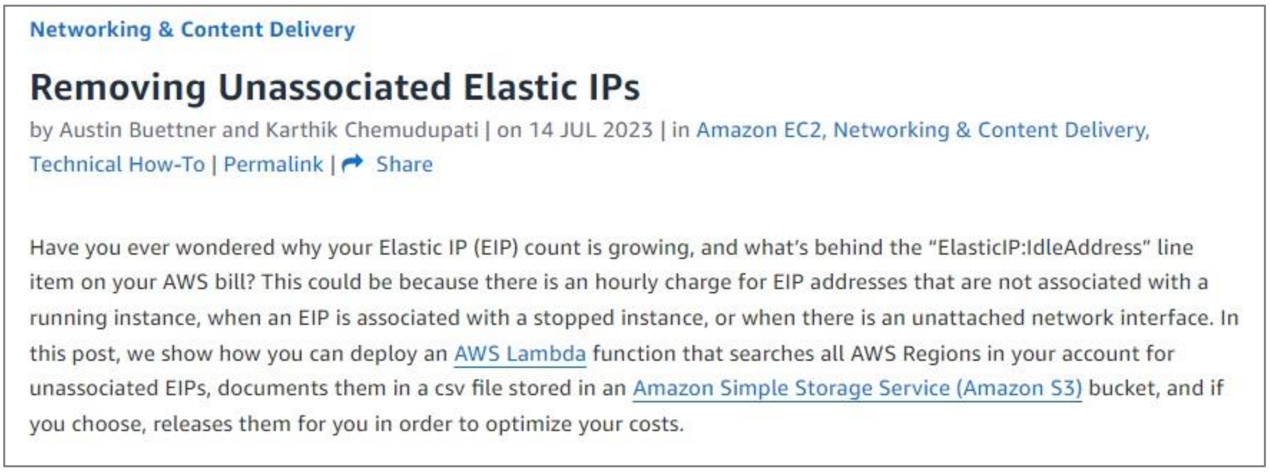
- Why the number input is the worst input
- Hypergrowth headaches (Ep. 485)

Featured on Meta

- Planned maintenance scheduled for Wednesday, 21 September, 00:30-03:00 UTC...
- Recent Color Contrast Changes and Accessibility Updates

Related

0	What is linux peth* and virbr* network interface?
5	Test connectivity of a Interface
2	Boot blocked during 1min30sec because /etc/network/interfaces
2	Two network interfaces (eth0 and eth1) of same linux machine can't ping each other



Networking & Content Delivery

Removing Unassociated Elastic IPs

by Austin Buettner and Karthik Chemudupati | on 14 JUL 2023 | in Amazon EC2, Networking & Content Delivery, Technical How-To | Permalink | Share

Have you ever wondered why your Elastic IP (EIP) count is growing, and what's behind the "ElasticIP:IdleAddress" line item on your AWS bill? This could be because there is an hourly charge for EIP addresses that are not associated with a running instance, when an EIP is associated with a stopped instance, or when there is an unattached network interface. In this post, we show how you can deploy an AWS Lambda function that searches all AWS Regions in your account for unassociated EIPs, documents them in a csv file stored in an Amazon Simple Storage Service (Amazon S3) bucket, and if you choose, releases them for you in order to optimize your costs.

TPN

41. ***Phenomenon1*** – the tendency of X to Y.
42. ***Phen2*** – the tendency of X to Y.
43. ***Phen3*** – the tendency of X to Y.
44. ***Phen4*** – the tendency of X to Y.
45. ***Phen5*** – the tendency of X to Y.
46. ***Phen6*** – the tendency of X to Y.
47. ***Phen7*** – the tendency of X to Y.
48. ***Phen8*** – the tendency of X to Y.
49. ***Phen9*** – the tendency of X to Y.
50. ***Phen10*** – the tendency of X to Y.

Glossary

Term1

Description of what term means here.

Term2

Description of what term means here.

Term3

Description of what term means here.

Bibliography

I. Official

https://medium.com/awesome-cloud/aws-vpc-difference-between-internet-gateway-and-nat-gateway-c9177e710af6#:~:text=Internet%20Gateway%20(IGW)%20allows%20instances,IPs%20to%20access%20the%20internet.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

II. Unofficial

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

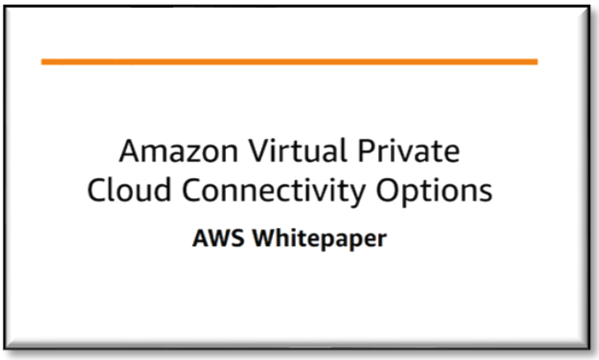
IV. General

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

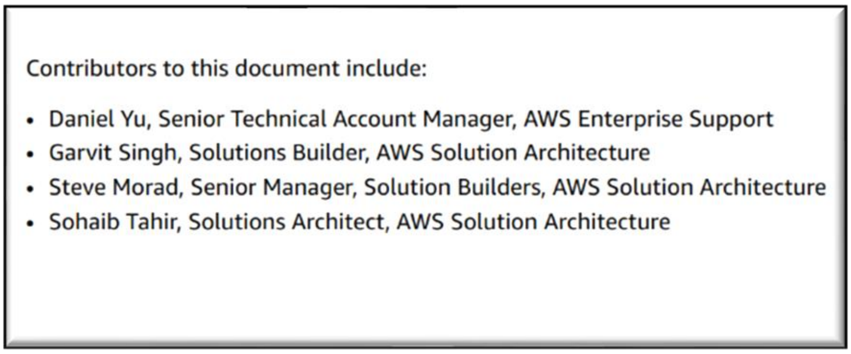
Module 3

Connect your VPC



Amazon Virtual Private
Cloud Connectivity Options

AWS Whitepaper



Contributors to this document include:

- Daniel Yu, Senior Technical Account Manager, AWS Enterprise Support
- Garvit Singh, Solutions Builder, AWS Solution Architecture
- Steve Morad, Senior Manager, Solution Builders, AWS Solution Architecture
- Sohaib Tahir, Solutions Architect, AWS Solution Architecture

TPN

51. ***Phenomenon1*** – the tendency of X to Y.
52. ***Phen2*** – the tendency of X to Y.
53. ***Phen3*** – the tendency of X to Y.
54. ***Phen4*** – the tendency of X to Y.
55. ***Phen5*** – the tendency of X to Y.
56. ***Phen6*** – the tendency of X to Y.
57. ***Phen7*** – the tendency of X to Y.
58. ***Phen8*** – the tendency of X to Y.
59. ***Phen9*** – the tendency of X to Y.
60. ***Phen10*** – the tendency of X to Y.

Glossary

Term1

Description of what term means here.

Term2

Description of what term means here.

Term3

Description of what term means here.

Bibliography

I. Official

https://medium.com/awesome-cloud/aws-vpc-difference-between-internet-gateway-and-nat-gateway-c9177e710af6#:~:text=Internet%20Gateway%20(IGW)%20allows%20instances,IPs%20to%20access%20the%20internet.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

II. Unofficial

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

IV. General

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

Module 4
Monitoring

TPN

61. ***Phenomenon1*** – the tendency of X to Y.
62. ***Phen2*** – the tendency of X to Y.
63. ***Phen3*** – the tendency of X to Y.
64. ***Phen4*** – the tendency of X to Y.
65. ***Phen5*** – the tendency of X to Y.
66. ***Phen6*** – the tendency of X to Y.
67. ***Phen7*** – the tendency of X to Y.
68. ***Phen8*** – the tendency of X to Y.
69. ***Phen9*** – the tendency of X to Y.
70. ***Phen10*** – the tendency of X to Y.

Glossary

Term1
Description of what term means here.

Term2
Description of what term means here.

Term3
Description of what term means here.

Bibliography

I. Official

https://medium.com/awesome-cloud/aws-vpc-difference-between-internet-gateway-and-nat-gateway-c9177e710af6#:~:text=Internet%20Gateway%20(IGW)%20allows%20instances,IPs%20to%20access%20the%20internet.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

II. Unofficial

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

IV. General

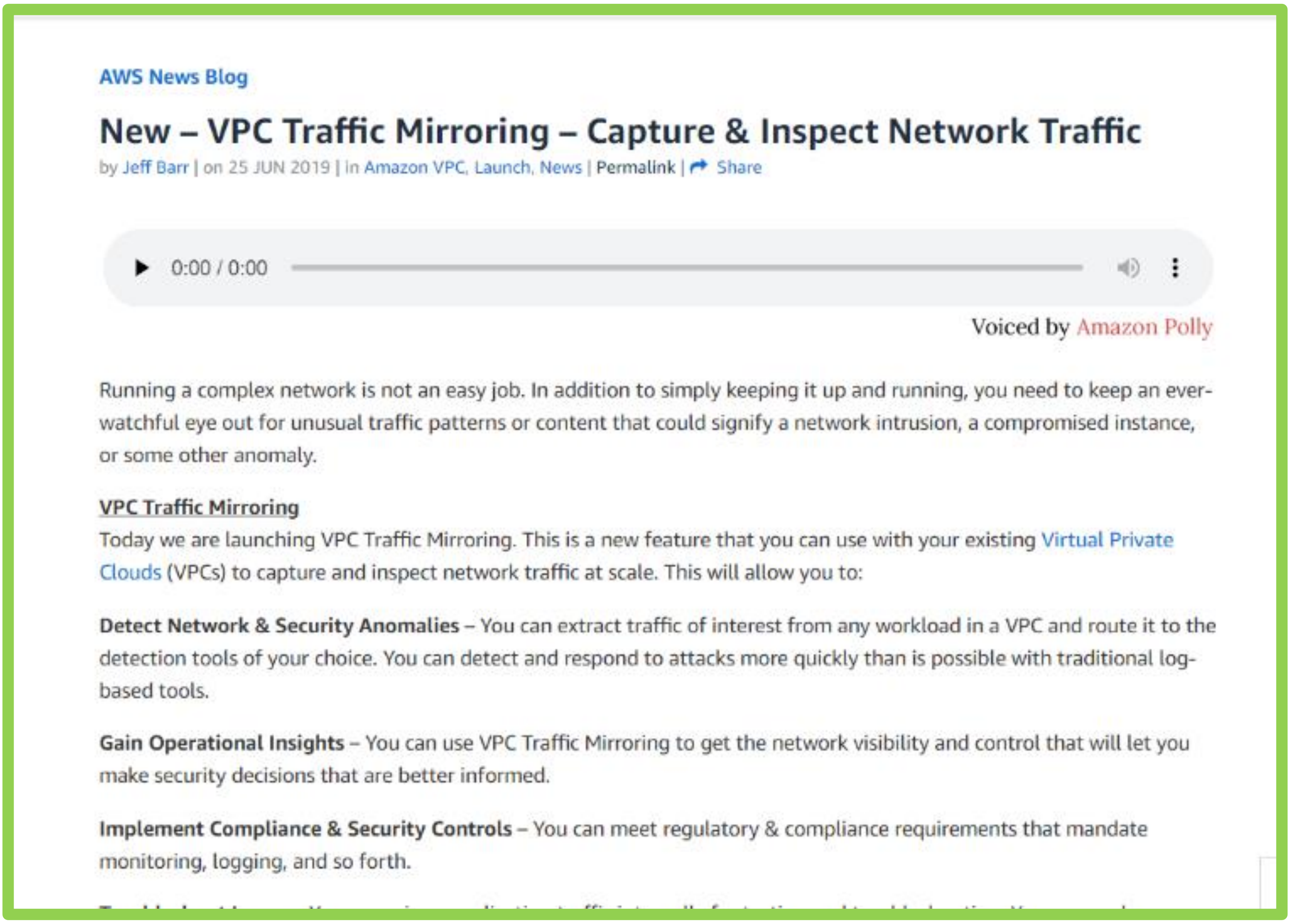
[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

Module 5

Security

Traffic Mirroring



AWS News Blog

New – VPC Traffic Mirroring – Capture & Inspect Network Traffic

by Jeff Barr | on 25 JUN 2019 | in Amazon VPC, Launch, News | Permalink | Share

Voiced by Amazon Polly

Running a complex network is not an easy job. In addition to simply keeping it up and running, you need to keep an ever-watchful eye out for unusual traffic patterns or content that could signify a network intrusion, a compromised instance, or some other anomaly.

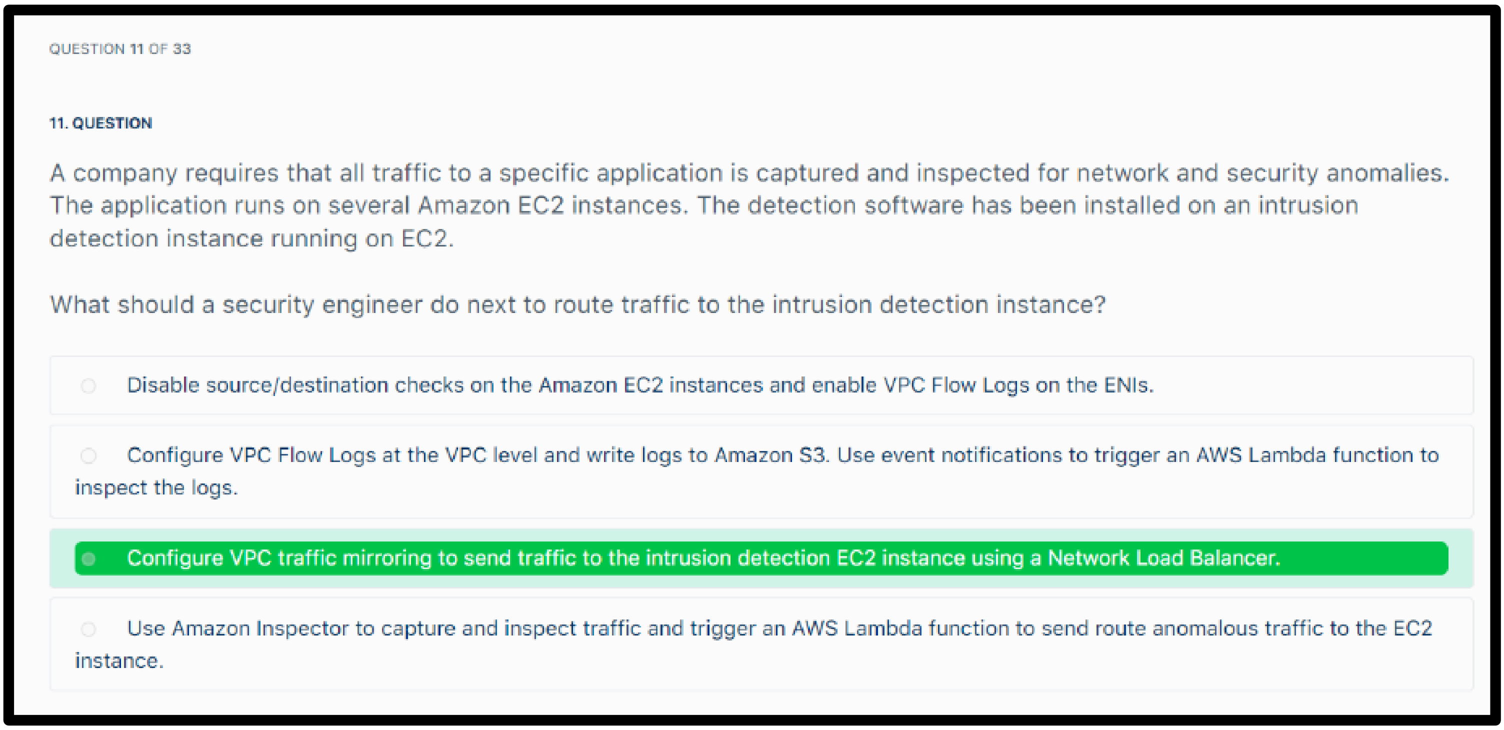
VPC Traffic Mirroring

Today we are launching VPC Traffic Mirroring. This is a new feature that you can use with your existing Virtual Private Clouds (VPCs) to capture and inspect network traffic at scale. This will allow you to:

Detect Network & Security Anomalies – You can extract traffic of interest from any workload in a VPC and route it to the detection tools of your choice. You can detect and respond to attacks more quickly than is possible with traditional log-based tools.

Gain Operational Insights – You can use VPC Traffic Mirroring to get the network visibility and control that will let you make security decisions that are better informed.

Implement Compliance & Security Controls – You can meet regulatory & compliance requirements that mandate monitoring, logging, and so forth.



QUESTION 11 OF 33

11. QUESTION

A company requires that all traffic to a specific application is captured and inspected for network and security anomalies. The application runs on several Amazon EC2 instances. The detection software has been installed on an intrusion detection instance running on EC2.

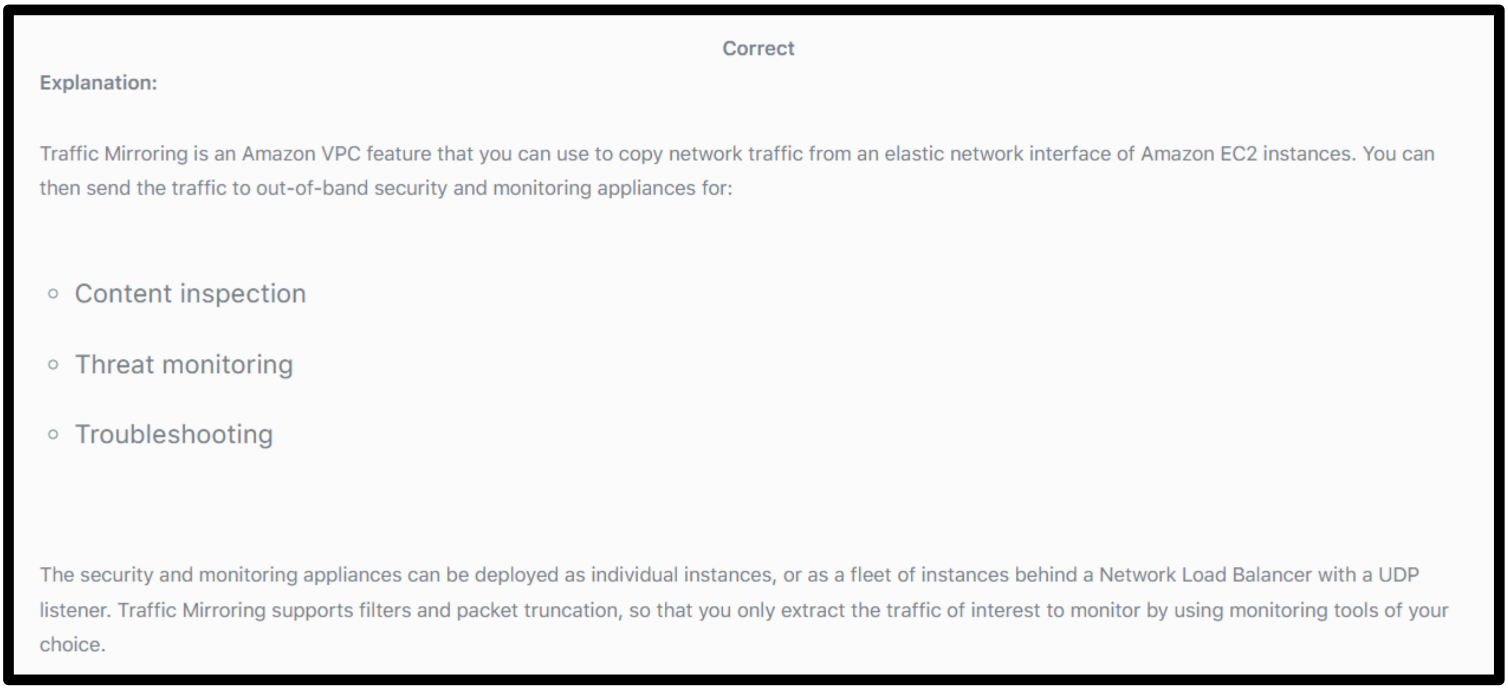
What should a security engineer do next to route traffic to the intrusion detection instance?

○ Disable source/destination checks on the Amazon EC2 instances and enable VPC Flow Logs on the ENIs.

○ Configure VPC Flow Logs at the VPC level and write logs to Amazon S3. Use event notifications to trigger an AWS Lambda function to inspect the logs.

● Configure VPC traffic mirroring to send traffic to the intrusion detection EC2 instance using a Network Load Balancer.

○ Use Amazon Inspector to capture and inspect traffic and trigger an AWS Lambda function to send route anomalous traffic to the EC2 instance.



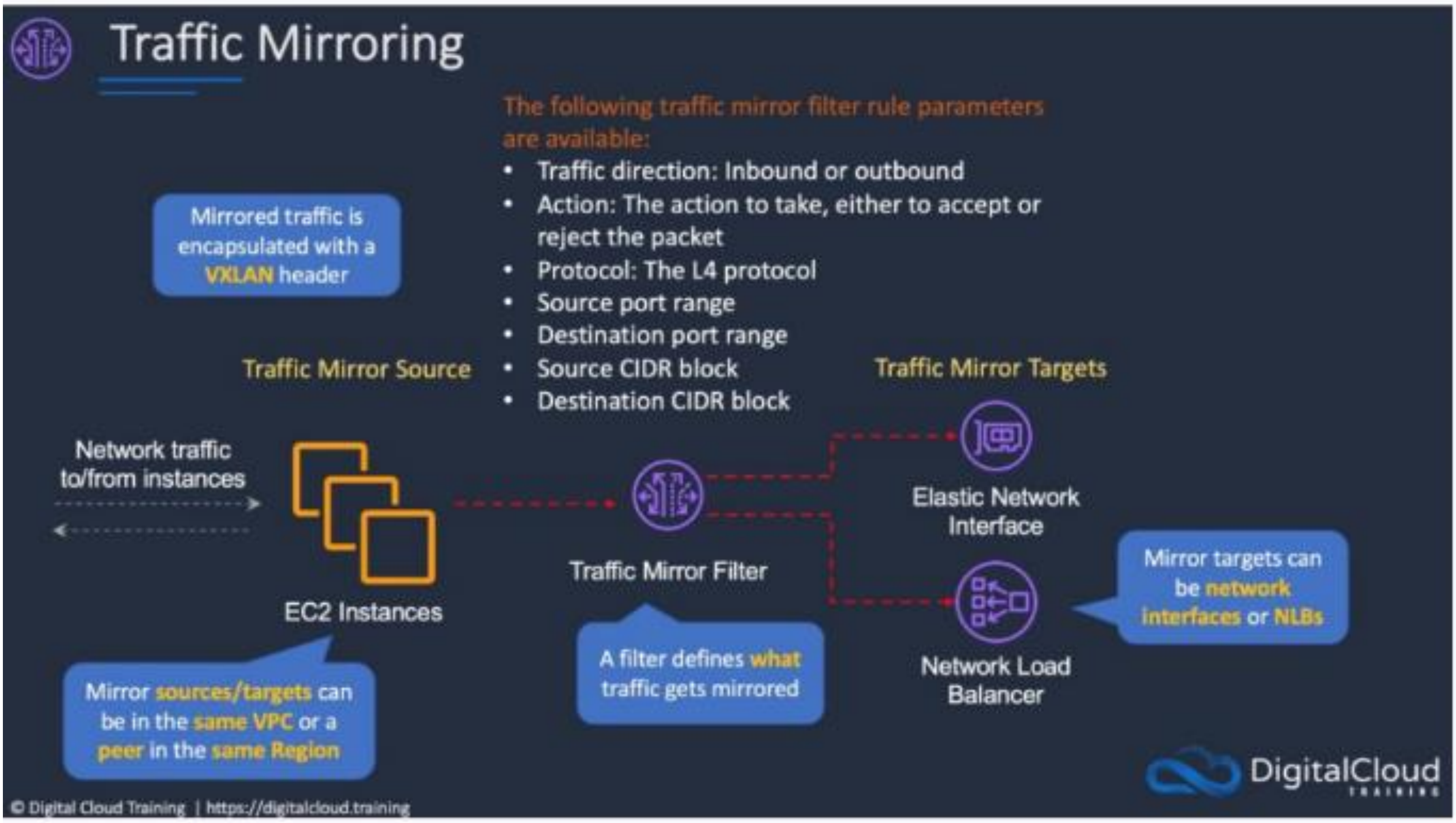
Correct

Explanation:

Traffic Mirroring is an Amazon VPC feature that you can use to copy network traffic from an elastic network interface of Amazon EC2 instances. You can then send the traffic to out-of-band security and monitoring appliances for:

- Content inspection
- Threat monitoring
- Troubleshooting

The security and monitoring appliances can be deployed as individual instances, or as a fleet of instances behind a Network Load Balancer with a UDP listener. Traffic Mirroring supports filters and packet truncation, so that you only extract the traffic of interest to monitor by using monitoring tools of your choice.



Traffic Mirroring

The following traffic mirror filter rule parameters are available:

- Traffic direction: Inbound or outbound
- Action: The action to take, either to accept or reject the packet
- Protocol: The L4 protocol
- Source port range
- Destination port range
- Source CIDR block
- Destination CIDR block

Mirrored traffic is encapsulated with a VXLAN header

Traffic Mirror Source

Network traffic to/from instances

EC2 Instances

Mirror sources/targets can be in the same VPC or a peer in the same Region

Traffic Mirror Filter

A filter defines what traffic gets mirrored

Traffic Mirror Targets

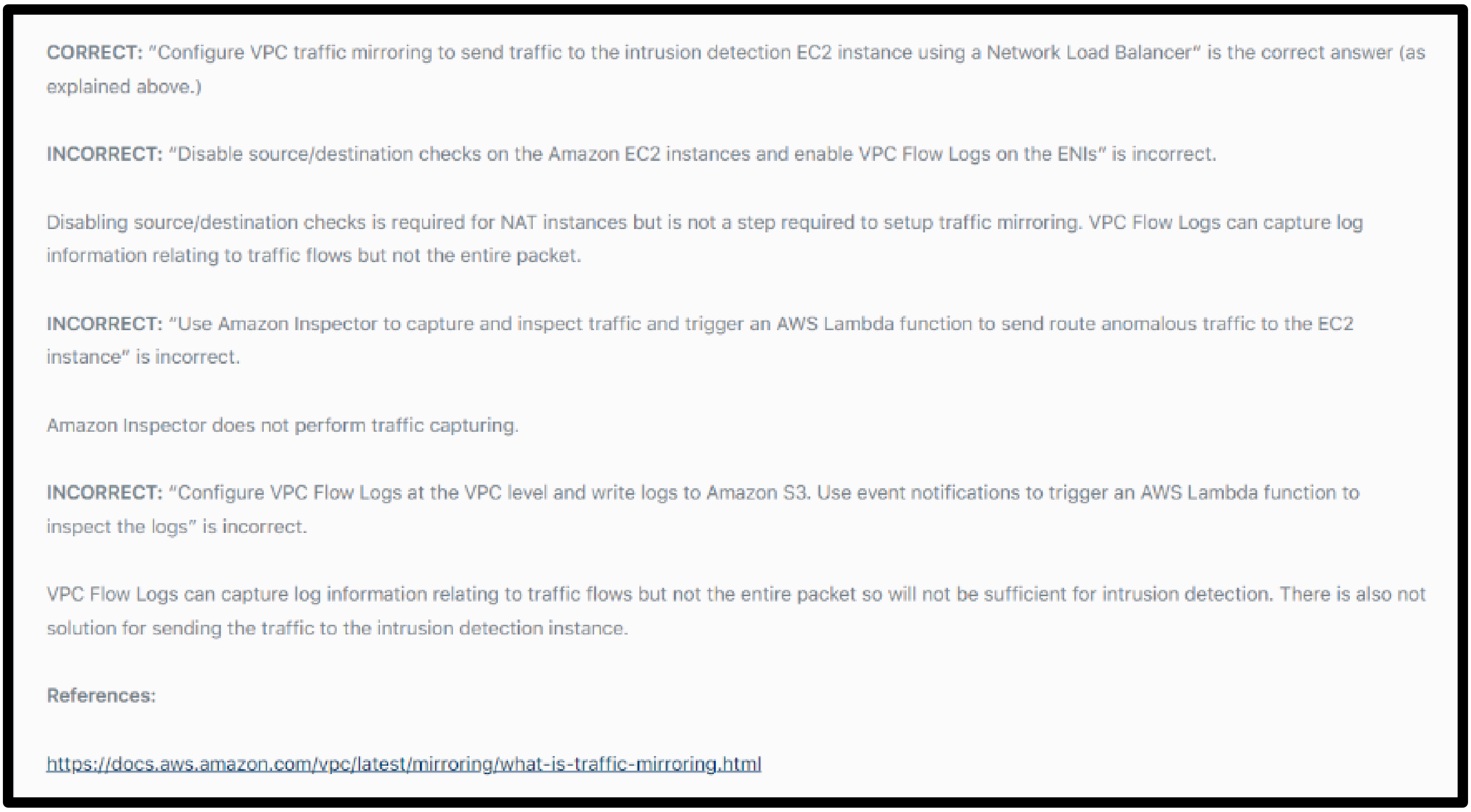
Elastic Network Interface

Network Load Balancer

Mirror targets can be network interfaces or NLBs

DigitalCloud TRAINING

© Digital Cloud Training | https://digitalcloud.training



CORRECT: "Configure VPC traffic mirroring to send traffic to the intrusion detection EC2 instance using a Network Load Balancer" is the correct answer (as explained above.)

INCORRECT: "Disable source/destination checks on the Amazon EC2 instances and enable VPC Flow Logs on the ENIs" is incorrect.

Disabling source/destination checks is required for NAT instances but is not a step required to setup traffic mirroring. VPC Flow Logs can capture log information relating to traffic flows but not the entire packet.

INCORRECT: "Use Amazon Inspector to capture and inspect traffic and trigger an AWS Lambda function to send route anomalous traffic to the EC2 instance" is incorrect.

Amazon Inspector does not perform traffic capturing.

INCORRECT: "Configure VPC Flow Logs at the VPC level and write logs to Amazon S3. Use event notifications to trigger an AWS Lambda function to inspect the logs" is incorrect.

VPC Flow Logs can capture log information relating to traffic flows but not the entire packet so will not be sufficient for intrusion detection. There is also not solution for sending the traffic to the intrusion detection instance.

References:

https://docs.aws.amazon.com/vpc/latest/mirroring/what-is-traffic-mirroring.html

TPN

71. ***Phenomenon1*** – the tendency of X to Y.
72. ***Phen2*** – the tendency of X to Y.
73. ***Phen3*** – the tendency of X to Y.
74. ***Phen4*** – the tendency of X to Y.
75. ***Phen5*** – the tendency of X to Y.
76. ***Phen6*** – the tendency of X to Y.
77. ***Phen7*** – the tendency of X to Y.
78. ***Phen8*** – the tendency of X to Y.
79. ***Phen9*** – the tendency of X to Y.

80. ***Phen10*** – the tendency of X to Y.

Glossary

Term1

Description of what term means here.

Term2

Description of what term means here.

Term3

Description of what term means here.

Bibliography

I. Official
II. Unofficial
III. Critical
IV. General

I. Official

https://medium.com/awesome-cloud/aws-vpc-difference-between-internet-gateway-and-nat-gateway-c9177e710af6#:~:text=Internet%20Gateway%20(IGW)%20allows%20instances,IPs%20to%20access%20the%20internet.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

II. Unofficial

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

IV. General

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

Pre-configuration versus Essence

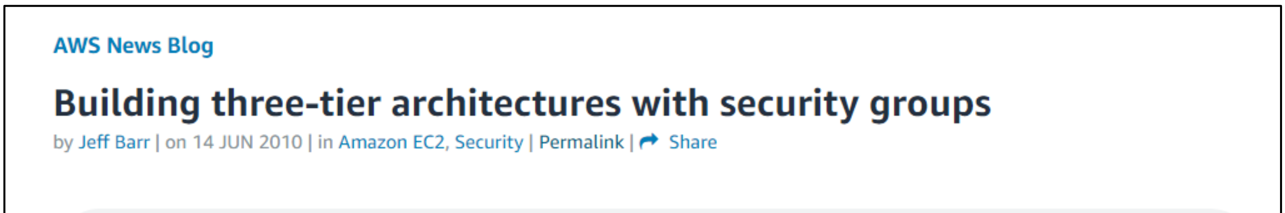
Divide talk into two parts. First part is called *WHAT IS THE CASE*

Second part is called *WHAT IS NECESSARILY THE CASE*.

Say that before the second part you need to go away. Try some practice exams. Forget which firewall has which properties. Work up an appetite for the second video.

There are two kinds of firewalls on AWS. The first is called a NACL and the second a Security Group. NACL just stands for Network Access Control List.

Jeff Barr wrote an article in June 2010:



AWS News Blog

Building three-tier architectures with security groups

by Jeff Barr | on 14 JUN 2010 | in Amazon EC2, Security | Permalink | Share

The article's focus is building a three tier architecture. This is not our focus. However, it is helpful for us that Jeff Barr gives a recap of Security Groups, writing:

> A security group is a semi-stateful firewall (more on this in a moment) that contains one or more rules defining which traffic is permitted into an instance. Rules contain the following elements:
>
> - The permitted protocol (TCP or UDP)
> - The permitted *destination* port range (more on this in a moment, too)
> - The permitted *source* IP address range or *originating* security group

SGs are rarely described as "semi stateful" in educational materials[1].

SPOT THE DIFFERENCE

It's very tempting to think that the rules in a Security Group reference domain names. For example, you have a group of instances in one SG. They are put in front of a network load balancer. Surely, you put the domain name

[1] Piper and Clinton write that 'a security group acts as a stateful firewall' (2021: 99).

of the NLB in the rules. But this is wrong: we put *IP addresses* in the rules of Security Groups.

ENSURE SUFFICIENT PROTECTION

It's your job as a Solutions Architect to *ensure sufficient protection*. When you're learning about AWS, it's common to go through THREE differentiating features of SGs and NACLs:

1. **EFFECT**

   NACLs can have ALLOW and DENY RULES; SGs only have ALLOW rules ("rules as relaxers*"*).

2. **STATEFULNESS**

   SGs are stateful; NACLs are stateless.

3. **PROTECTEE**

   SGs protect instances[2]; NACLs protect subnets.

ESP (*Ensure Sufficient Protection*) stands for Effect, Statefulness and Protectee. Regarding (3) I want to add a note about why it is that these two protectors are able to find employment.

I will mark the following words of Piper and Clinton (2021: 98) as E1:

> **(E1)** Every ENI **must** have at least one security group associated with it.

[2] The most important thing is instances; Ashish Patel's famous article rightly only mentions *instance*. But they also protect 'elastic load balancer listeners' (Piper and Clinton 2021 p323).

It is, however, important to point out that SGs cannot be associated with *network* load balancers. Neal Davis's Digital Cloud Training has a mock exam question for SCS-01 that tests this subtlety.

So, ENIs *demand* security groups. This is how SGs gain employment. I believe this makes it acceptable to hold (in informal contexts, at least) that:

> Every *EC2 instance* must have at least one security group.

This follows from another principle: an EC2 instance must have an ENI. Just like instances themselves, SGs must exist in a VPC. (An interesting question is *Can an SG exist with zero instances as part of it?)*

Just as ENIs *demand* a Security Group, subnets *demand* a NACL. The VPC user guide states:

> **(E2)** Each subnet in your VPC **must** be associated with a [NACL].

As well this issue about demanded association, we can ask two further questions about each protector:

> **CAN IT MOONLIGHT**? Can the firewall have more than one protectee?
>
> **CAN IT BE A TEAM PLAYER?** Can more than one firewall protect a particular protectee?

Security Groups can certainly moonlight. I believe this explains the name; multiple instances can be united into one security group.

Security Groups can also be team players. In other words, an SG can be part of a *team* of SGs protecting an instance[3]. The following question needs to be answered:

> Does an instance need to have multiple ENIs to achieve this?

[3] There was recently a discussion about this on a forum. It raises a question over how the combined set of rules is assessed. Crucially:

> If there is more than one rule for a specific port, we apply the most permissive rule. For example, if you have a rule that allows access to TCP port 22 (SSH) from IP address 203.0.113.1 and another rule that allows access to TCP port 22 from everyone, everyone has access to TCP port 22.

See https://serverfault.com/questions/483938/multiple-ec2-security-groups-permissive-or-restrictive

The user guide for VPC tells us:

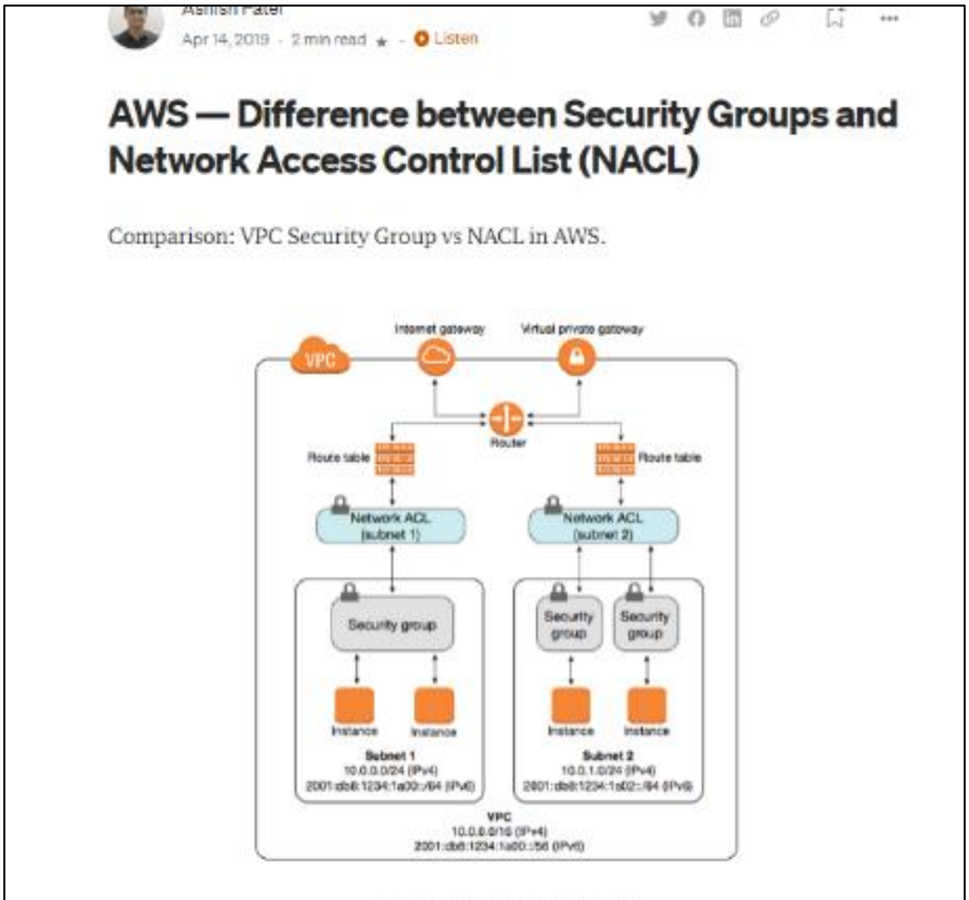
> You can associate a network ACL with multiple subnets.
>
> However, a subnet can be associated with only one network ACL at a time. When you associate a network ACL with a subnet, the previous association is removed.

So, NACLs can moonlight but cannot be team players. A NACL can be responsible for many subnets. If it is, it necessarily is the *only* NACL for all its subnets. A NACL never has to get along with other NACLs on the protection team for a subnet.

Moonlighting is done by the **m**asses; **t**eam-playing is more… **t**ouchy. *Only* SGs can form protection teams for single instances. Put another way:

> A subnet can have only one NACL associated with it.

Piper and Clinton 2021: 101



Ashish Patel

Apr 14, 2019 · 2 min read ★ · Listen

AWS — Difference between Security Groups and Network Access Control List (NACL)

Comparison: VPC Security Group vs NACL in AWS.

Unfortunately, these features can appear arbitrary. For example, *it just so happens* that SGs are stateful. We therefore have to remember the three features by rote, and a student is to—essentially—get disciplined by practice questions until they finally stick.

The point of this essay is to elucidate the Features. Elucidation involves reminding the reader of certain things, at the right time, such that the Features no longer appear arbitrary. They appear necessary. Following the elucidation, the problem of memorising the Features will be dissolved.

THOUGHT NO. 1 – SUBNETS ARE NOT RESOURCES

Consider what it is that each firewall protects. Security groups protect instances. These are virtual servers – they can be spun up and terminated easily in AWS. We can attach EBS volumes to instances; we can put databases on them; instances can constitute the WEB TIER of a three-tier stack; instances possess IPv4 addresses.

What is a subnet? I have friends who explain subnets using an analogy. Just as you slice up a pizza, you can slice up a CIDR block. A CIDR block is a range of IP addresses, and you're given one with every VPC. To make this CIDR block more manageable, we slice it up—into subnets.

Subnet—subnetwork—is a range of *potential* IP addresses. When a subnet comes into existence, it's not the case that a bunch of ACTUAL instances pop up. There *are* now, though, POTENTIALLY instances within that subnet.

Instances are obligate parasites—they obliged to inhabit subnets. It's impossible for instances to float free. To ask "have you got round to putting that instance in a subnet yet?" betrays a complete misunderstanding of the concept of an instance.

So, instances are substantial things. Subnetworks are not. Subnets are ranges of IP addresses—they are ranges of POTENTIAL instances.

First, let's think about **(3) PROTECTEE**. In one sense, there's not even a question to pose here. There's going to be *some* firewall existing on the instance. AWS must call it *something*. We might say:

> Which firewall is called which, is just an accidental feature. *It is what it is*. The one on the instances is called a Security Group. Accept it.

And if we ask, "why do Security Groups protect instances?" this is trivial. The answer is: that's what it ***means*** to be a security group. AWS provided a stipulative definition of what Security Groups are.

That being so, there's still room to ask "why did they choose this name over other names?" Why *SECURITY GROUP?* When you could

a) …mirror *Network Access Control List*, opting for *Instance* Access Control List.

b) …call it an *EC2 Firewall*.

c) …avoid calling something that is one single thing, a GROUP!

Another reason I ask the first question is that ACLs traditionally have objects, or **resources**, as their protectee. For example, in Active Directory, you have ACLs on **files** and **directories**. An instance has greater claim to being a RESOURCE than a subnet. So, if ACLs protect resources, the firewall at the *instance* level should be called an ACL.

These are all quibbles about the necessity of **(3) PROTECTEE.** So, let's address this first.

We can address (c) immediately. Group of what? Group of *instances*. *Many* instances can be attached to an SG. To say that we have the same rules governing a set of instances, it's very helpful to say they're a group. We could even say the instances are a *security* group(!)[4]. The term "security group" is thus brutally minimalistic.

NACLs protect subnets, which are more like containers. We can think of them as being larger than an instance. Grouping subnets according to security configuration is… not something that interests us. So, we have *zero* reason to call this firewall a Group. No grouping of subnets is possible or desirable.

We've answered c now. We haven't properly addressed (a) yet since we haven't given a reason for not having the names mirror one another. (a) and (b) will be answered in the course of our next topic: why it has to be *SGs* which only have allow-rules.

[4] Frankly, I'm not certain whether it is so-called because we have a group of *instances* or a group of *rules*. It suits us here to consider it the former.

Now we zoom up to **(1) EFFECT**. One of the firewalls has *BOTH* allow and deny rules. This is the one we'll call an Access Control List. Why?

Firstly, you're **CONTROLLING** access. And CONTROL involves the ability to reach for an ALLOW or DENY rule. So, to really be an Access ***Control*** List is to have both options in your armoury. (This is a nice heuristic at least…)

Traditionally, ACLs were on **objects** such as files or directories. They use ACLs in Active Directory. The question was *Which users are allowed to access this file?* In fact, it's strange, on reflection, that AWS are taking a SUBNET and treating it as if it's an object.

With those traditional ACLs on objects, there's no traffic involved. It's purely about objects accessible, or perceivable, by a user. Network traffic doesn't come into it at all. If "ACL" has no suggestion of traffic, it certainly has no suggestion of being STATEFUL. And this property of "ACL" is **extremely helpful** for AWS. "ACL" is an *excellent* term for these outer firewalls. This is because these outer firewalls are not stateful.

The term "ACL" flourishes for a second reason. In the traditional world, with ACLs and OBJECTS, you had this capability:

Distinguishing for Denial (DfD)

The capacity to distinguish one user and deny them access.

We need a term that implies DfD. That term is: ACL. It happens to be the case that Security Groups *cannot* DfD. You literally cannot create deny-rules on Security Groups. So, you *certainly* cannot DfD.

Instances are resources. Yet we simply cannot call their firewall an ACL. *That* would imply a DfD capability. And this is a capability Security Groups lack.

We've dealt with (1) EFFECT at this point. The firewall with CONTROL in its name has the choice: of ALLOW or DENY rules. And consider the firewall that has ONLY allow-rules. Because this neuters the firewalls of a DfD capacity, it would be irresponsible to call it an ACL. So, it cannot be the NACL that has ONLY allow-rules.

Notice how we've covered **(3) PROTECTEE also.** We're progressively making the **NECESSITY** of ESP more vivid. One kind of firewall lacks DENY rules. Such a firewall simply *cannot* be called an ACL. This is because (1) ACLs are user-centric and they confer a DfD capacity on those who use them. (2) A**C**Ls provide the full breadth of control (ALLOW ***and*** DENY) for that user. Clearly, to denote those inner, intelligent instance-firewalls as "ACLs" is **inconceivable**.

Now let's move onto the second item in **ESP**, **(2) STATEFULNESS**. Security Groups are stateful. But *why*? Why does *this* firewall **have to be** the one that is stateful?

Stateful firewalls monitor the state of open network connections. It is *sometimes* said that the firewall is intelligent or has a memory. It's usually put in terms of **keeping track** of connections.[5]

If a firewall is **stateful**, we don't need **bountiful bounded rules**. Stateful firewalls keep it neat and tidy. They are smart and efficient. For example, suppose you add a rule that allows INBOUND traffic of a particular sort. You naturally then think: *now I must add a rule for the outbound traffic*. The security group says "**No, relax**". It's taken care of.

It works the other way too. Suppose you add a rule allowing *OUTBOUND* traffic. You then think "I must let the traffic back in". **No, relax**. You don't. The firewall's being stateful means that REPLY TRAFFIC is taken care of.

Do not get confused. What we're talking about here is not

> *allowing some particular kind of traffic in both directions.*

To an observer, it might look as if this is all we are doing. But what you cannot observe is that the first traffic is a REQUEST and the second a RESPONSE. These concepts are indispensable to understanding statefulness. If you just think in terms of the direction of traffic, without contemplating its nature, you will **fail** to capture statefulness.

[5] For example, the entry for Stateful Firewall on Wikipedia says 'a stateful firewall… individually tracks sessions of network connections traversing it'.

Consider an analogy. I order a book on Amazon.co.uk asking for it to be delivered through my letter box. The next day or so, someone shoves an envelope into my house. This is acceptable.

Consider if we reverse the order of events: first they put a leaflet through my door, and then I'm supposed to go on their website. No! That's cold calling. The change in the order of events changes everything. The identity of the initiator of the communication changes everything.

Talk of REQUEST and RESPONSE is indispensable to understanding statefulness. To an observer, it's merely bidirectional communication—but it's more than this. We might say that Statefulness consists in this:

Statefulness (1)

If we ALLOW [specific traffic *x*] to request something,

then we'll automatically make arrangements for the **REPLY TRAFFIC**.

What happens is that we define some positive event. This is an **initial request**, which is permitted. Then, we make arrangements for a follow-up to the event.

That initial request is crucial. Statefulness is **not** merely making arrangements for reply traffic to:

> *any* requests that come in

for example. Statefulness is in fact making arrangements for reply traffic to

> requests we've allowed

This means that the **initial request**—the thing we arrange the follow up for—must not be out the blue. The initial request must be somewhat *expected*—it must be defined in a rule.

If you think about it, the defining rule must be an ALLOW rule. If it was a deny rule, nothing happens. So, there's nothing for which to arrange the follow-up for. Therefore, the initial request must be described in an allow rule.

Recall **Statefulness (1)**. The first part stated:

> **If** we allow [specific traffic x]

However, we don't tend to selectively allow using NACLs. By 'selectively allow' I simply mean pinpointing entities, to allow them in. (This is the opposite of the DfD capacity.) NACLs certainly *can do*—NACLs have ALLOW rules for us to utilise. This gives NACLs a remarkable adaptability. They can DfD and DfA.

I believe, however, that **using NACLs to DfA is problematic**. I am not yet fully able to express why, however I believe that it is conventional to use NACLs to DfD. My hypotheses about why this convention exists involve the speculations that

(1) one subnet is likely to contain *multiple* instances.

> Collectively, the instances require a large amount (variety?) of traffic to be allowed into the subnet. Thus, DfA-ing would be inefficient and cumbersome.

(2) Duplicating the DfA strategy on two, encapsulating firewalls, is pointless.

> Plus, it means we have to re-configure two firewalls every time we need to re-configure.

(3) If we used a DfA strategy on the outer firewall (the NACL) at all, administrators might come to rely on it.

> They might allow their SGs to be extremely permissive, safe in the knowledge that the NACL blanket-bans traffic.

(4) One item of traffic is harmless to one instance in the subnet while being harmful to the other.

> I don't know it this is possible, especially as requests tend to be addressed to specific servers. Packets, or

> whatever we're talking about here, don't "go for a gander" around the subnet. Anyhow, perhaps there's something in this.

If it *is* the case that using NACLs to DfA is problematic, then we won't be defining ALLOW rules. Therefore, we won't be

> ALLOWing [specific traffic *x*] to request something

And if we're not doing *that*, then there's no point in having a capability to make arrangements for the reply traffic. In other words, there's no point in *being stateful*. In stark contrast, SGs *must* DfA. They *must* "selectively allow" and so this is the *only* firewall where statefulness makes sense.

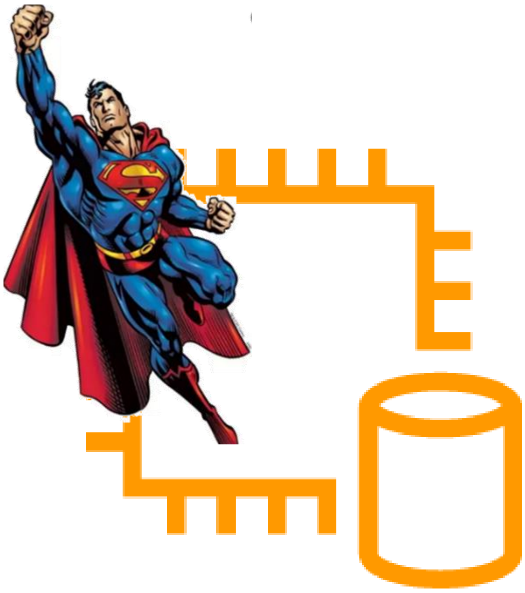
NACL NAIVITE

At some point in your AWS journey, you'll be hit with NACL Naiveté. This involves a specific belief about NACLs, which is false. It creeps in by considering this:

> (1) If a firewall has **ALLOW exclusivity**
>
> then it blocks traffic by default.

That looks compelling. Really, it's only half-true with SGs. AWS configure their instances to *CAPTURE THE DAY* and exert themselves on the world.

Their SGs configured with a bias for ***ACTION***. We won't be lounging about mulling over advice: the list of inbound rules is *wiped clean*. And there's a NO-HOLDING BACK outbound rule. *Get out there!*



So, if a firewall possesses **ALLOW exclusivity**, it doesn't necessarily block traffic by default. Consider SGs. They allow all outbound traffic.

However, the conditional above (1) is true for the inbound rules of SGs. For that reason, the conditional has some truth. The conditional therefore pops up again and again, and it leads to the following thought.

(2) If a firewall lacks **ALLOW exclusivity**,

then it doesn't block traffic by default.

This appears to be reasonable. It is in some ways the opposite of (1).

Well, NACLs lack allow exclusivity. Therefore, it must be the case that they do not block traffic by default. That is, they're very permissive. *This* is **NACL Naiveté**.

However, the jump from (1) to (2) is a fallacy. Just because hosepipes result in puddles doesn't mean that a lack of hosepipes results in a lack of puddles. The jump was like this:

(1) $P \rightarrow Q$
(2) $\neg P \rightarrow \neg Q$

P is "allow exclusivity" and Q is "denying by default".

However, NACLs *don't* lack the property that is

Denying by default.

It turns out there are *many* ways to skin the "deny by default" cat. There are many ways to achieve "default denial of traffic".

To explain how NACLs achieve it, let me quickly mention two bizarre things.

First, every single NACL contains something remarkable. Tucked away at the bottom of the list of rules is an ***INDESTRUCTIBLE*** RULE. Like some sort of horcrux, it just list at the bottom of the list eternally. Oh, and it ***DENIES***. Everything.

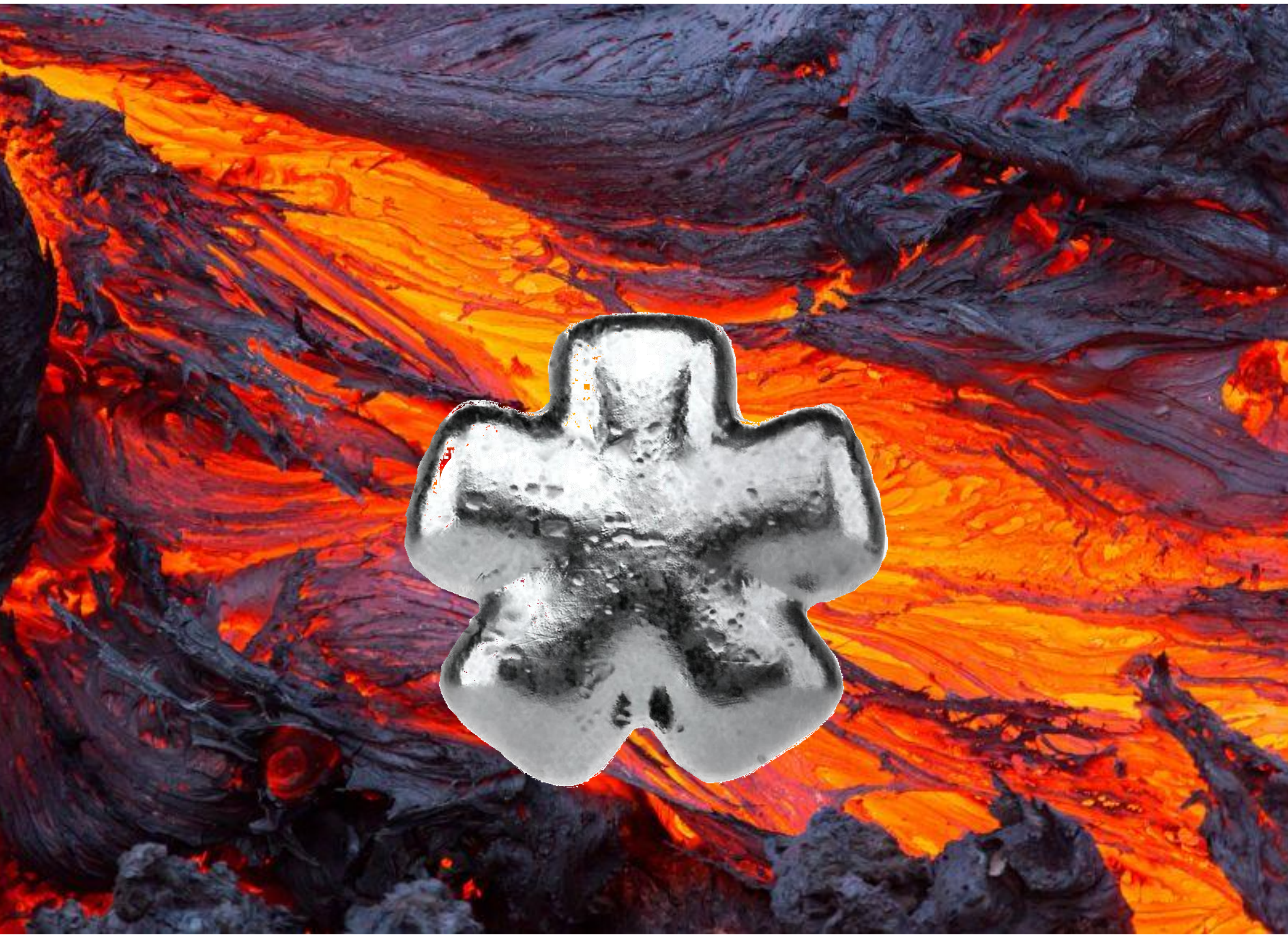


Figure 15b. A captive Default Deny (colourised). So-called "default domestication" was made illegal in March 2013.

To this day, nobody has been able to spot a NACL lacking a Default Deny. So, that's the first bizarre thing. All NACLs contain a Default Deny rule, denoted by an asterisk.

The **second bizarre thing** is this. If you poke around your VPC while it still has that new-VPC-smell, you'll notice a NACL sat in the corner.

That's right—AWS kindly pre-install a NACL ready to be used. This is very nice of them, because it means that if you launch any subnets, these subnets can go to this NACL in order to fulfil their DUTY to wear a NACL. You're able to get things going quickly.

And what AWS kindly do is essentially "cover up" the default deny rule. They do this by inserting an ALLOW rule. This ALLOW rule overrides the *DEFAULT DENY* and allows all traffic into and out of the subnet.

The scary DEFAULT DENY still exists but it's been covered up. There is a wide-ranging ALLOW rule laying over the top of it, preventing it from being effective. I call this rule that goes over the top the **Apron Allow**.

The Apron Allow means that the pre-installed NACL is very permissive indeed. It allows everything in and out of the subnet. I've now explained the two bizarre things: (1) the *DEFAULT DENY* and the pre-installed, default NACL.

The NACL being pre-configured in this way causes many educational AWS resources to simply state:

> NACLs ***are*** default allow.

This is cute because it's the exact opposite of SGs, which are default-deny. But it's not necessarily true.

You might create a NACL, to sit alongside the default one. If you do, things won't quite be the same. The *DEFAULT DENY* rule will certainly lurk at the bottom of your created NACL.

But they'll be no Apron Allow. This means the indestructible *DEFAULT DENY* will not be overridden. Its negative power will be unleashed unto the world. The NACL will be more like default deny.

We can now see why it is misleading to describing NACLs as default-allow.

The Apron-Allow rule is a luxury reserved for the *default* NACL. Created—or **custom**—NACLs will not contain it.

I told you there were many ways to skin the "default deny" cat. One way is to eliminate the very concept of a deny rule (as SGs do); another way is to have a *DEFAULT DENY* (as NACLs do). Don't fall prey to NACL **NACL Naiveté**.

Two Explanatory Forces

With the two AWS firewalls, there are two great explanatory forces.

It's helpful to draw a distinction between the ESSENCE of the firewall and the way it's been pre-configured.

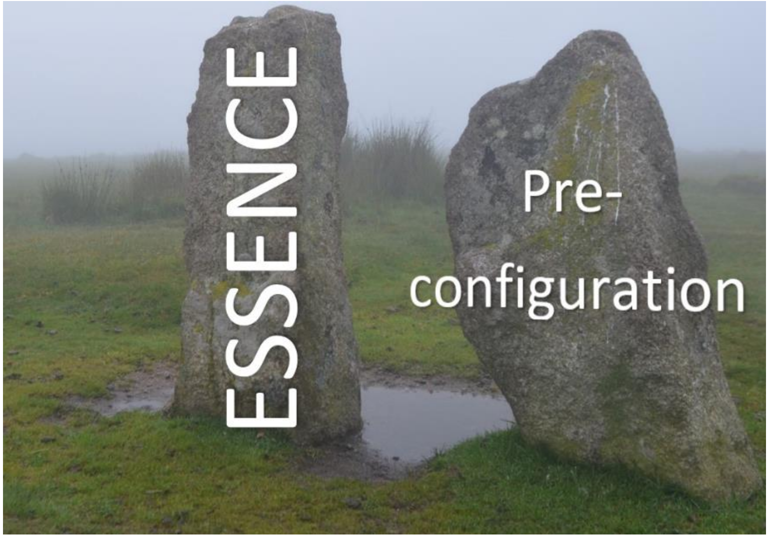
A feature is essential if you cannot remove it through configuration. For example, allow-exclusivity is essential to SGs. Allow-exclusivity is what is *means* to be an SG.

Some features, on the other hand, arise from the way the firewall has been configured—the way it happens to have been set up. Sometimes, it really is the firewall's pre-configuration that does the explanatory work. We've seen two examples of this. NACLs are given the Apron Allow rule, and SGs are given a BIAS FOR ACTION.

I believe it's this distinction that should loom large in your mind: essence versus pre-configuration. They're equally important explanatory forces. For example, sometimes the **Allow-exclusivity** of Security groups explains something. And sometimes the **statelessness** of NACLs explains something. Explains what?

- Explains why you need to choose a particular option on your SAA exam.
- Explains why your architecture needs to be (re-)configured.

Try to ask whether the essential nature of the firewalls is explaining things, or the firewall's pre-configuration.



Most AWS education resources fail to make this distinction at all. Instead, the distinction at the forefront is that between:

- Default allow and
- Default deny.

This is the distinction at the forefront. However, I believe that having this as the fundamental distinction can be misleading. Here is an example of writers using Default-allow and default-deny as the fundamental distinction. Piper and Clinton (2021) write:

> By default, a security group will
>
> - deny all incoming traffic while
> - permitting all outgoing traffic.
>
> p36

The SG *only* permits outgoing traffic because AWS have pre-configured it, installing a BIAS FOR ACTION rule (BfA). The SG certainly does the two things described. It denies so-and-so and allows such-and-such. But the former is achieved by the *absence* of any rules, the latter is achieved by the *presence* of a rule. This is quite an important difference.

Two really quite different senses of "default" are run together here. Suppose you delete the BfA rule. It would follow that it would now be inappropriate to say the SG:

> Permits all outbound traffic *by default.*

Meanwhile, it will *never* cease to be appropriate to say that it:

> Denies all inbound traffic *by default.*

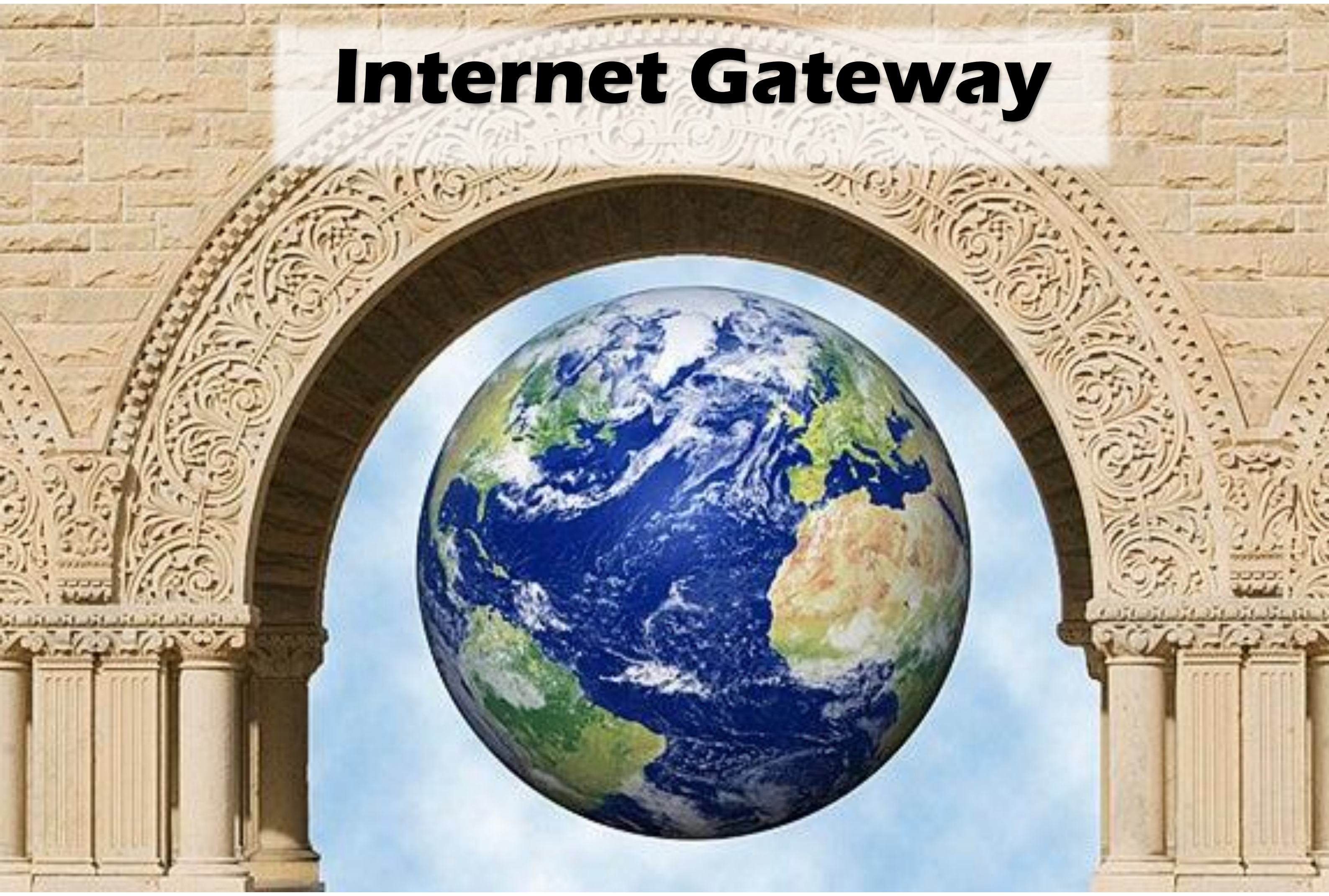
So, the first description ceased to be appropriate quite quickly, the second remains. Note how it would still be appropriate to say the SG "denies inbound traffic by default" *even* if you've added rules that allow in traffic. It remains appropriate because the firewall possesses Allow-exclusivity.

Sure, a new SG will start its life permitting outbound traffic. But that doesn't mean SGs have an ongoing tendency to permit outbound traffic. Meanwhile SGs *do* have an ongoing tendency to deny traffic. So, the sentence above takes some very different properties and treats them as if their on an equal footing. This is just one example of what happens when the fundamental distinction between default-allow and default-deny.

In a NACL rule, you can specify **only a CIDR** as the source or destination. This is unlike a security group rule, for which you can specify **another security group** for the source or destination.

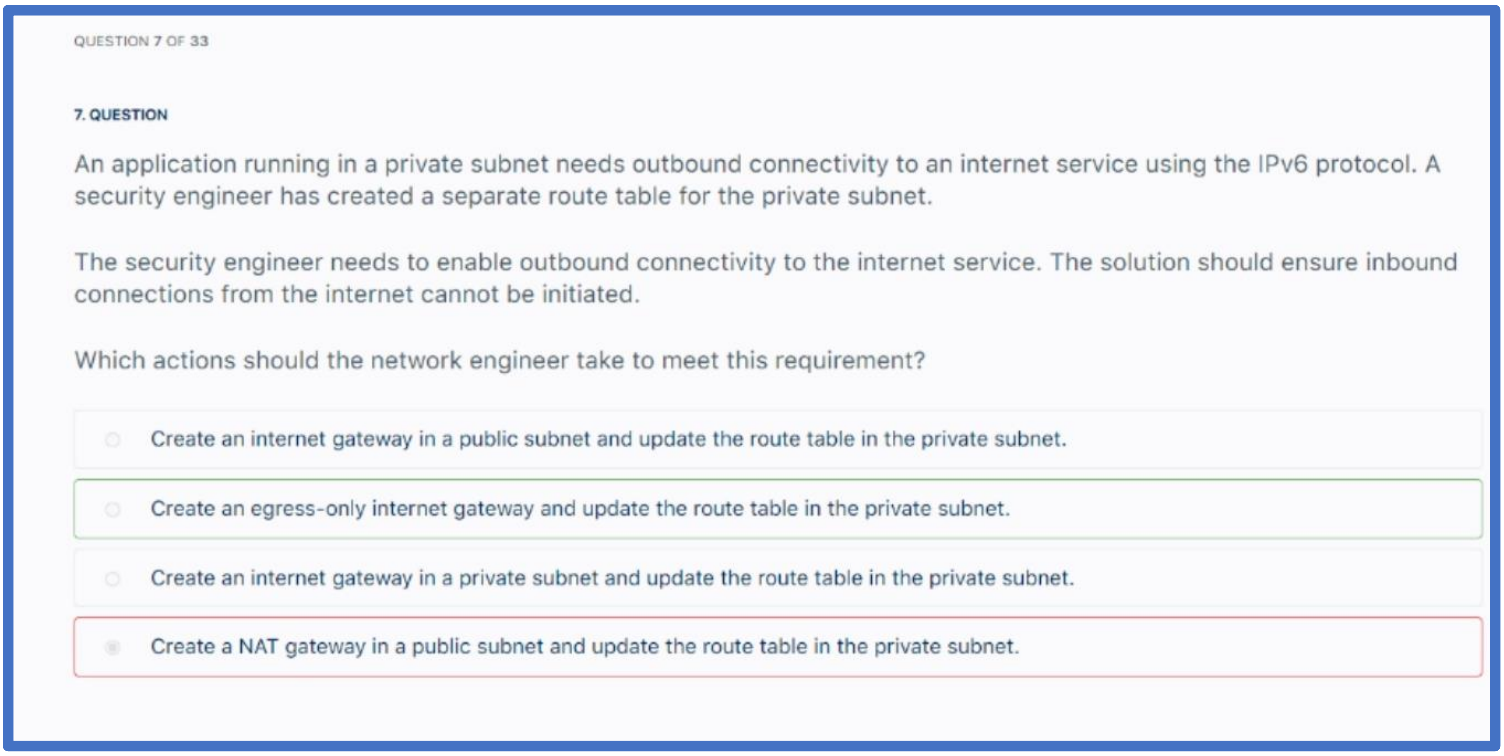
How can we remember this?

Internet Gateway



Egress-only Internet Gateway

Have a look at this exam question:



QUESTION 7 OF 33

7. QUESTION

An application running in a private subnet needs outbound connectivity to an internet service using the IPv6 protocol. A security engineer has created a separate route table for the private subnet.

The security engineer needs to enable outbound connectivity to the internet service. The solution should ensure inbound connections from the internet cannot be initiated.

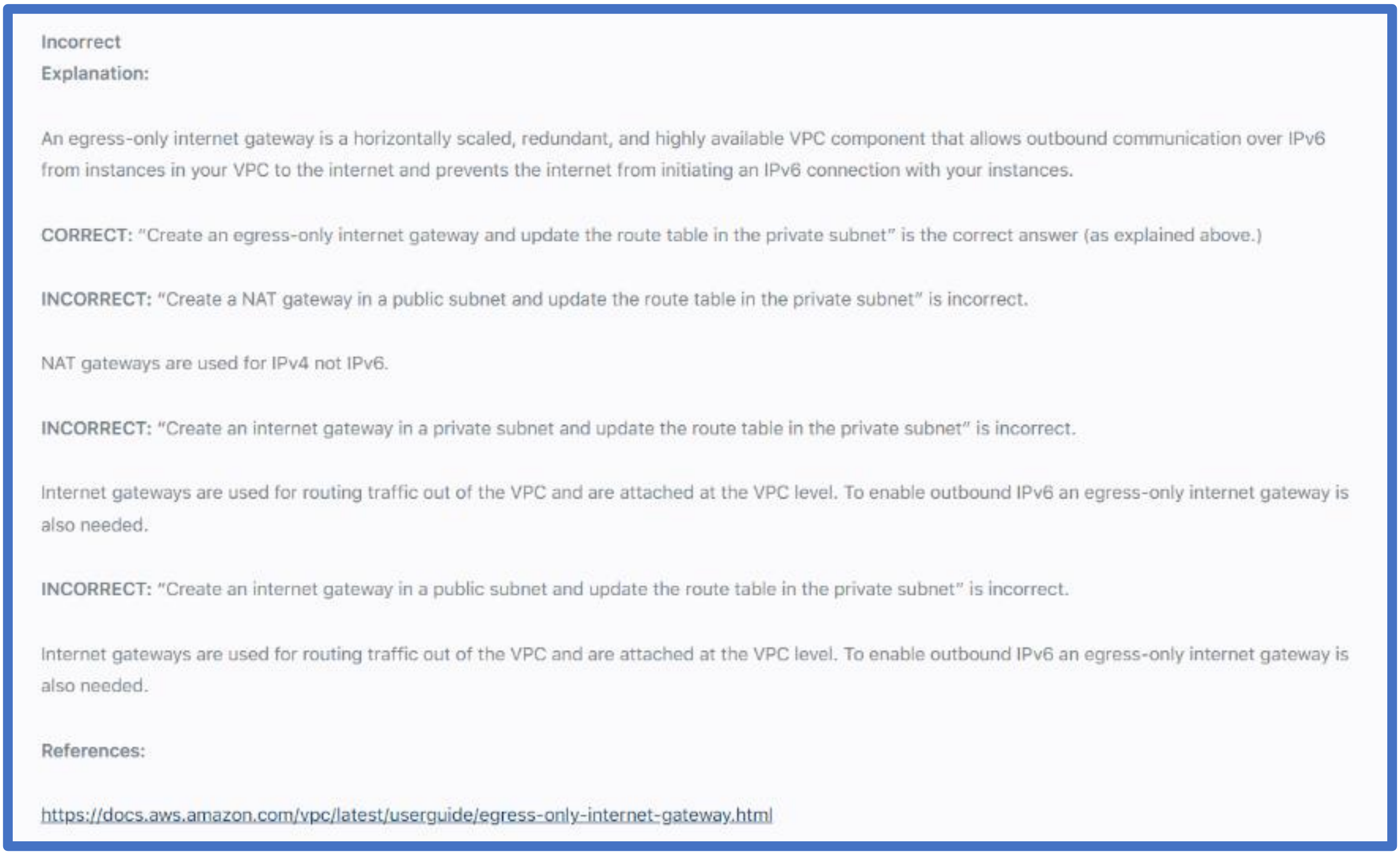
Which actions should the network engineer take to meet this requirement?

○ Create an internet gateway in a public subnet and update the route table in the private subnet.

○ Create an egress-only internet gateway and update the route table in the private subnet.

○ Create an internet gateway in a private subnet and update the route table in the private subnet.

● Create a NAT gateway in a public subnet and update the route table in the private subnet.



Incorrect

Explanation:

An egress-only internet gateway is a horizontally scaled, redundant, and highly available VPC component that allows outbound communication over IPv6 from instances in your VPC to the internet and prevents the internet from initiating an IPv6 connection with your instances.

CORRECT: "Create an egress-only internet gateway and update the route table in the private subnet" is the correct answer (as explained above.)

INCORRECT: "Create a NAT gateway in a public subnet and update the route table in the private subnet" is incorrect.

NAT gateways are used for IPv4 not IPv6.

INCORRECT: "Create an internet gateway in a private subnet and update the route table in the private subnet" is incorrect.

Internet gateways are used for routing traffic out of the VPC and are attached at the VPC level. To enable outbound IPv6 an egress-only internet gateway is also needed.

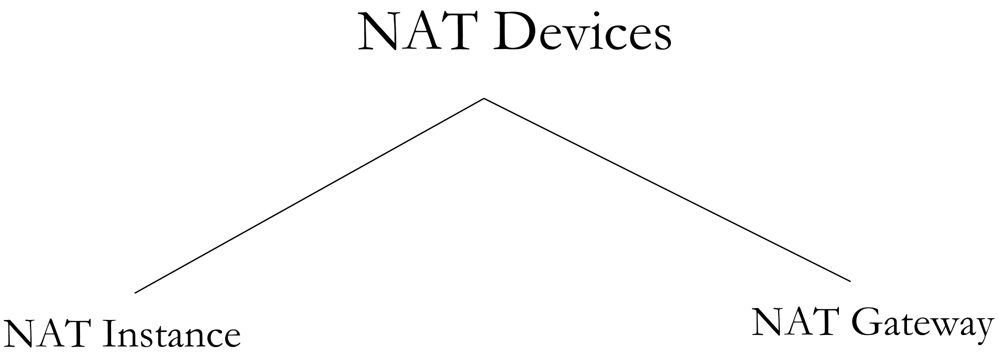
INCORRECT: "Create an internet gateway in a public subnet and update the route table in the private subnet" is incorrect.

Internet gateways are used for routing traffic out of the VPC and are attached at the VPC level. To enable outbound IPv6 an egress-only internet gateway is also needed.

References:

https://docs.aws.amazon.com/vpc/latest/userguide/egress-only-internet-gateway.html

Richard Mortier explaining Network Address Translation (NAT) in this video.



NAT Devices

NAT Instance

NAT Gateway

What is the purpose of an NAT device?

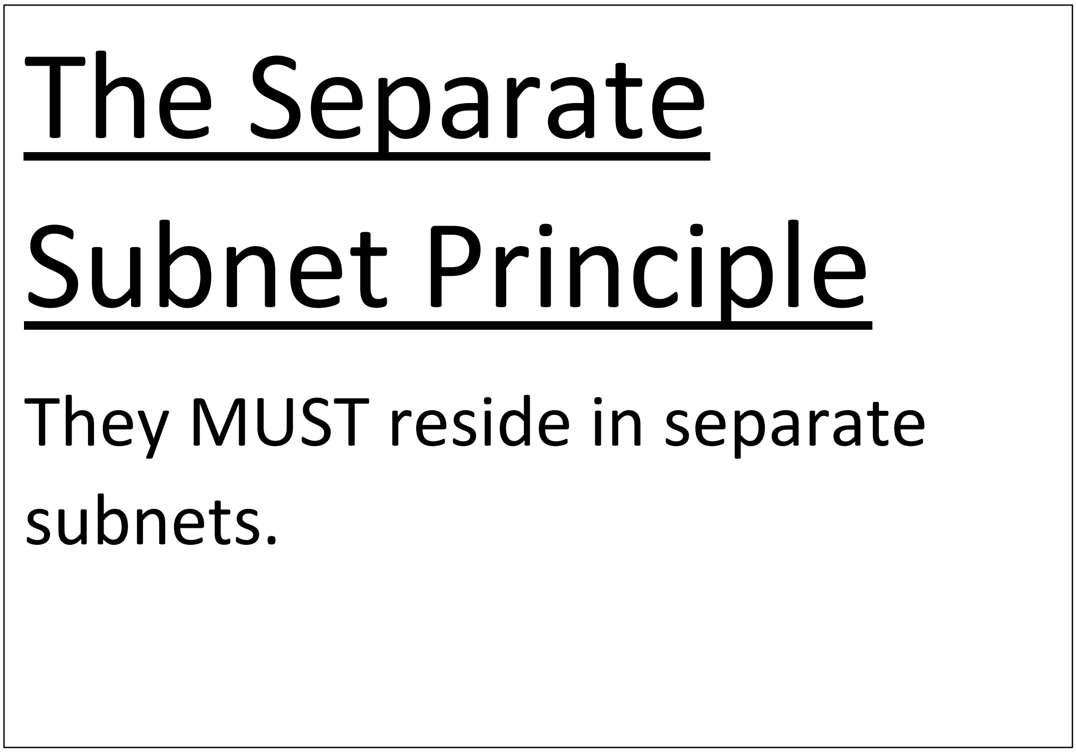
Why?

At first, the private IP address of the NAT device is within the packet.

But eventually, the IGW says “no mate, take that out”. And the public IP of the NAT device is put in instead.

See page 111 of Piper and Clinton.

The NAT Gateway

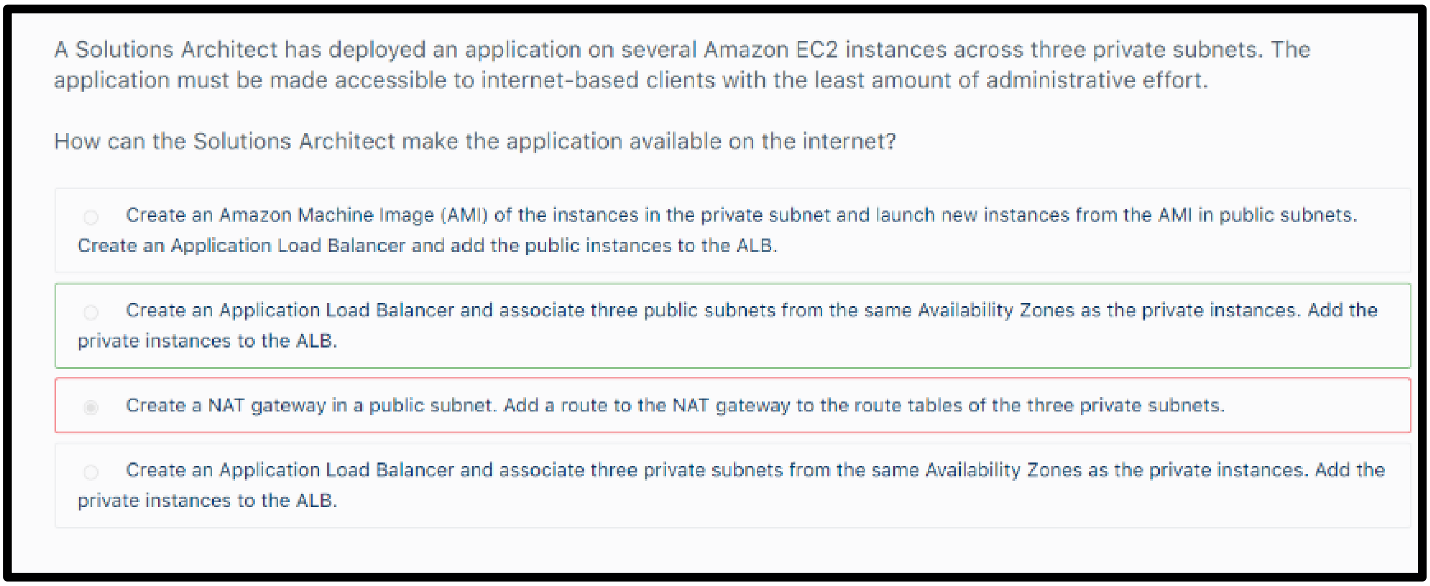


The Separate Subnet Principle

They MUST reside in separate subnets.

The differing targets in the default routes.

The NAT
Gateway



A Solutions Architect has deployed an application on several Amazon EC2 instances across three private subnets. The application must be made accessible to internet-based clients with the least amount of administrative effort.

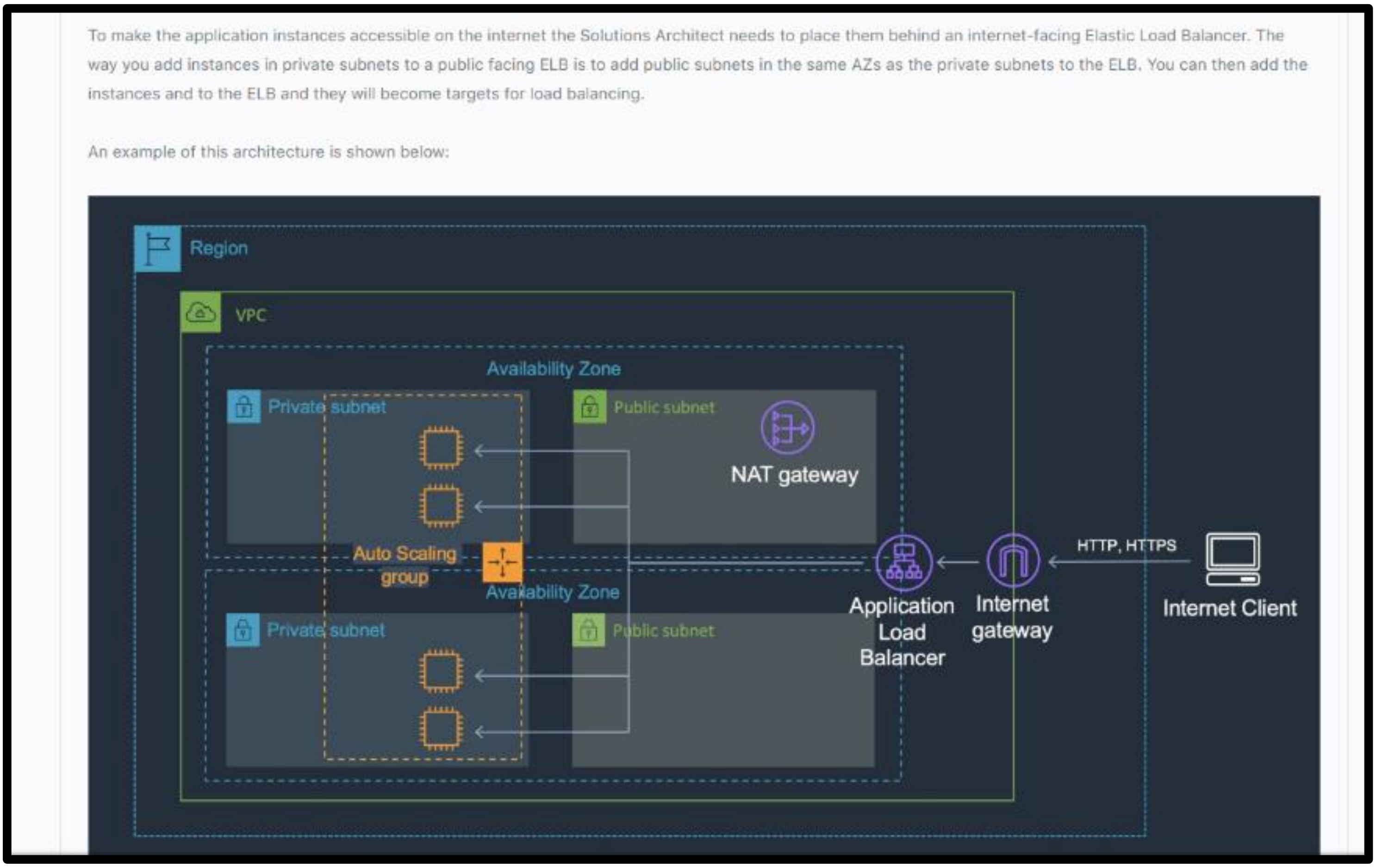
How can the Solutions Architect make the application available on the internet?

○ Create an Amazon Machine Image (AMI) of the instances in the private subnet and launch new instances from the AMI in public subnets. Create an Application Load Balancer and add the public instances to the ALB.

○ Create an Application Load Balancer and associate three public subnets from the same Availability Zones as the private instances. Add the private instances to the ALB.

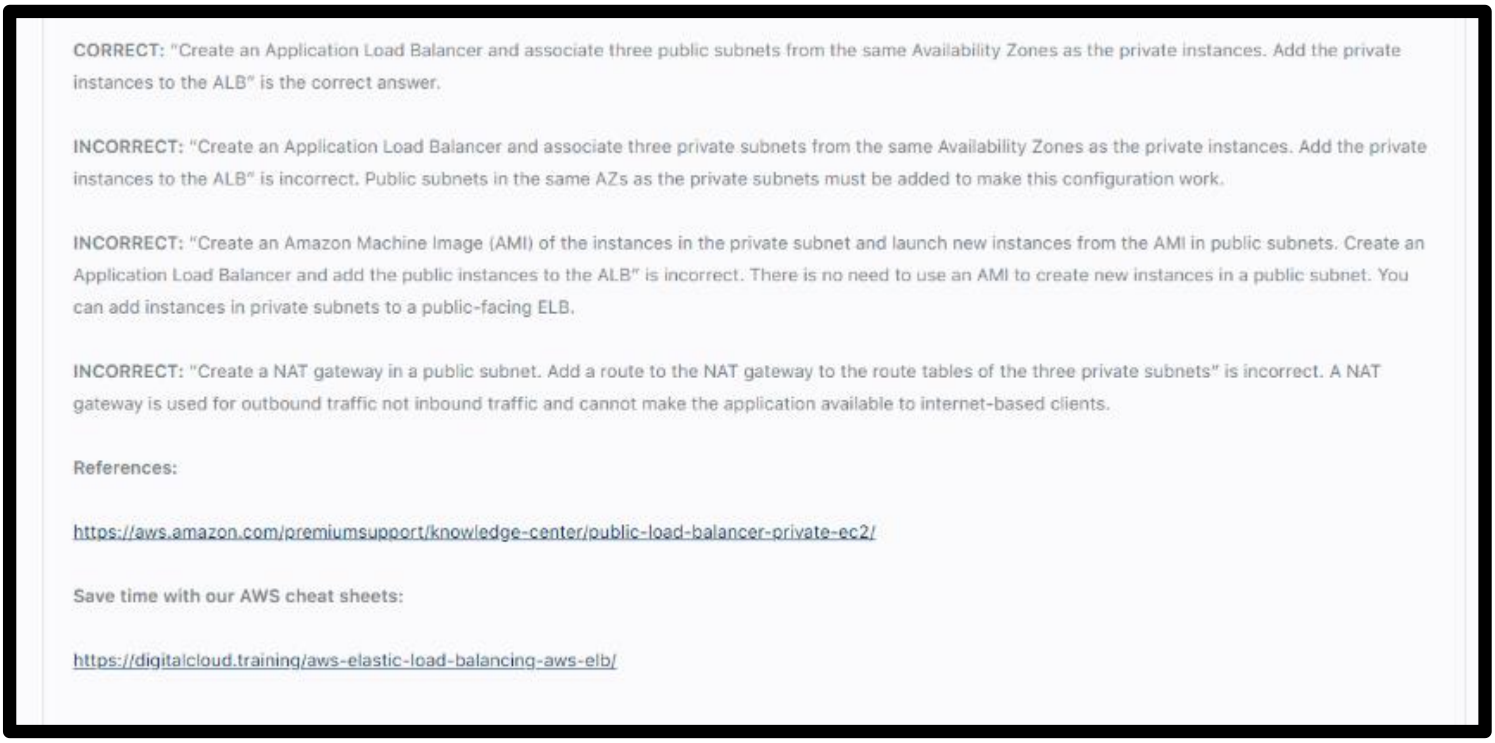
● Create a NAT gateway in a public subnet. Add a route to the NAT gateway to the route tables of the three private subnets.

○ Create an Application Load Balancer and associate three private subnets from the same Availability Zones as the private instances. Add the private instances to the ALB.



To make the application instances accessible on the internet the Solutions Architect needs to place them behind an internet-facing Elastic Load Balancer. The way you add instances in private subnets to a public facing ELB is to add public subnets in the same AZs as the private subnets to the ELB. You can then add the instances and to the ELB and they will become targets for load balancing.

An example of this architecture is shown below:



CORRECT: "Create an Application Load Balancer and associate three public subnets from the same Availability Zones as the private instances. Add the private instances to the ALB" is the correct answer.

INCORRECT: "Create an Application Load Balancer and associate three private subnets from the same Availability Zones as the private instances. Add the private instances to the ALB" is incorrect. Public subnets in the same AZs as the private subnets must be added to make this configuration work.

INCORRECT: "Create an Amazon Machine Image (AMI) of the instances in the private subnet and launch new instances from the AMI in public subnets. Create an Application Load Balancer and add the public instances to the ALB" is incorrect. There is no need to use an AMI to create new instances in a public subnet. You can add instances in private subnets to a public-facing ELB.

INCORRECT: "Create a NAT gateway in a public subnet. Add a route to the NAT gateway to the route tables of the three private subnets" is incorrect. A NAT gateway is used for outbound traffic not inbound traffic and cannot make the application available to internet-based clients.

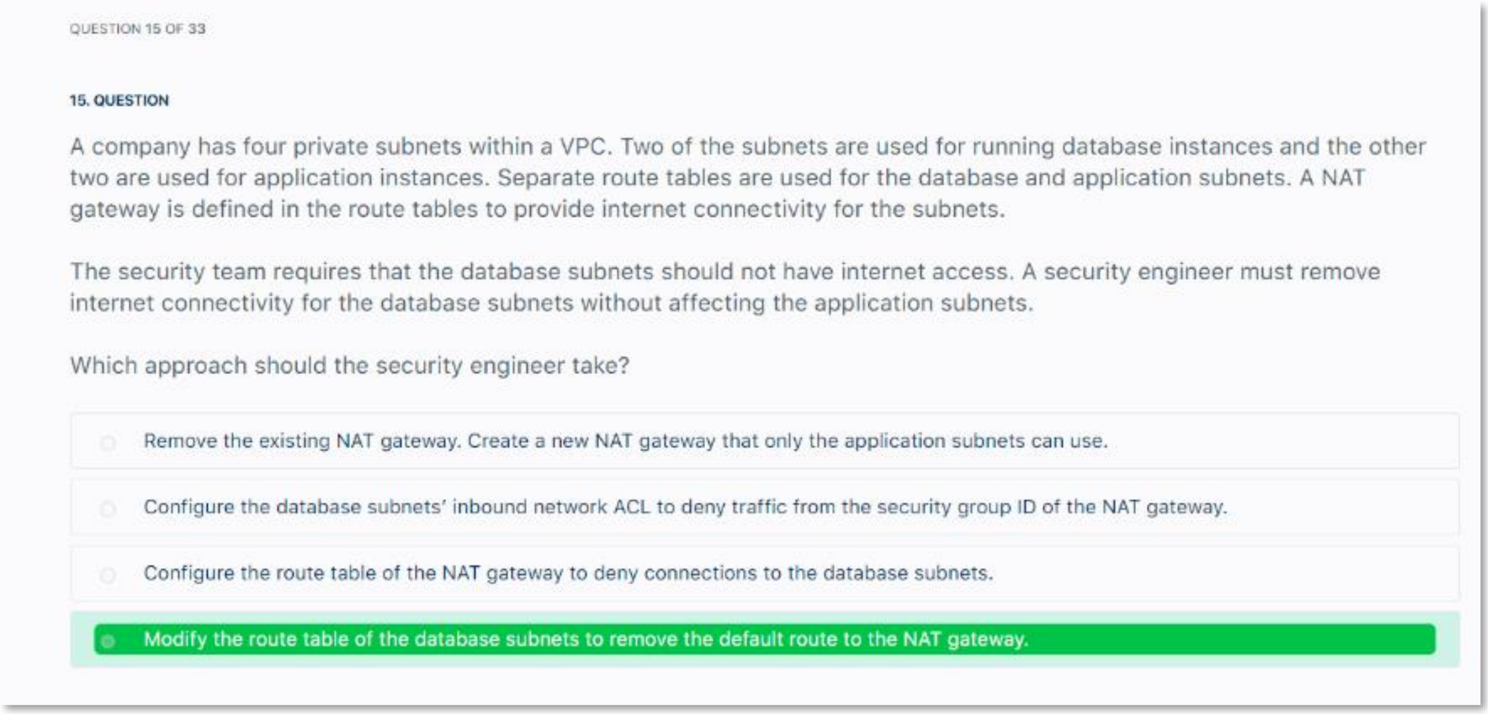
References:

https://aws.amazon.com/premiumsupport/knowledge-center/public-load-balancer-private-ec2/

Save time with our AWS cheat sheets:

https://digitalcloud.training/aws-elastic-load-balancing-aws-elb/

"A NAT Gateway is used for ***OUTBOUND*** traffic"



QUESTION 15 OF 33

15. QUESTION

A company has four private subnets within a VPC. Two of the subnets are used for running database instances and the other two are used for application instances. Separate route tables are used for the database and application subnets. A NAT gateway is defined in the route tables to provide internet connectivity for the subnets.

The security team requires that the database subnets should not have internet access. A security engineer must remove internet connectivity for the database subnets without affecting the application subnets.

Which approach should the security engineer take?

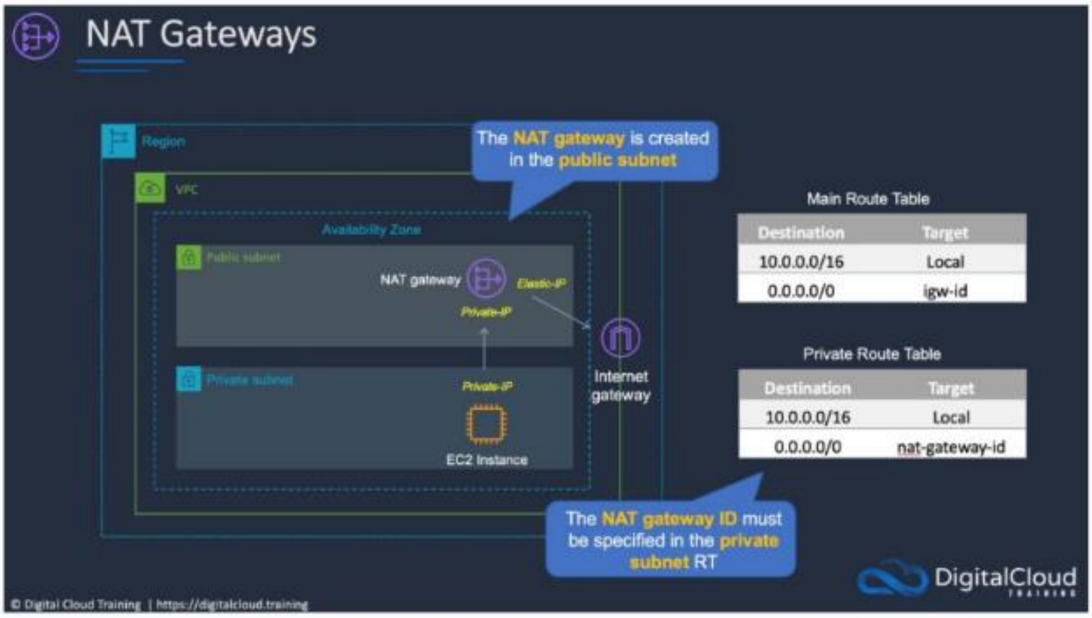
- ○ Remove the existing NAT gateway. Create a new NAT gateway that only the application subnets can use.
- ○ Configure the database subnets' inbound network ACL to deny traffic from the security group ID of the NAT gateway.
- ○ Configure the route table of the NAT gateway to deny connections to the database subnets.
- ● Modify the route table of the database subnets to remove the default route to the NAT gateway.

Correct

Explanation:

As seen in the diagram below, the NAT gateway is deployed in a public subnet and the route tables of private subnets are updated with a route pointing to the NAT gateway for all traffic for which another more specific route is not defined.

Therefore, the only change that needs to be made is to remove the route table entry for the NAT gateway from the route table of the private subnets in which the database instances are running.



The NAT gateway is created in the public subnet

Main Route Table

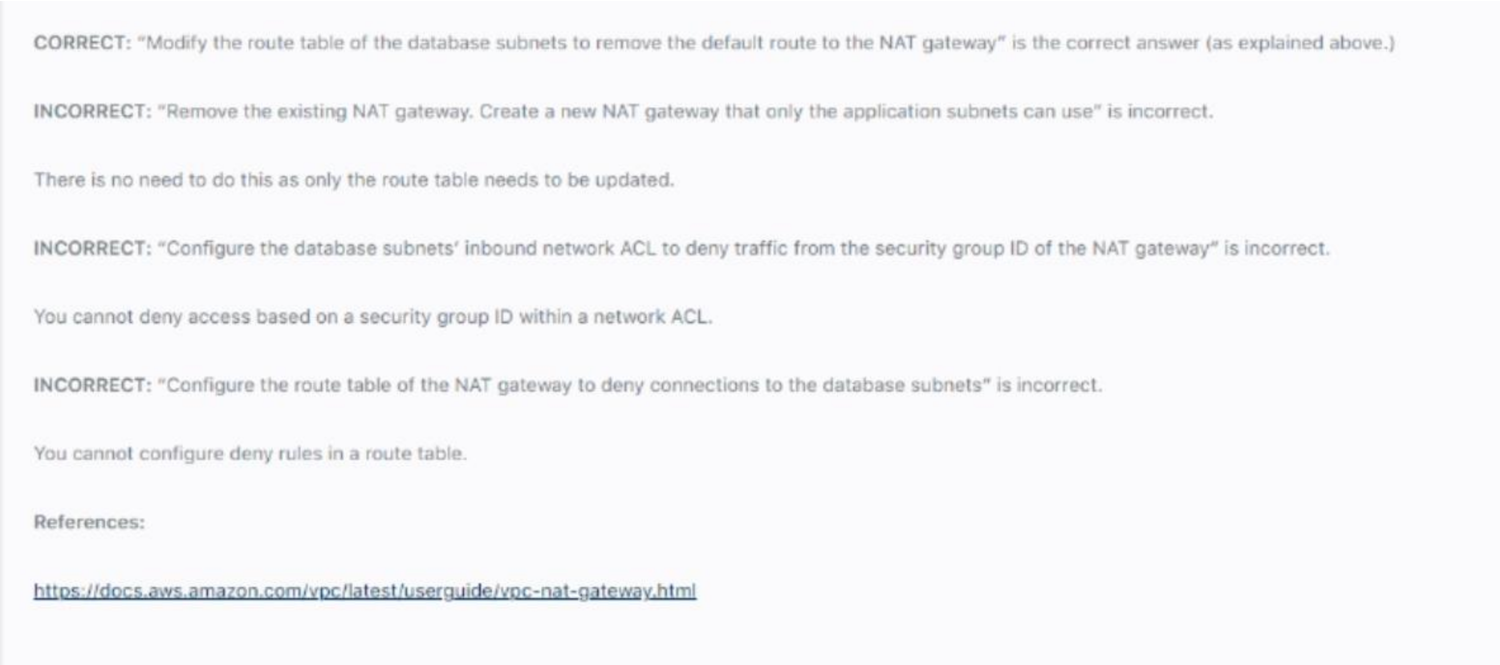
Destination	Target
10.0.0.0/16	Local
0.0.0.0/0	igw-id

Private Route Table

Destination	Target
10.0.0.0/16	Local
0.0.0.0/0	nat-gateway-id

The NAT gateway ID must be specified in the private subnet RT

© Digital Cloud Training | https://digitalcloud.training



CORRECT: "Modify the route table of the database subnets to remove the default route to the NAT gateway" is the correct answer (as explained above.)

INCORRECT: "Remove the existing NAT gateway. Create a new NAT gateway that only the application subnets can use" is incorrect.

There is no need to do this as only the route table needs to be updated.

INCORRECT: "Configure the database subnets' inbound network ACL to deny traffic from the security group ID of the NAT gateway" is incorrect.

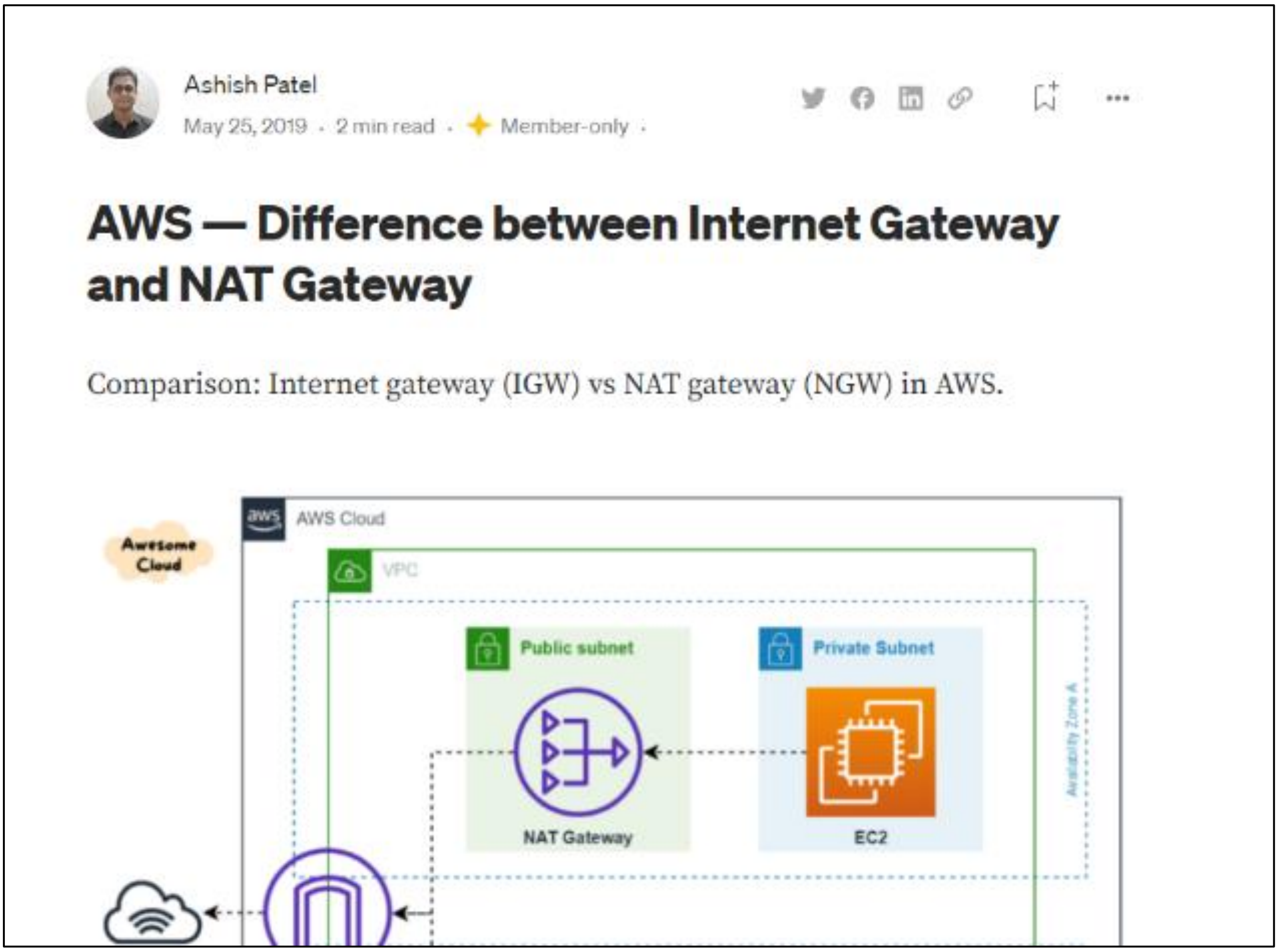
You cannot deny access based on a security group ID within a network ACL.

INCORRECT: "Configure the route table of the NAT gateway to deny connections to the database subnets" is incorrect.

You cannot configure deny rules in a route table.

References:

https://docs.aws.amazon.com/vpc/latest/userguide/vpc-nat-gateway.html

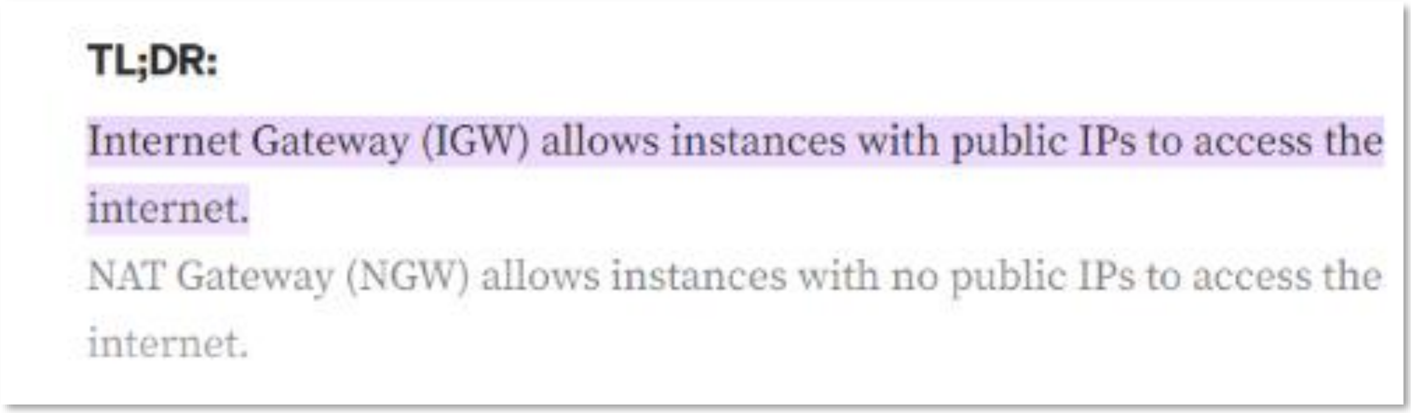


Ashish Patel

May 25, 2019 · 2 min read · ✦ Member-only ·

AWS — Difference between Internet Gateway and NAT Gateway

Comparison: Internet gateway (IGW) vs NAT gateway (NGW) in AWS.



TL;DR:

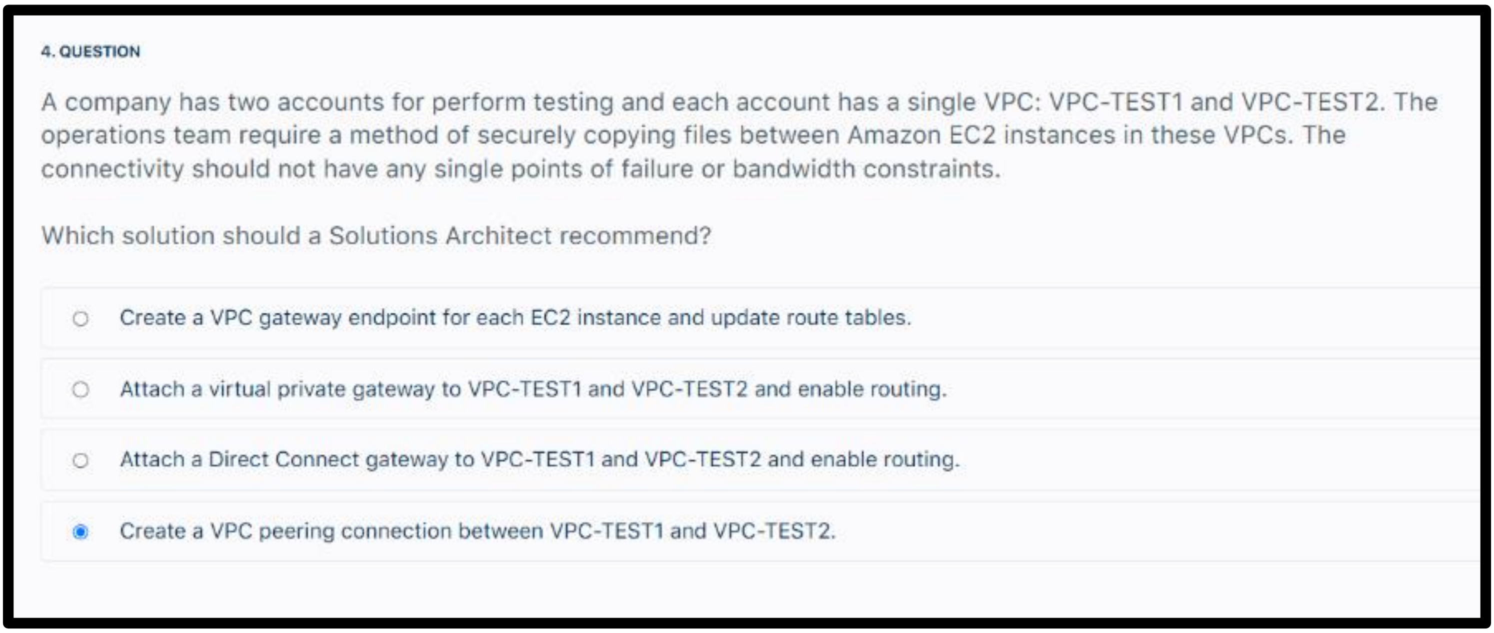
Internet Gateway (IGW) allows instances with public IPs to access the internet.

NAT Gateway (NGW) allows instances with no public IPs to access the internet.

Allocate-address

Associate-address

Describe-network-interfaces



4. QUESTION

A company has two accounts for perform testing and each account has a single VPC: VPC-TEST1 and VPC-TEST2. The operations team require a method of securely copying files between Amazon EC2 instances in these VPCs. The connectivity should not have any single points of failure or bandwidth constraints.

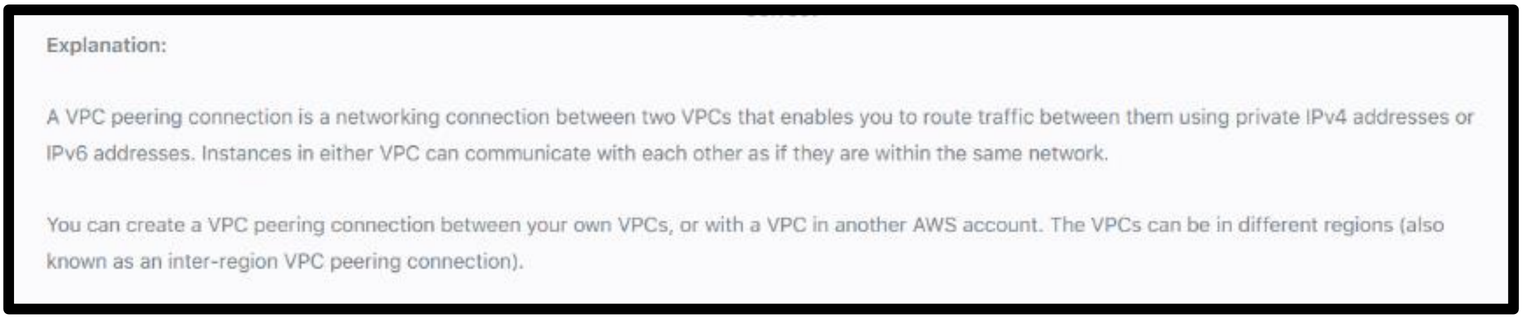
Which solution should a Solutions Architect recommend?

○ Create a VPC gateway endpoint for each EC2 instance and update route tables.

○ Attach a virtual private gateway to VPC-TEST1 and VPC-TEST2 and enable routing.

○ Attach a Direct Connect gateway to VPC-TEST1 and VPC-TEST2 and enable routing.

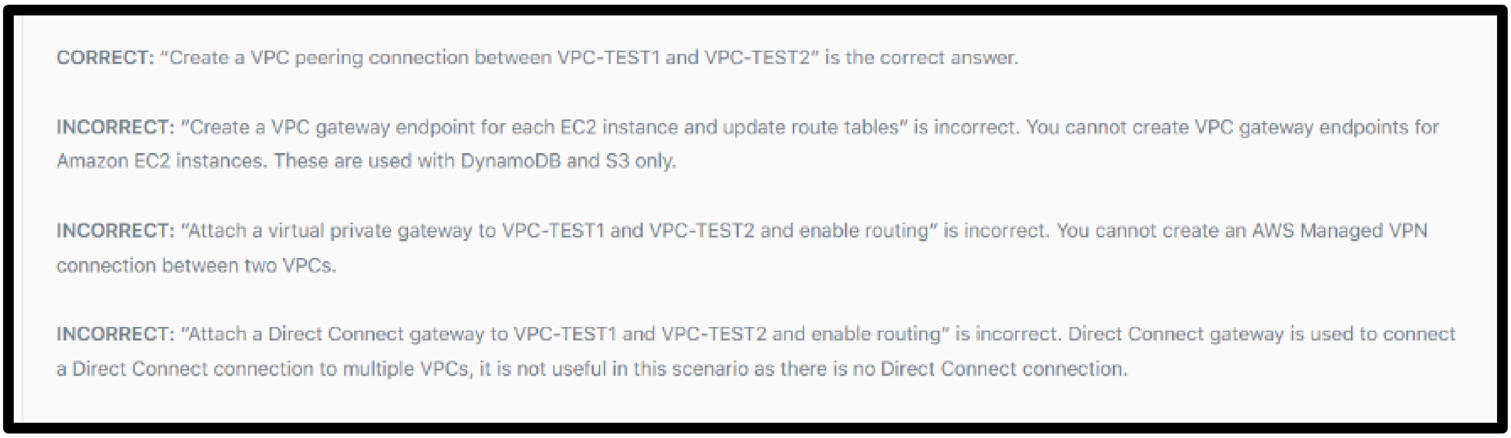
● Create a VPC peering connection between VPC-TEST1 and VPC-TEST2.



Explanation:

A VPC peering connection is a networking connection between two VPCs that enables you to route traffic between them using private IPv4 addresses or IPv6 addresses. Instances in either VPC can communicate with each other as if they are within the same network.

You can create a VPC peering connection between your own VPCs, or with a VPC in another AWS account. The VPCs can be in different regions (also known as an inter-region VPC peering connection).

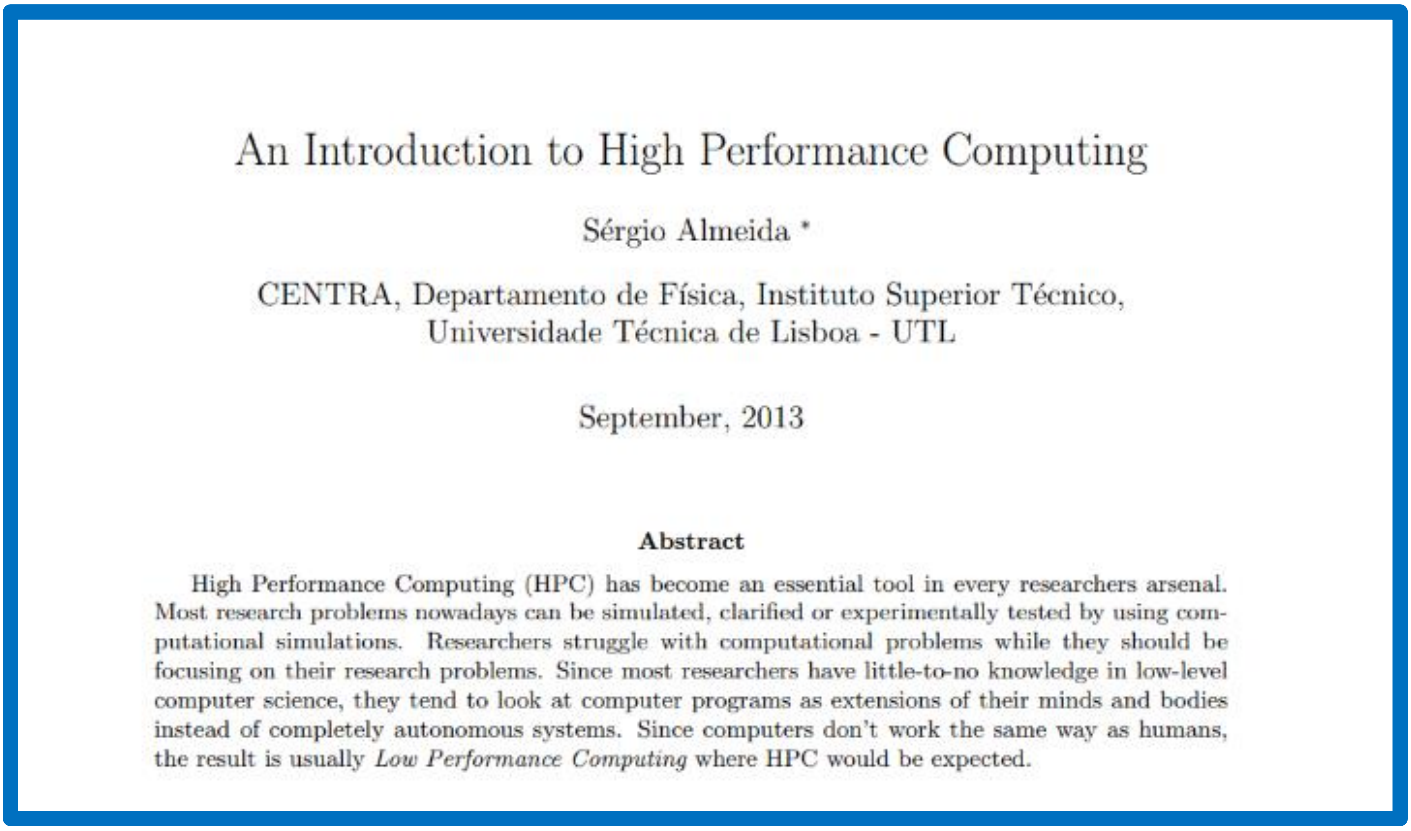


CORRECT: "Create a VPC peering connection between VPC-TEST1 and VPC-TEST2" is the correct answer.

INCORRECT: "Create a VPC gateway endpoint for each EC2 instance and update route tables" is incorrect. You cannot create VPC gateway endpoints for Amazon EC2 instances. These are used with DynamoDB and S3 only.

INCORRECT: "Attach a virtual private gateway to VPC-TEST1 and VPC-TEST2 and enable routing" is incorrect. You cannot create an AWS Managed VPN connection between two VPCs.

INCORRECT: "Attach a Direct Connect gateway to VPC-TEST1 and VPC-TEST2 and enable routing" is incorrect. Direct Connect gateway is used to connect a Direct Connect connection to multiple VPCs, it is not useful in this scenario as there is no Direct Connect connection.



An Introduction to High Performance Computing

Sérgio Almeida *

CENTRA, Departamento de Física, Instituto Superior Técnico,
Universidade Técnica de Lisboa - UTL

September, 2013

Abstract

High Performance Computing (HPC) has become an essential tool in every researchers arsenal. Most research problems nowadays can be simulated, clarified or experimentally tested by using computational simulations. Researchers struggle with computational problems while they should be focusing on their research problems. Since most researchers have little-to-no knowledge in low-level computer science, they tend to look at computer programs as extensions of their minds and bodies instead of completely autonomous systems. Since computers don't work the same way as humans, the result is usually *Low Performance Computing* where HPC would be expected.

TPN

81. ***Phenomenon1*** – the tendency of X to Y.
82. ***Phen2*** – the tendency of X to Y.
83. ***Phen3*** – the tendency of X to Y.
84. ***Phen4*** – the tendency of X to Y.
85. ***Phen5*** – the tendency of X to Y.
86. ***Phen6*** – the tendency of X to Y.
87. ***Phen7*** – the tendency of X to Y.
88. ***Phen8*** – the tendency of X to Y.
89. ***Phen9*** – the tendency of X to Y.
90. ***Phen10*** – the tendency of X to Y.

Glossary

Term1

Description of what term means here.

Term2

Description of what term means here.

Term3

Description of what term means here.

Bibliography

I. Official

https://medium.com/awesome-cloud/aws-vpc-difference-between-internet-gateway-and-nat-gateway-c9177e710af6#:~:text=Internet%20Gateway%20(IGW)%20allows%20instances,IPs%20to%20access%20the%20internet.

[Surname1]
Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]
Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

II. Unofficial

[Surname1]
Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

IV. General

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

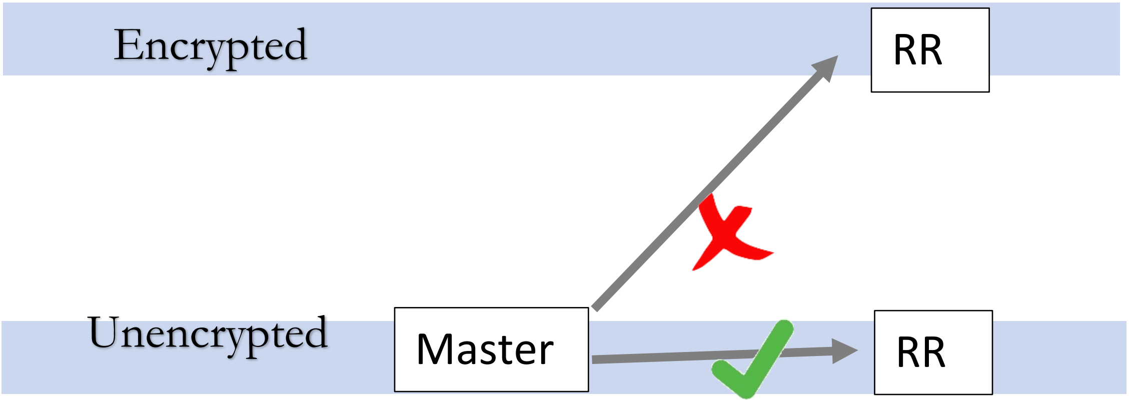
Relational Database Service

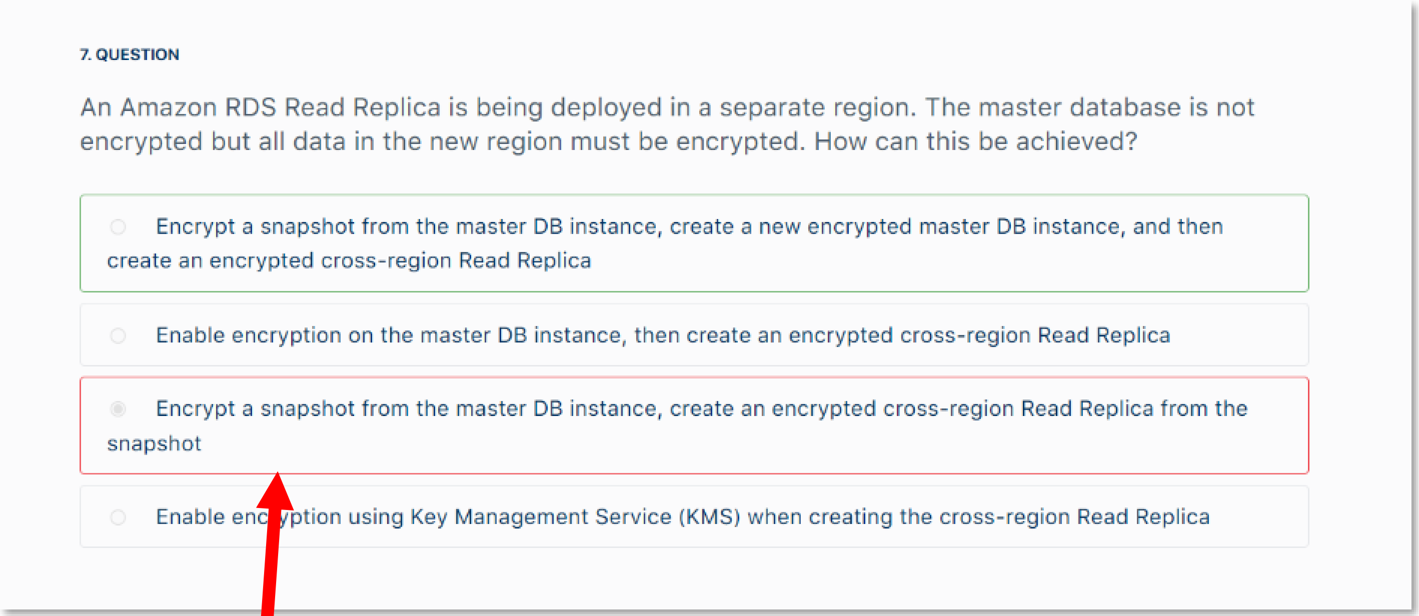
Entreaty against Elevation

(to *Encrypted*)

You **CANNOT** create an ***encrypted*** Read Replica (RR) from

an ***unencrypted*** master DB instance.





7. QUESTION

An Amazon RDS Read Replica is being deployed in a separate region. The master database is not encrypted but all data in the new region must be encrypted. How can this be achieved?

○ Encrypt a snapshot from the master DB instance, create a new encrypted master DB instance, and then create an encrypted cross-region Read Replica

○ Enable encryption on the master DB instance, then create an encrypted cross-region Read Replica

● Encrypt a snapshot from the master DB instance, create an encrypted cross-region Read Replica from the snapshot

○ Enable encryption using Key Management Service (KMS) when creating the cross-region Read Replica

You cannot even proceed by encrypting a *snapshot*!

The third option from the top is marked as incorrect.

Taken from Neal Davis's Digital Cloud Training (accessed 2022).

Can you think why? The idea is that you MUST create a new master. It is *not* enough to encrypt the snapshot ("instead of the master", runs the thought). Think about it: the snapshot is just a representation (image, perhaps) of the original database—and that original master is not encrypted.

I'm looking for the option that involves the creation of a distinct ***master***!

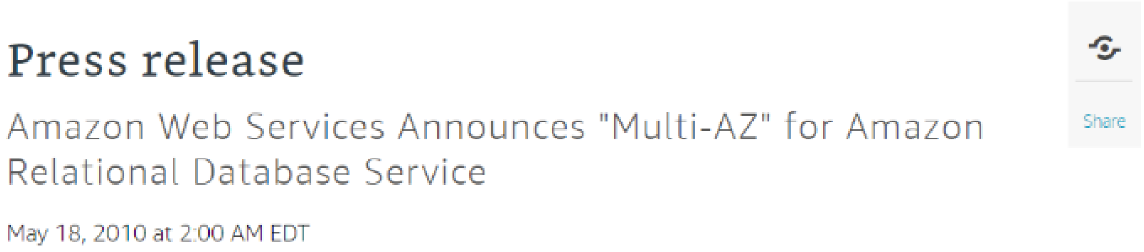
EMBRACE

Principle

If you want to EMBRACE ENCRYPTION, then you must do so from the beginning.

*You **cannot** enable encryption after launch time for the master DB instance.*

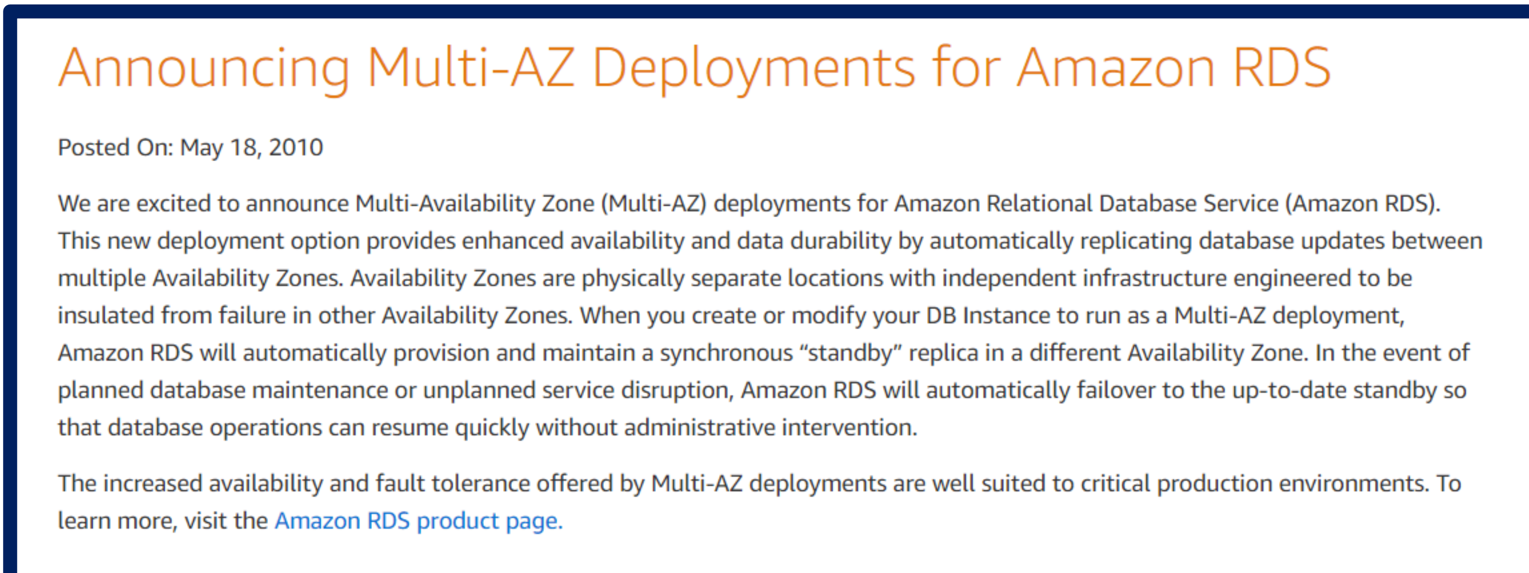
Pay attention to the distinction between:



Press release

Amazon Web Services Announces "Multi-AZ" for Amazon Relational Database Service

May 18, 2010 at 2:00 AM EDT

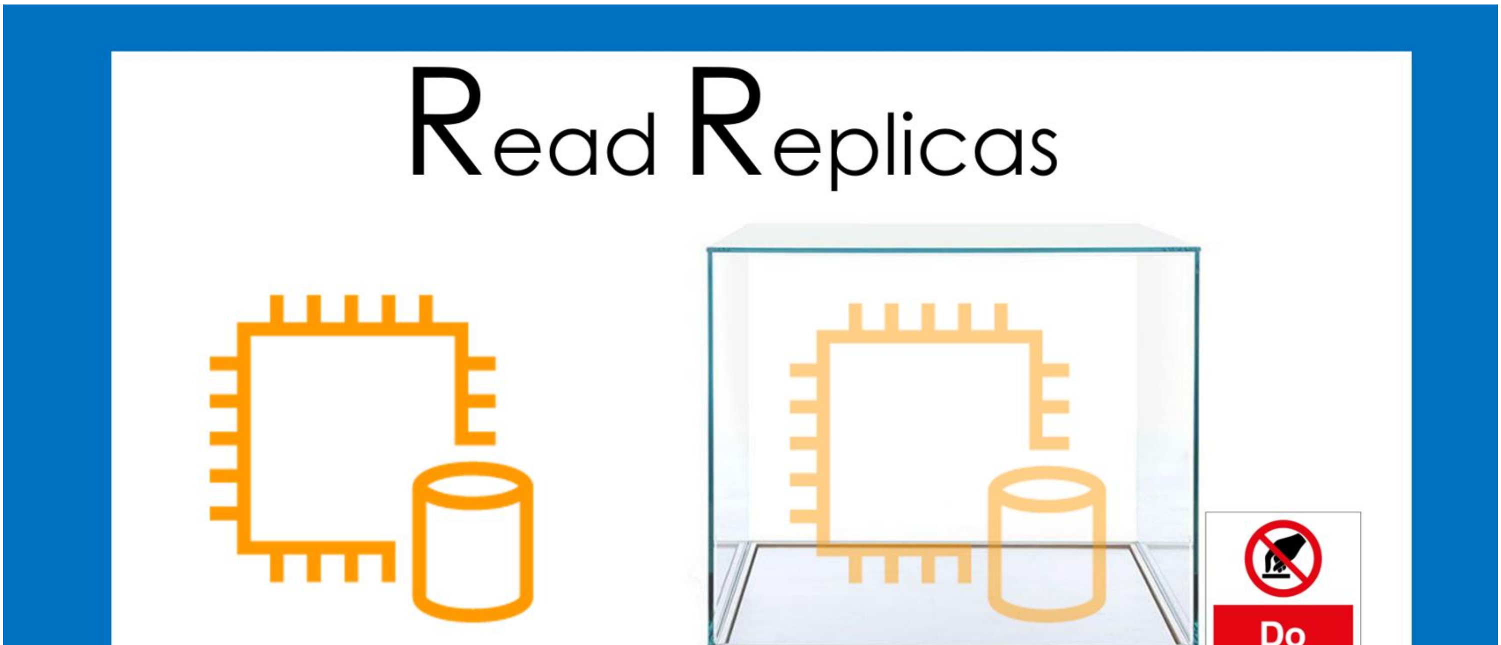


Announcing Multi-AZ Deployments for Amazon RDS

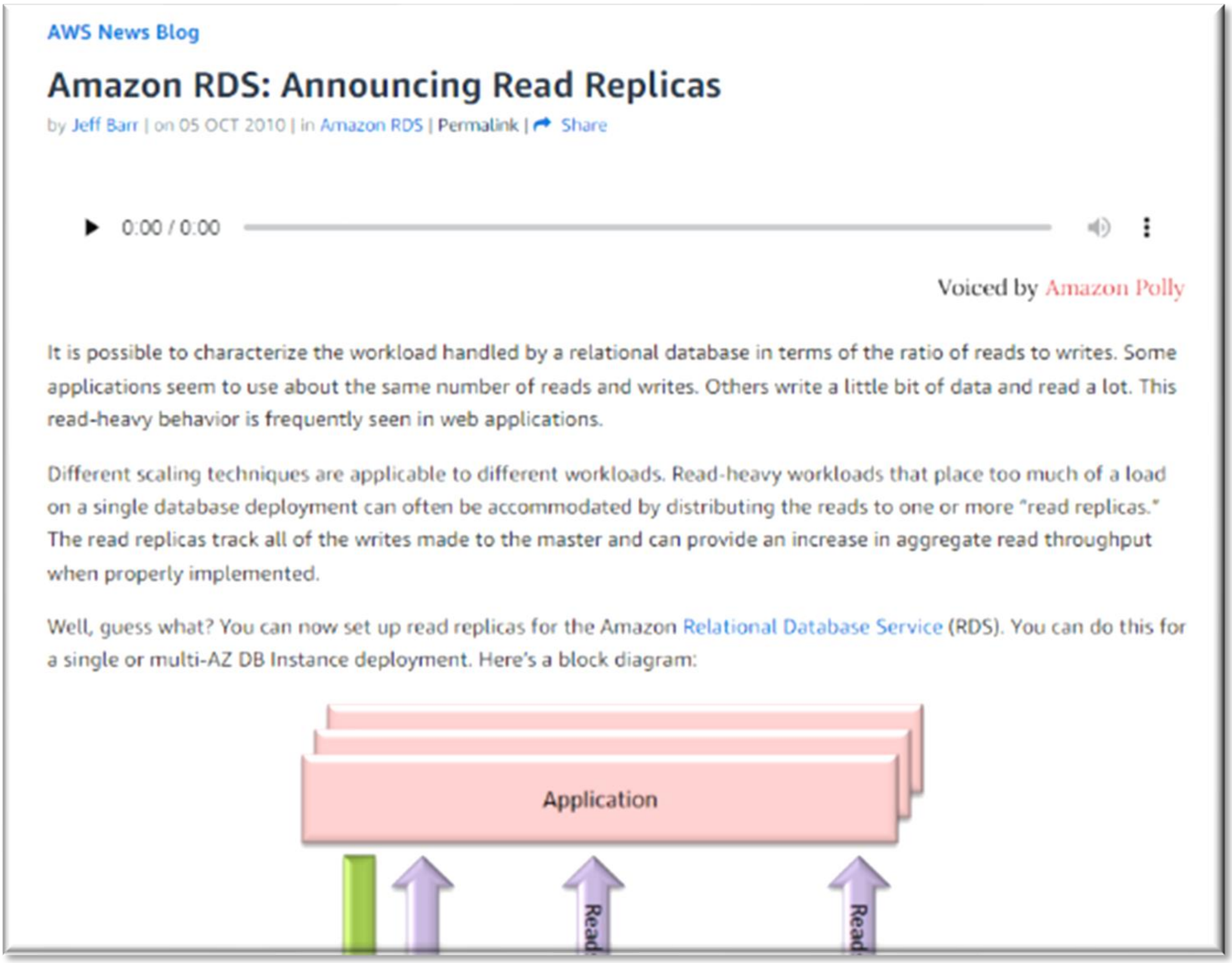
Posted On: May 18, 2010

We are excited to announce Multi-Availability Zone (Multi-AZ) deployments for Amazon Relational Database Service (Amazon RDS). This new deployment option provides enhanced availability and data durability by automatically replicating database updates between multiple Availability Zones. Availability Zones are physically separate locations with independent infrastructure engineered to be insulated from failure in other Availability Zones. When you create or modify your DB Instance to run as a Multi-AZ deployment, Amazon RDS will automatically provision and maintain a synchronous “standby” replica in a different Availability Zone. In the event of planned database maintenance or unplanned service disruption, Amazon RDS will automatically failover to the up-to-date standby so that database operations can resume quickly without administrative intervention.

The increased availability and fault tolerance offered by Multi-AZ deployments are well suited to critical production environments. To learn more, visit the Amazon RDS product page.



Read Replicas



AWS News Blog

Amazon RDS: Announcing Read Replicas

by Jeff Barr | on 05 OCT 2010 | in Amazon RDS | Permalink | Share

Voiced by Amazon Polly

It is possible to characterize the workload handled by a relational database in terms of the ratio of reads to writes. Some applications seem to use about the same number of reads and writes. Others write a little bit of data and read a lot. This read-heavy behavior is frequently seen in web applications.

Different scaling techniques are applicable to different workloads. Read-heavy workloads that place too much of a load on a single database deployment can often be accommodated by distributing the reads to one or more "read replicas." The read replicas track all of the writes made to the master and can provide an increase in aggregate read throughput when properly implemented.

Well, guess what? You can now set up read replicas for the Amazon Relational Database Service (RDS). You can do this for a single or multi-AZ DB Instance deployment. Here's a block diagram:

We've now looked at two principles (or constraints) relating to RDS. The names I've given them are my own. These two *CONSTRAINTS* are:

1. EEE

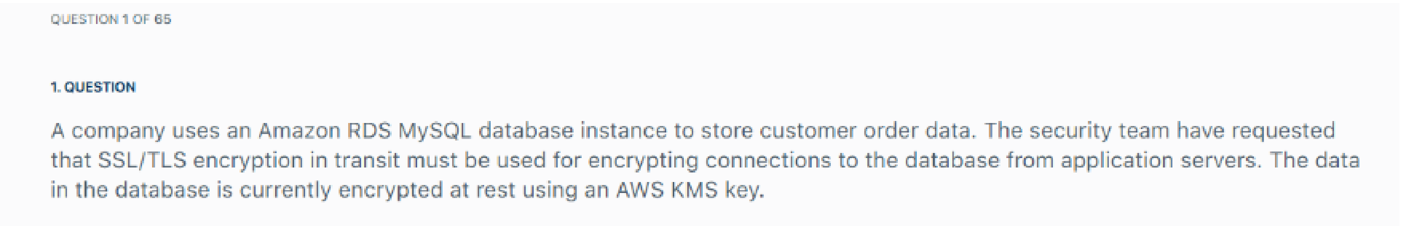
Entreaty against elevation, to encrypted status

This specifically concerns the creation of RRs from master instances

2. EE

The Embrace Encryption Principle

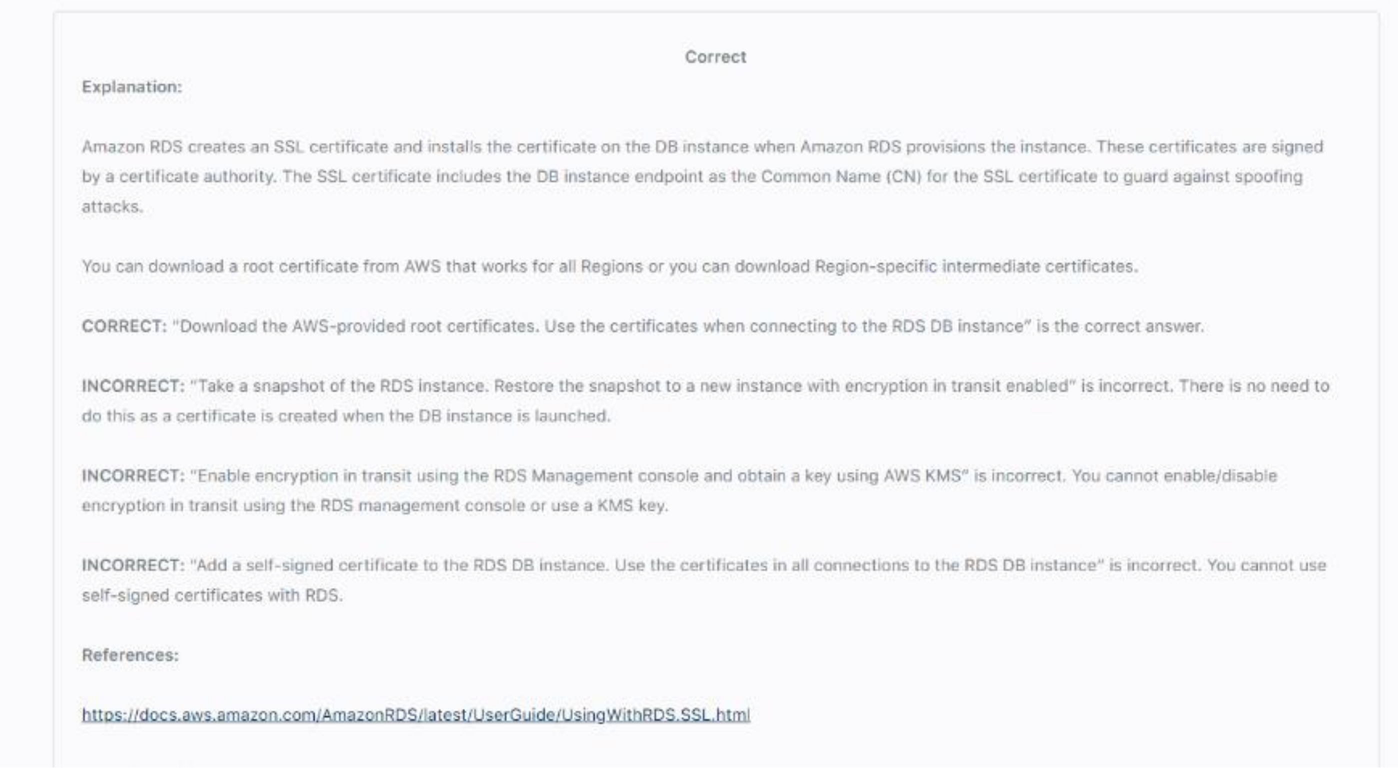
A master RDS instance cannot switch to being encrypted. It must be encrypted from the offset.



QUESTION 1 OF 65

1. QUESTION

A company uses an Amazon RDS MySQL database instance to store customer order data. The security team have requested that SSL/TLS encryption in transit must be used for encrypting connections to the database from application servers. The data in the database is currently encrypted at rest using an AWS KMS key.



Correct

Explanation:

Amazon RDS creates an SSL certificate and installs the certificate on the DB instance when Amazon RDS provisions the instance. These certificates are signed by a certificate authority. The SSL certificate includes the DB instance endpoint as the Common Name (CN) for the SSL certificate to guard against spoofing attacks.

You can download a root certificate from AWS that works for all Regions or you can download Region-specific intermediate certificates.

CORRECT: "Download the AWS-provided root certificates. Use the certificates when connecting to the RDS DB instance" is the correct answer.

INCORRECT: "Take a snapshot of the RDS instance. Restore the snapshot to a new instance with encryption in transit enabled" is incorrect. There is no need to do this as a certificate is created when the DB instance is launched.

INCORRECT: "Enable encryption in transit using the RDS Management console and obtain a key using AWS KMS" is incorrect. You cannot enable/disable encryption in transit using the RDS management console or use a KMS key.

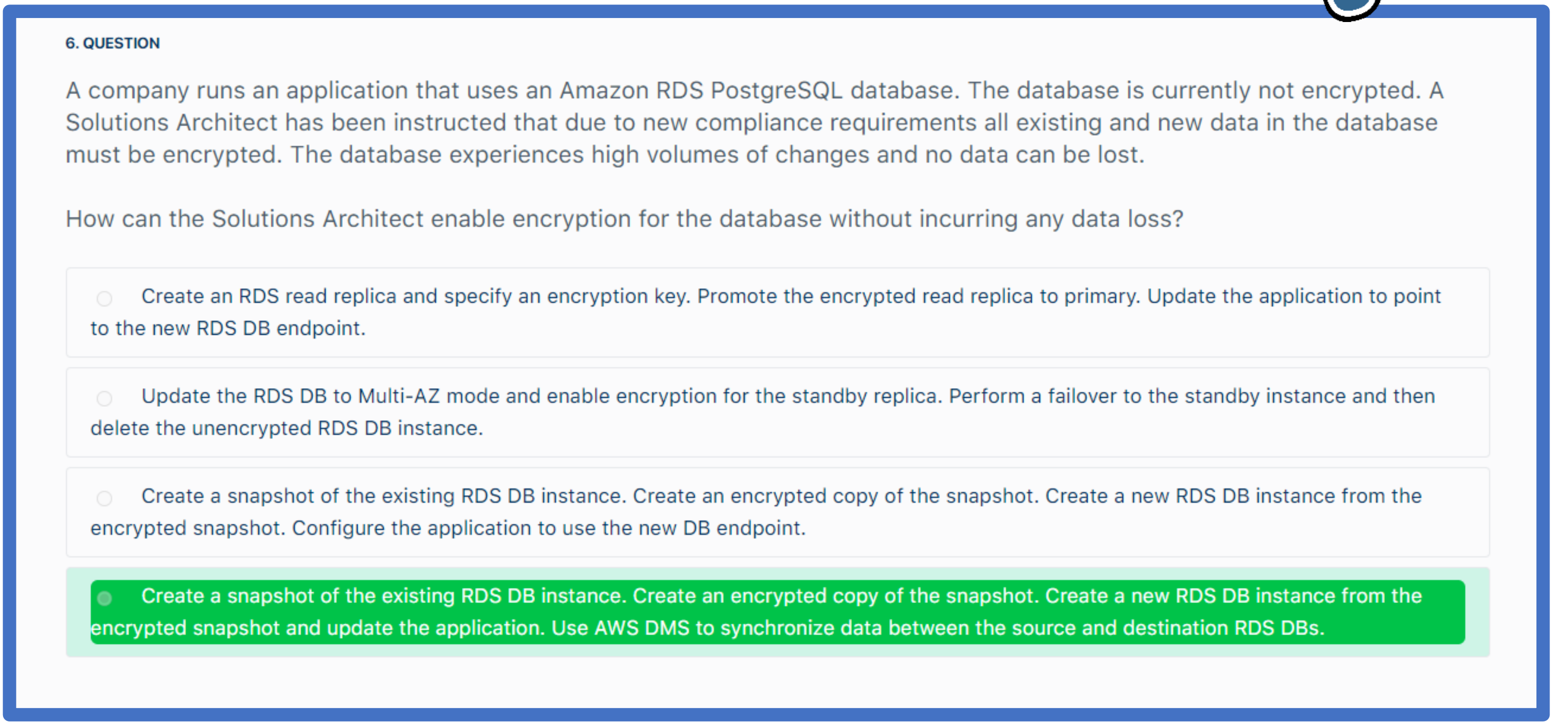
INCORRECT: "Add a self-signed certificate to the RDS DB instance. Use the certificates in all connections to the RDS DB instance" is incorrect. You cannot use self-signed certificates with RDS.

References:

https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/UsingWithRDS.SSL.html

375g ℮
MULTIGRAIN HOOPS
CROWNFIELD
MULTIGRAIN
HOOPS
AZ

Deployments



6. QUESTION

A company runs an application that uses an Amazon RDS PostgreSQL database. The database is currently not encrypted. A Solutions Architect has been instructed that due to new compliance requirements all existing and new data in the database must be encrypted. The database experiences high volumes of changes and no data can be lost.

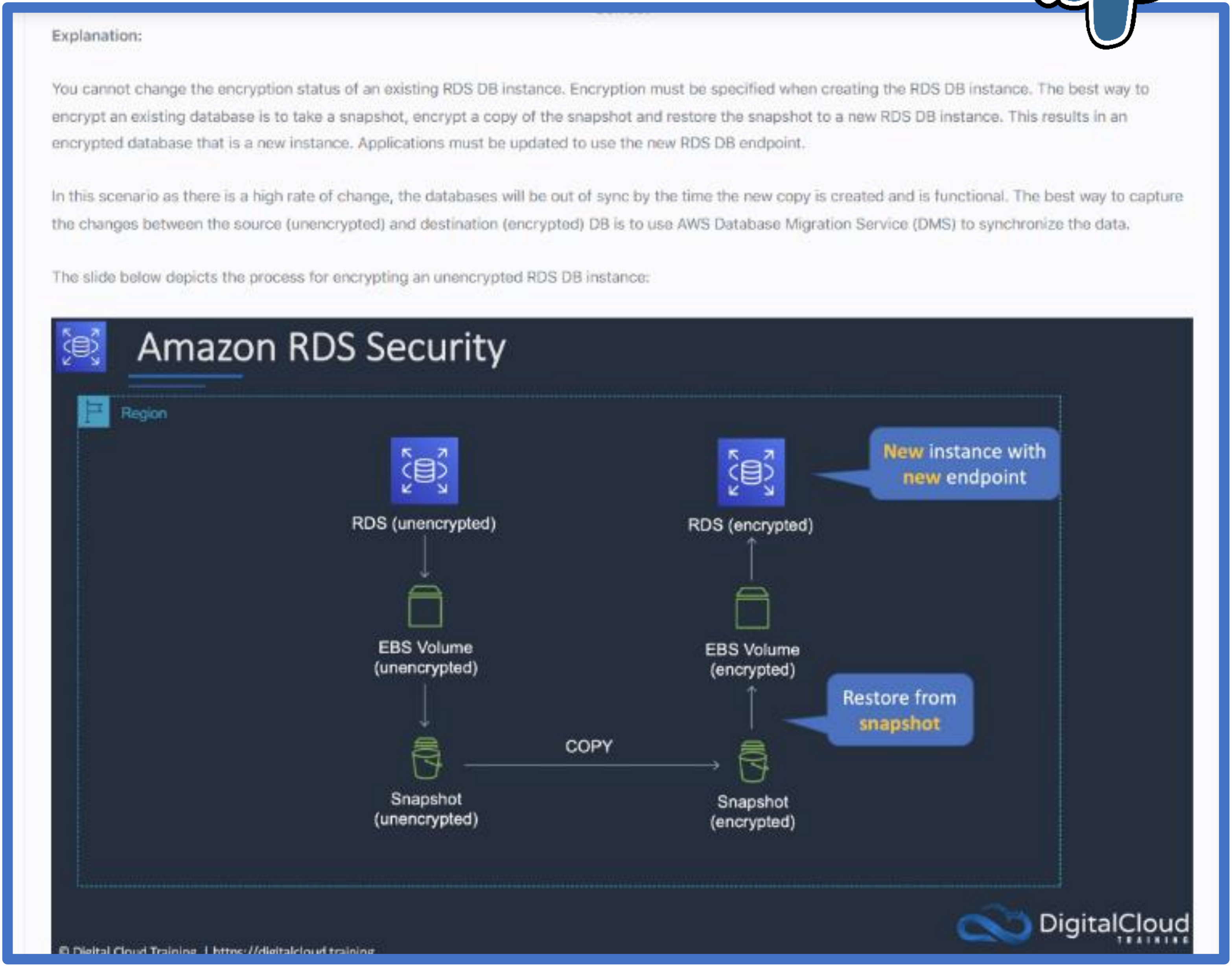
How can the Solutions Architect enable encryption for the database without incurring any data loss?

○ Create an RDS read replica and specify an encryption key. Promote the encrypted read replica to primary. Update the application to point to the new RDS DB endpoint.

○ Update the RDS DB to Multi-AZ mode and enable encryption for the standby replica. Perform a failover to the standby instance and then delete the unencrypted RDS DB instance.

○ Create a snapshot of the existing RDS DB instance. Create an encrypted copy of the snapshot. Create a new RDS DB instance from the encrypted snapshot. Configure the application to use the new DB endpoint.

● Create a snapshot of the existing RDS DB instance. Create an encrypted copy of the snapshot. Create a new RDS DB instance from the encrypted snapshot and update the application. Use AWS DMS to synchronize data between the source and destination RDS DBs.



Explanation:

You cannot change the encryption status of an existing RDS DB instance. Encryption must be specified when creating the RDS DB instance. The best way to encrypt an existing database is to take a snapshot, encrypt a copy of the snapshot and restore the snapshot to a new RDS DB instance. This results in an encrypted database that is a new instance. Applications must be updated to use the new RDS DB endpoint.

In this scenario as there is a high rate of change, the databases will be out of sync by the time the new copy is created and is functional. The best way to capture the changes between the source (unencrypted) and destination (encrypted) DB is to use AWS Database Migration Service (DMS) to synchronize the data.

The slide below depicts the process for encrypting an unencrypted RDS DB instance:

Amazon RDS Security

Region

RDS (unencrypted)

EBS Volume (unencrypted)

Snapshot (unencrypted)

COPY

Snapshot (encrypted)

EBS Volume (encrypted)

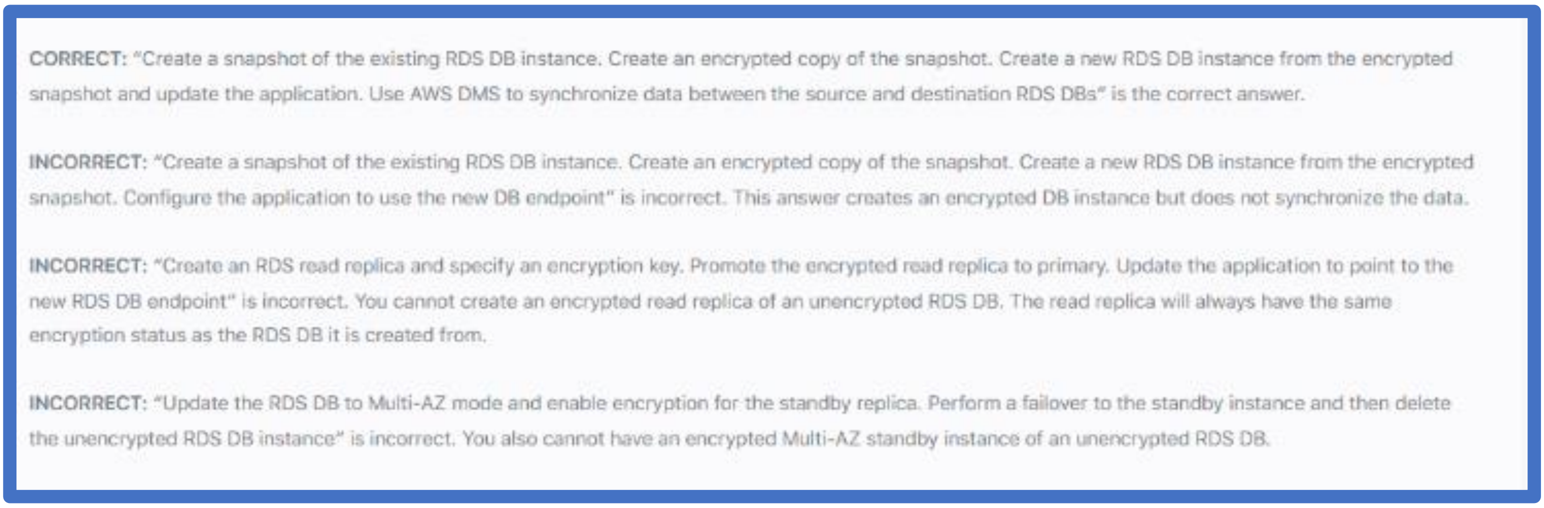
RDS (encrypted)

New instance with new endpoint

Restore from snapshot

DigitalCloud TRAINING

© Digital Cloud Training | https://digitalcloud.training



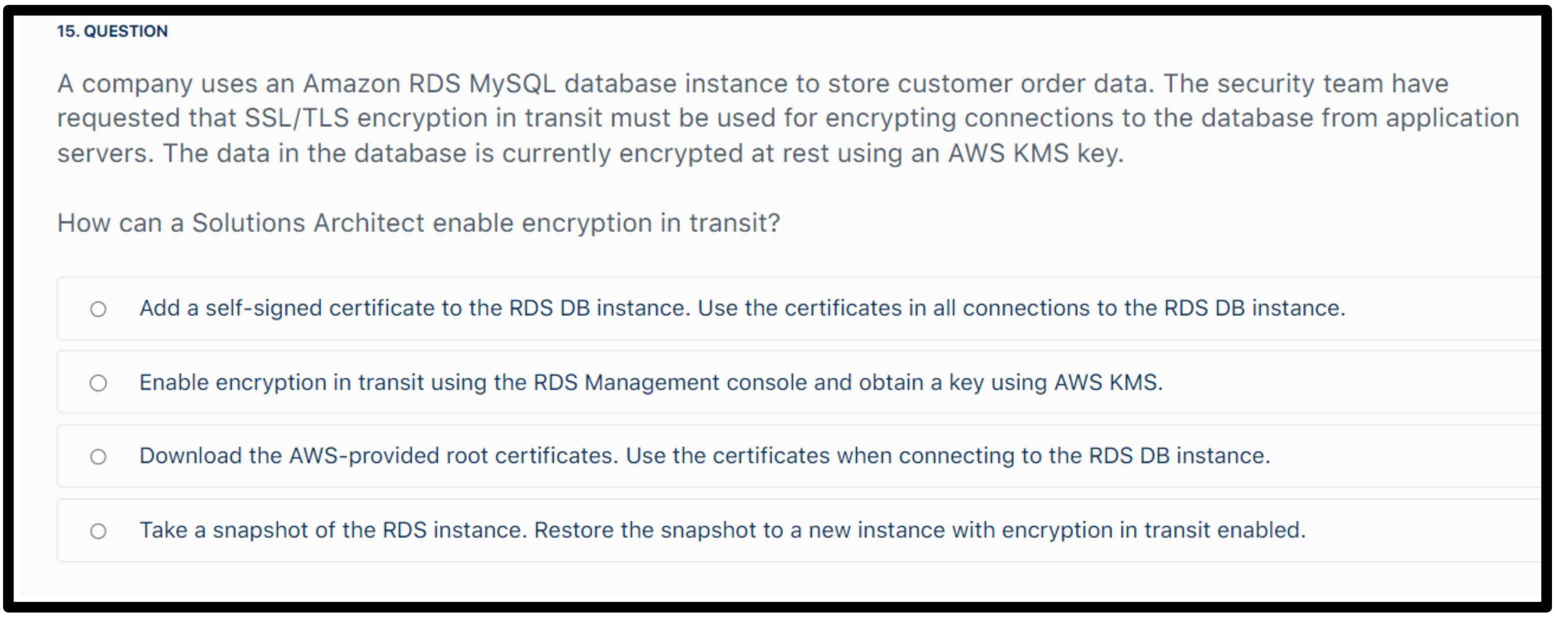
CORRECT: "Create a snapshot of the existing RDS DB instance. Create an encrypted copy of the snapshot. Create a new RDS DB instance from the encrypted snapshot and update the application. Use AWS DMS to synchronize data between the source and destination RDS DBs" is the correct answer.

INCORRECT: "Create a snapshot of the existing RDS DB instance. Create an encrypted copy of the snapshot. Create a new RDS DB instance from the encrypted snapshot. Configure the application to use the new DB endpoint" is incorrect. This answer creates an encrypted DB instance but does not synchronize the data.

INCORRECT: "Create an RDS read replica and specify an encryption key. Promote the encrypted read replica to primary. Update the application to point to the new RDS DB endpoint" is incorrect. You cannot create an encrypted read replica of an unencrypted RDS DB. The read replica will always have the same encryption status as the RDS DB it is created from.

INCORRECT: "Update the RDS DB to Multi-AZ mode and enable encryption for the standby replica. Perform a failover to the standby instance and then delete the unencrypted RDS DB instance" is incorrect. You also cannot have an encrypted Multi-AZ standby instance of an unencrypted RDS DB.

In the above question, the option I would probably go for (other than the correct one) would be the one that states: "create a new RDS DB instance from the encrypted snapshot". We are told, in the explanation, that "this answer creates an encrypted DB instance but does not synchronize the data". So, it really is essential that we have the synchronisation capability provided by AWS DMS.



15. QUESTION

A company uses an Amazon RDS MySQL database instance to store customer order data. The security team have requested that SSL/TLS encryption in transit must be used for encrypting connections to the database from application servers. The data in the database is currently encrypted at rest using an AWS KMS key.

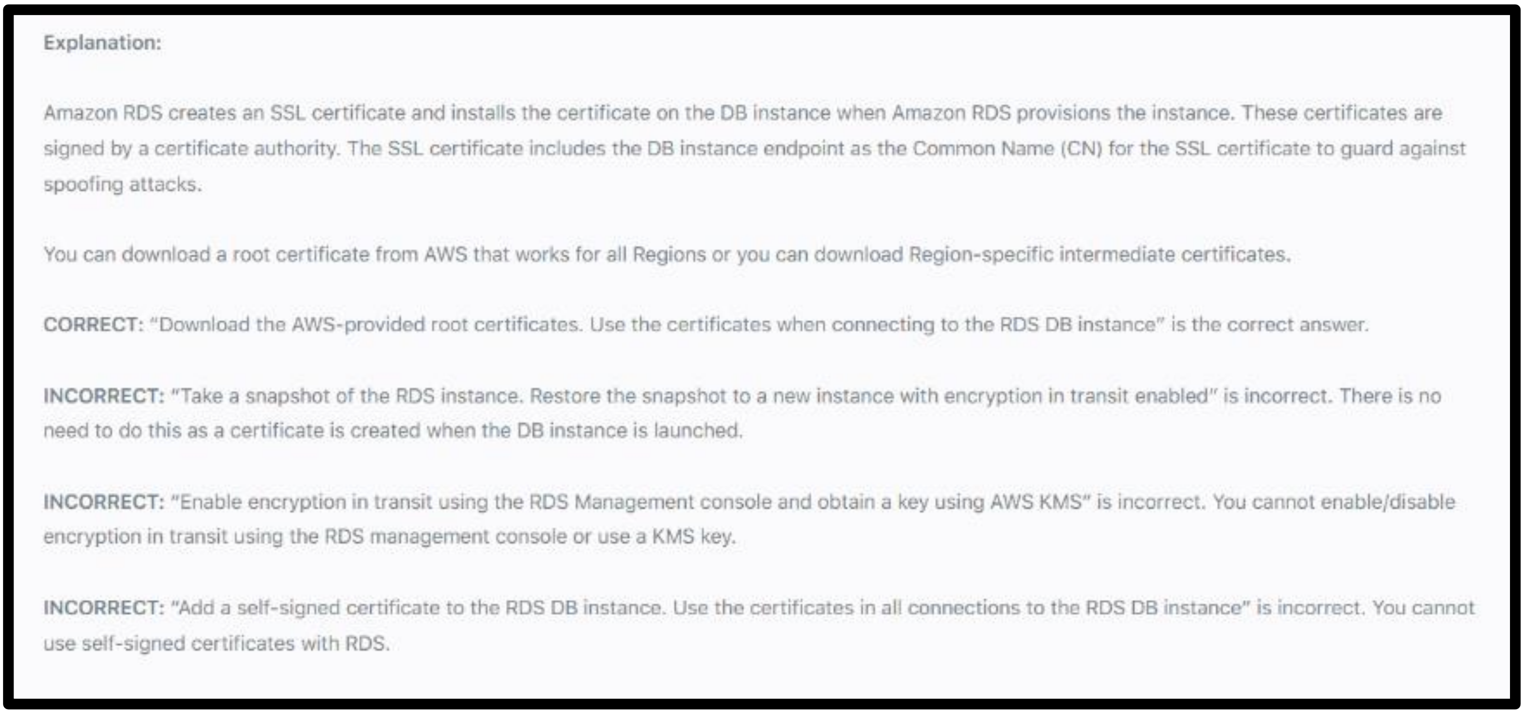
How can a Solutions Architect enable encryption in transit?

○ Add a self-signed certificate to the RDS DB instance. Use the certificates in all connections to the RDS DB instance.

○ Enable encryption in transit using the RDS Management console and obtain a key using AWS KMS.

○ Download the AWS-provided root certificates. Use the certificates when connecting to the RDS DB instance.

○ Take a snapshot of the RDS instance. Restore the snapshot to a new instance with encryption in transit enabled.



Explanation:

Amazon RDS creates an SSL certificate and installs the certificate on the DB instance when Amazon RDS provisions the instance. These certificates are signed by a certificate authority. The SSL certificate includes the DB instance endpoint as the Common Name (CN) for the SSL certificate to guard against spoofing attacks.

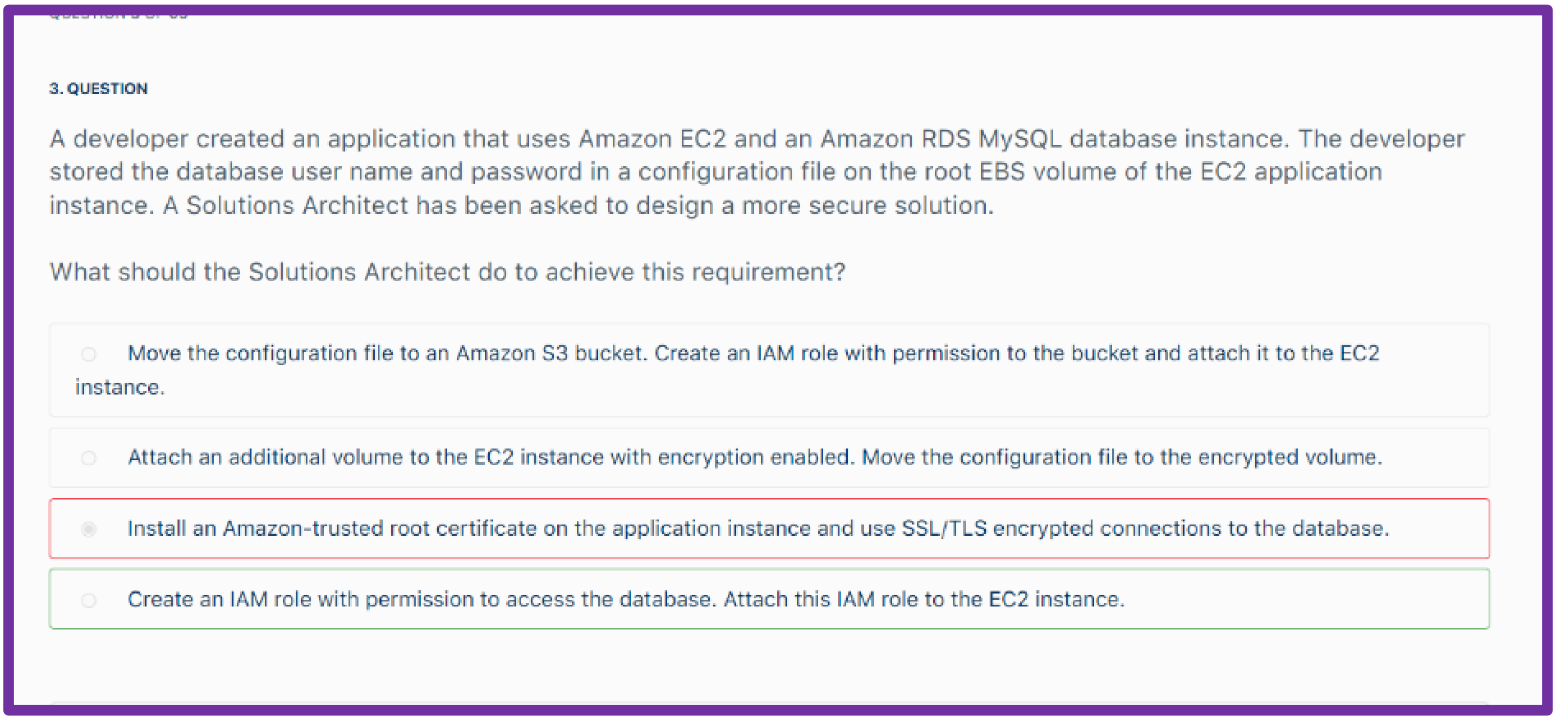
You can download a root certificate from AWS that works for all Regions or you can download Region-specific intermediate certificates.

CORRECT: "Download the AWS-provided root certificates. Use the certificates when connecting to the RDS DB instance" is the correct answer.

INCORRECT: "Take a snapshot of the RDS instance. Restore the snapshot to a new instance with encryption in transit enabled" is incorrect. There is no need to do this as a certificate is created when the DB instance is launched.

INCORRECT: "Enable encryption in transit using the RDS Management console and obtain a key using AWS KMS" is incorrect. You cannot enable/disable encryption in transit using the RDS management console or use a KMS key.

INCORRECT: "Add a self-signed certificate to the RDS DB instance. Use the certificates in all connections to the RDS DB instance" is incorrect. You cannot use self-signed certificates with RDS.



3. QUESTION

A developer created an application that uses Amazon EC2 and an Amazon RDS MySQL database instance. The developer stored the database user name and password in a configuration file on the root EBS volume of the EC2 application instance. A Solutions Architect has been asked to design a more secure solution.

What should the Solutions Architect do to achieve this requirement?

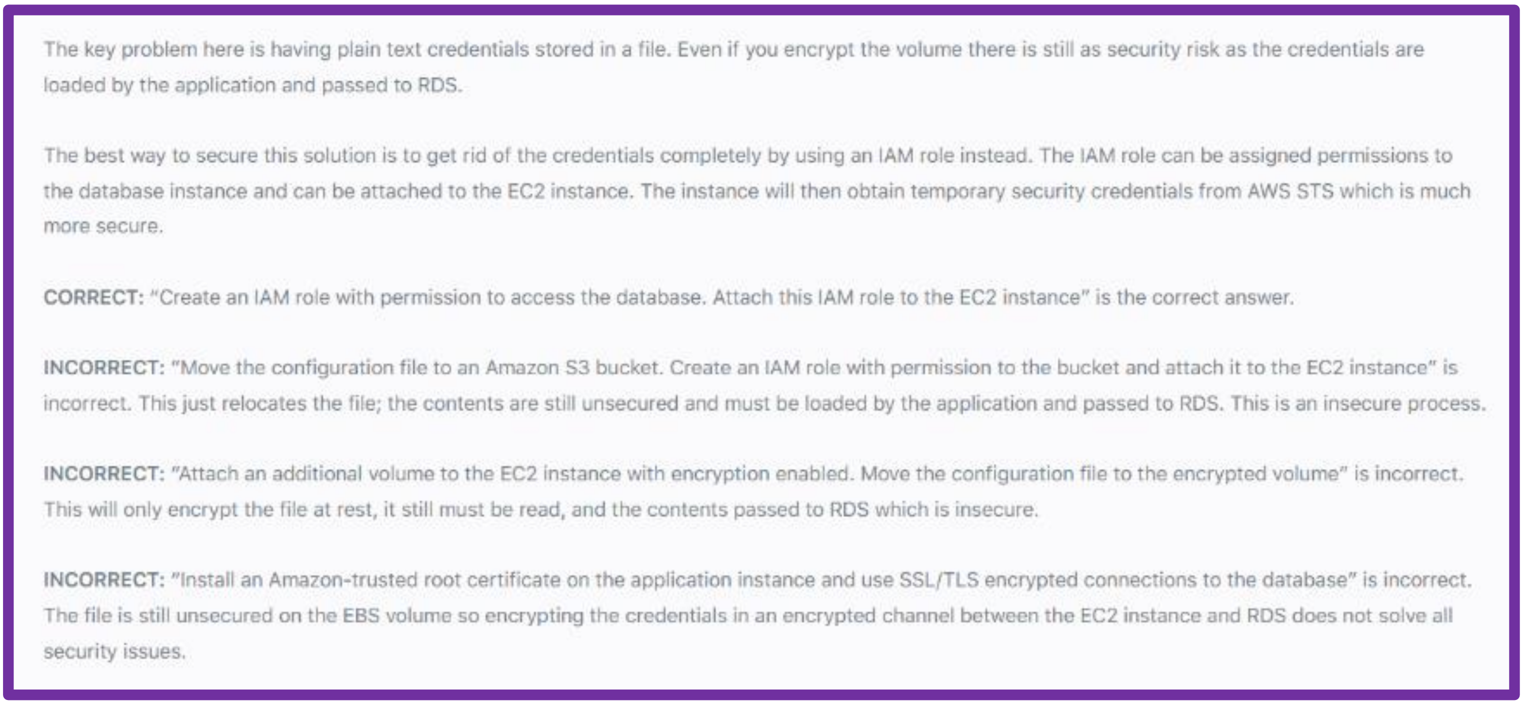
○ Move the configuration file to an Amazon S3 bucket. Create an IAM role with permission to the bucket and attach it to the EC2 instance.

○ Attach an additional volume to the EC2 instance with encryption enabled. Move the configuration file to the encrypted volume.

● Install an Amazon-trusted root certificate on the application instance and use SSL/TLS encrypted connections to the database.

○ Create an IAM role with permission to access the database. Attach this IAM role to the EC2 instance.

`



The key problem here is having plain text credentials stored in a file. Even if you encrypt the volume there is still as security risk as the credentials are loaded by the application and passed to RDS.

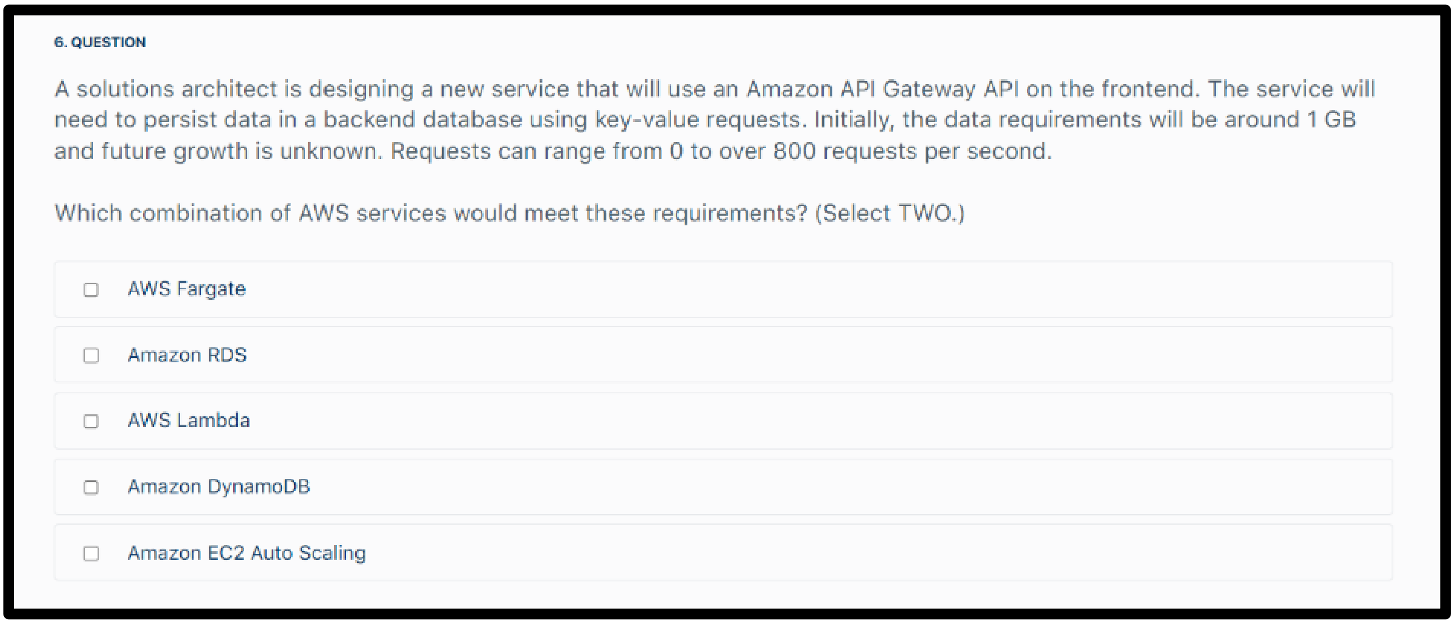
The best way to secure this solution is to get rid of the credentials completely by using an IAM role instead. The IAM role can be assigned permissions to the database instance and can be attached to the EC2 instance. The instance will then obtain temporary security credentials from AWS STS which is much more secure.

CORRECT: "Create an IAM role with permission to access the database. Attach this IAM role to the EC2 instance" is the correct answer.

INCORRECT: "Move the configuration file to an Amazon S3 bucket. Create an IAM role with permission to the bucket and attach it to the EC2 instance" is incorrect. This just relocates the file; the contents are still unsecured and must be loaded by the application and passed to RDS. This is an insecure process.

INCORRECT: "Attach an additional volume to the EC2 instance with encryption enabled. Move the configuration file to the encrypted volume" is incorrect. This will only encrypt the file at rest, it still must be read, and the contents passed to RDS which is insecure.

INCORRECT: "Install an Amazon-trusted root certificate on the application instance and use SSL/TLS encrypted connections to the database" is incorrect. The file is still unsecured on the EBS volume so encrypting the credentials in an encrypted channel between the EC2 instance and RDS does not solve all security issues.



6. QUESTION

A solutions architect is designing a new service that will use an Amazon API Gateway API on the frontend. The service will need to persist data in a backend database using key-value requests. Initially, the data requirements will be around 1 GB and future growth is unknown. Requests can range from 0 to over 800 requests per second.

Which combination of AWS services would meet these requirements? (Select TWO.)

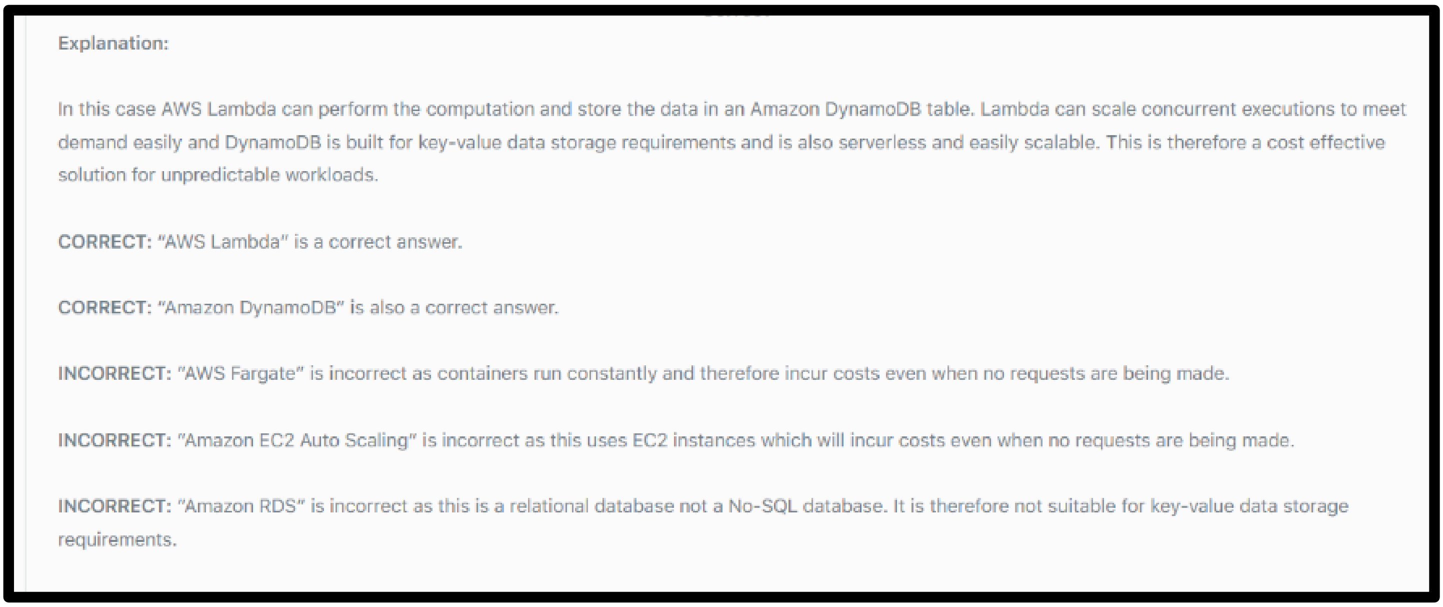
□ AWS Fargate

□ Amazon RDS

□ AWS Lambda

□ Amazon DynamoDB

□ Amazon EC2 Auto Scaling



Explanation:

In this case AWS Lambda can perform the computation and store the data in an Amazon DynamoDB table. Lambda can scale concurrent executions to meet demand easily and DynamoDB is built for key-value data storage requirements and is also serverless and easily scalable. This is therefore a cost effective solution for unpredictable workloads.

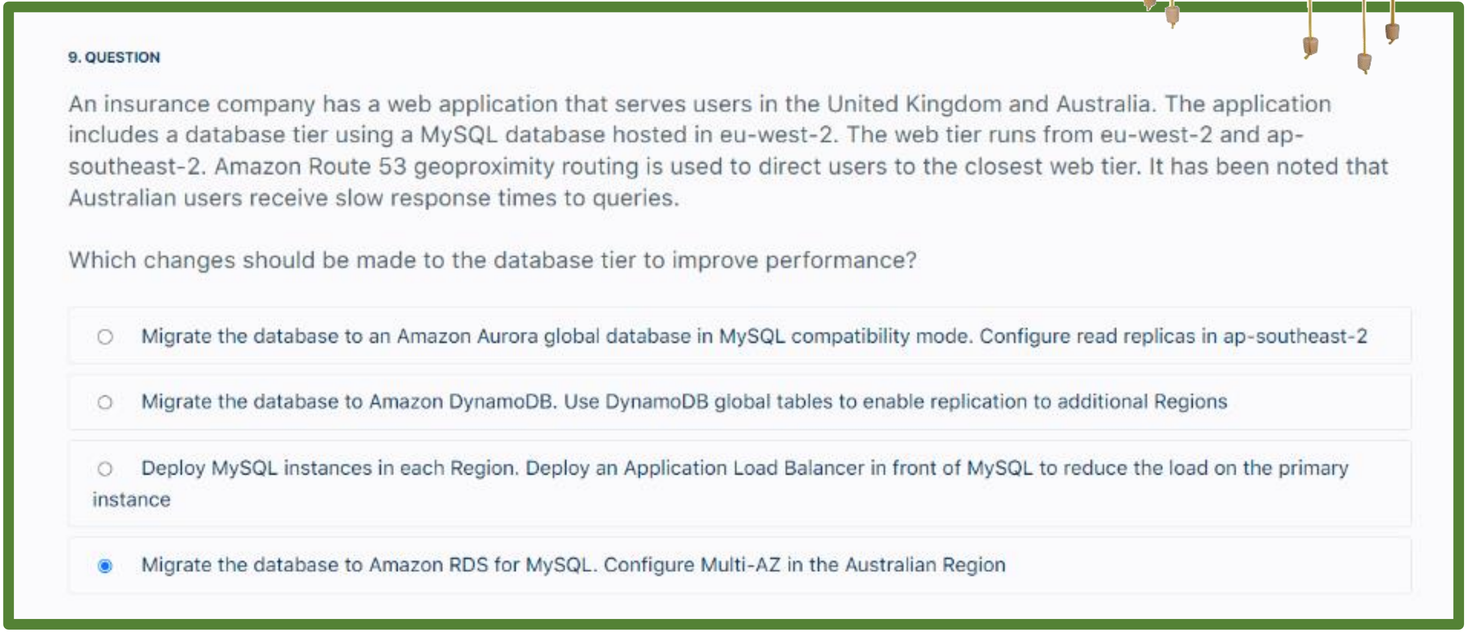
CORRECT: "AWS Lambda" is a correct answer.

CORRECT: "Amazon DynamoDB" is also a correct answer.

INCORRECT: "AWS Fargate" is incorrect as containers run constantly and therefore incur costs even when no requests are being made.

INCORRECT: "Amazon EC2 Auto Scaling" is incorrect as this uses EC2 instances which will incur costs even when no requests are being made.

INCORRECT: "Amazon RDS" is incorrect as this is a relational database not a No-SQL database. It is therefore not suitable for key-value data storage requirements.



9. QUESTION

An insurance company has a web application that serves users in the United Kingdom and Australia. The application includes a database tier using a MySQL database hosted in eu-west-2. The web tier runs from eu-west-2 and ap-southeast-2. Amazon Route 53 geoproximity routing is used to direct users to the closest web tier. It has been noted that Australian users receive slow response times to queries.

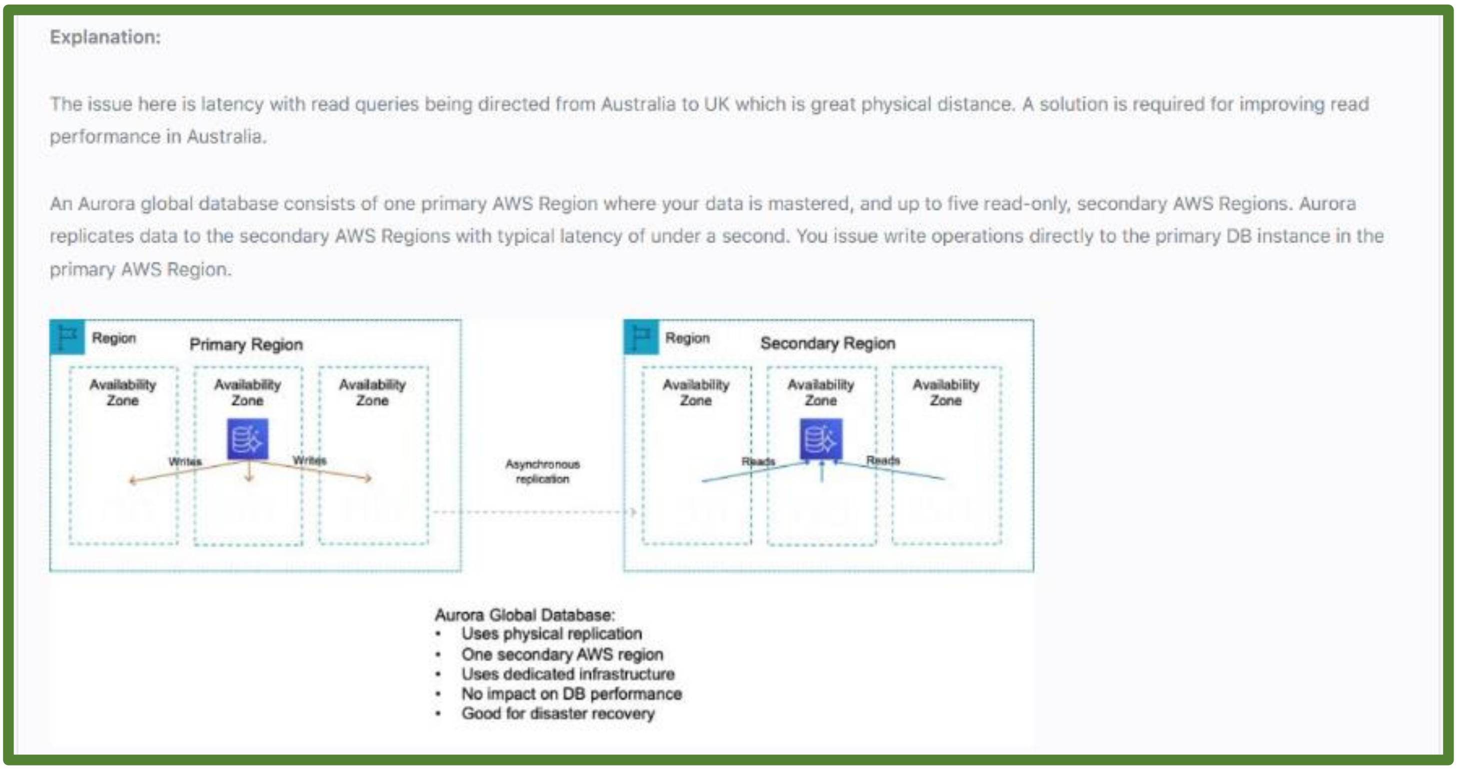
Which changes should be made to the database tier to improve performance?

○ Migrate the database to an Amazon Aurora global database in MySQL compatibility mode. Configure read replicas in ap-southeast-2

○ Migrate the database to Amazon DynamoDB. Use DynamoDB global tables to enable replication to additional Regions

○ Deploy MySQL instances in each Region. Deploy an Application Load Balancer in front of MySQL to reduce the load on the primary instance

● Migrate the database to Amazon RDS for MySQL. Configure Multi-AZ in the Australian Region



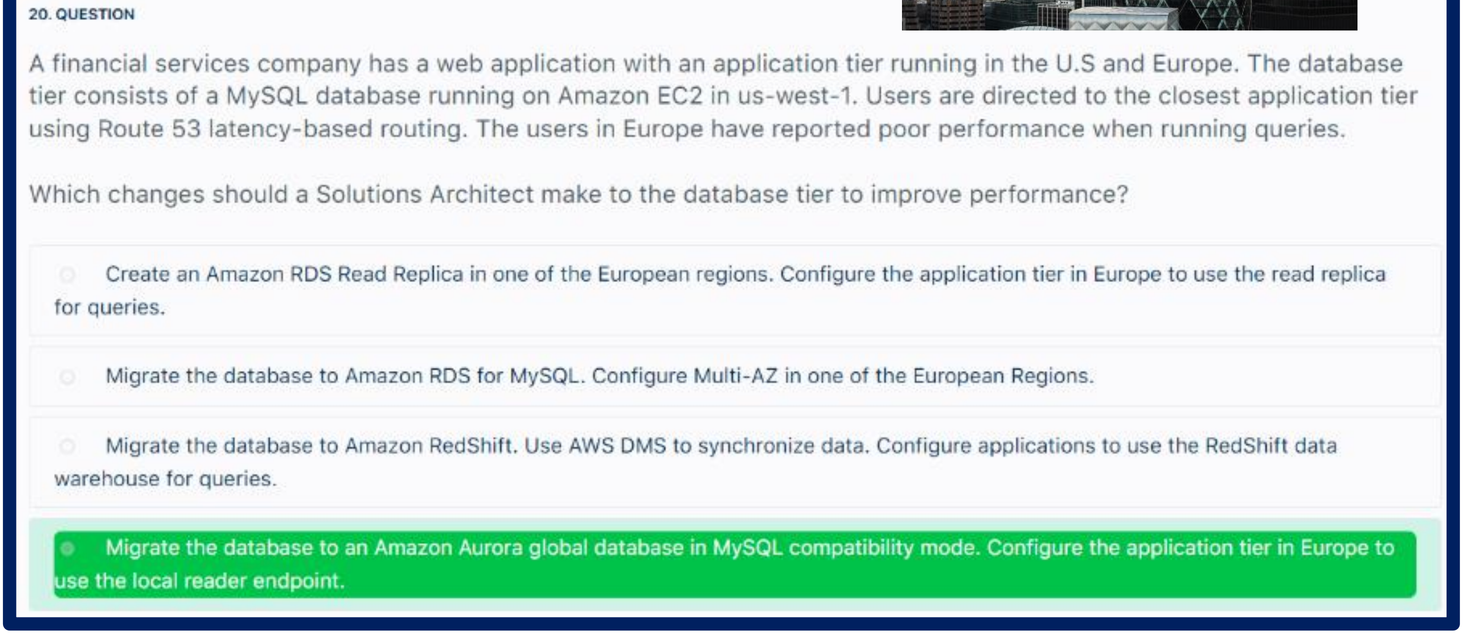
Explanation:

The issue here is latency with read queries being directed from Australia to UK which is great physical distance. A solution is required for improving read performance in Australia.

An Aurora global database consists of one primary AWS Region where your data is mastered, and up to five read-only, secondary AWS Regions. Aurora replicates data to the secondary AWS Regions with typical latency of under a second. You issue write operations directly to the primary DB instance in the primary AWS Region.

Aurora Global Database:

- Uses physical replication
- One secondary AWS region
- Uses dedicated infrastructure
- No impact on DB performance
- Good for disaster recovery



20. QUESTION

A financial services company has a web application with an application tier running in the U.S and Europe. The database tier consists of a MySQL database running on Amazon EC2 in us-west-1. Users are directed to the closest application tier using Route 53 latency-based routing. The users in Europe have reported poor performance when running queries.

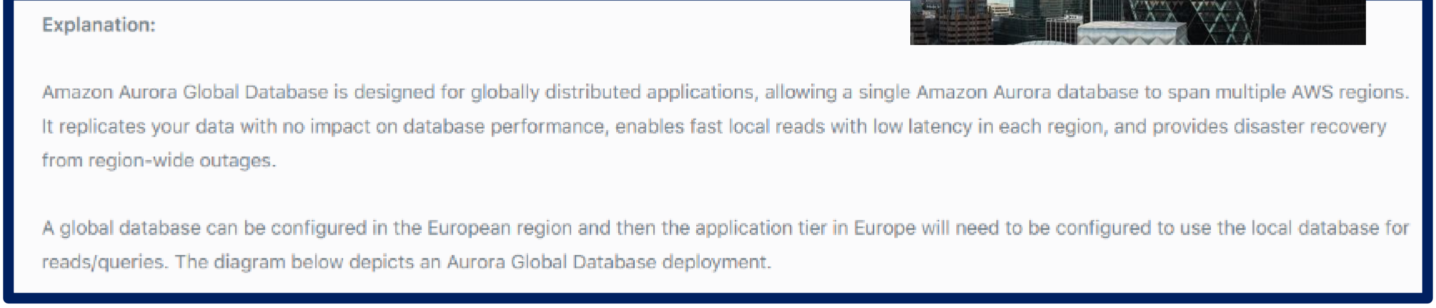
Which changes should a Solutions Architect make to the database tier to improve performance?

○ Create an Amazon RDS Read Replica in one of the European regions. Configure the application tier in Europe to use the read replica for queries.

○ Migrate the database to Amazon RDS for MySQL. Configure Multi-AZ in one of the European Regions.

○ Migrate the database to Amazon RedShift. Use AWS DMS to synchronize data. Configure applications to use the RedShift data warehouse for queries.

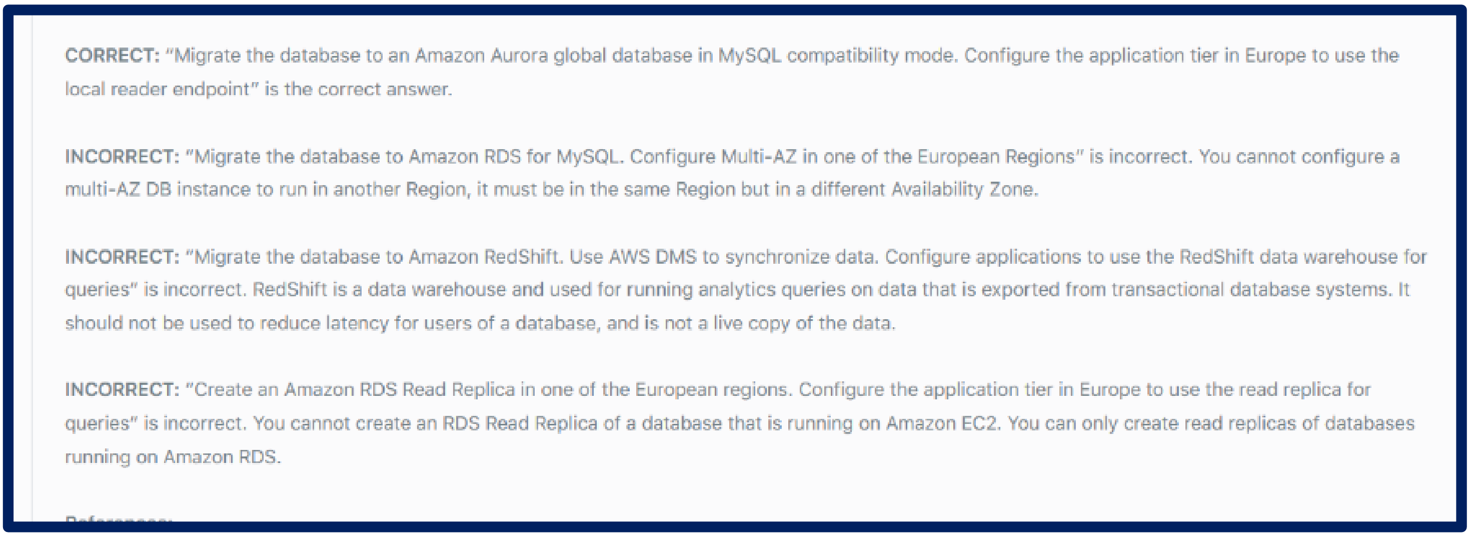
● Migrate the database to an Amazon Aurora global database in MySQL compatibility mode. Configure the application tier in Europe to use the local reader endpoint.



Explanation:

Amazon Aurora Global Database is designed for globally distributed applications, allowing a single Amazon Aurora database to span multiple AWS regions. It replicates your data with no impact on database performance, enables fast local reads with low latency in each region, and provides disaster recovery from region-wide outages.

A global database can be configured in the European region and then the application tier in Europe will need to be configured to use the local database for reads/queries. The diagram below depicts an Aurora Global Database deployment.



CORRECT: "Migrate the database to an Amazon Aurora global database in MySQL compatibility mode. Configure the application tier in Europe to use the local reader endpoint" is the correct answer.

INCORRECT: "Migrate the database to Amazon RDS for MySQL. Configure Multi-AZ in one of the European Regions" is incorrect. You cannot configure a multi-AZ DB instance to run in another Region, it must be in the same Region but in a different Availability Zone.

INCORRECT: "Migrate the database to Amazon RedShift. Use AWS DMS to synchronize data. Configure applications to use the RedShift data warehouse for queries" is incorrect. RedShift is a data warehouse and used for running analytics queries on data that is exported from transactional database systems. It should not be used to reduce latency for users of a database, and is not a live copy of the data.

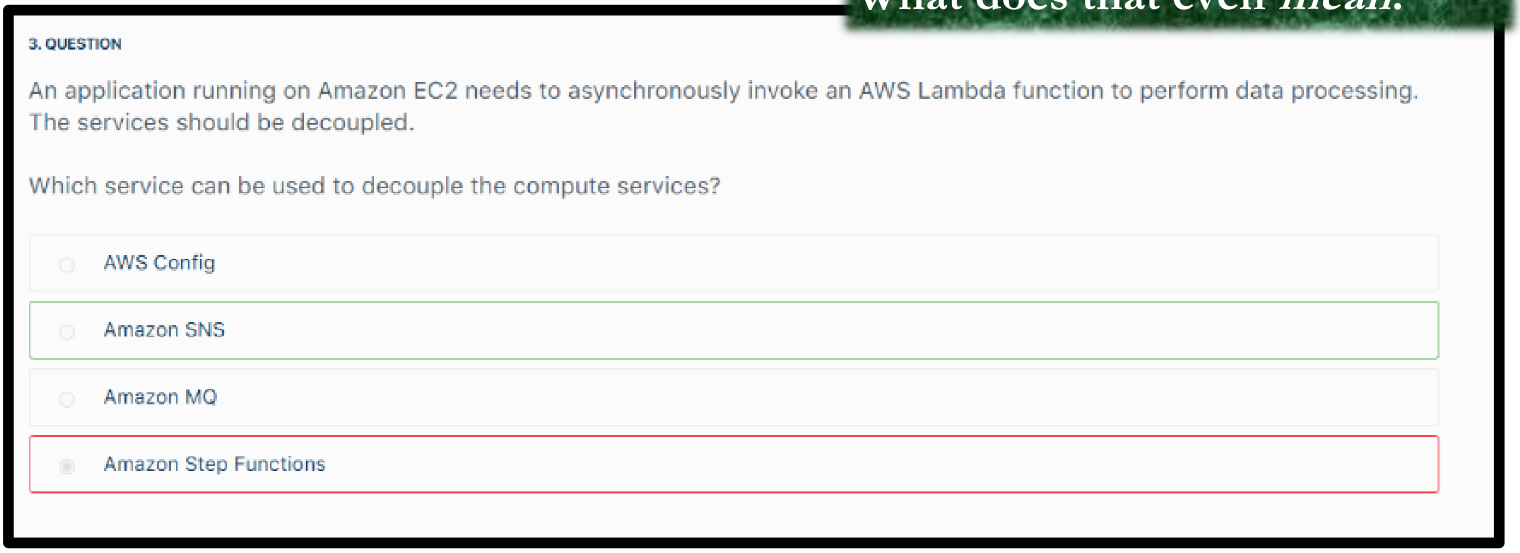
INCORRECT: "Create an Amazon RDS Read Replica in one of the European regions. Configure the application tier in Europe to use the read replica for queries" is incorrect. You cannot create an RDS Read Replica of a database that is running on Amazon EC2. You can only create read replicas of databases running on Amazon RDS.

References:

Simple Notification Service

“Asynchronously invoke”

What does that even *mean*?



3. QUESTION

An application running on Amazon EC2 needs to asynchronously invoke an AWS Lambda function to perform data processing. The services should be decoupled.

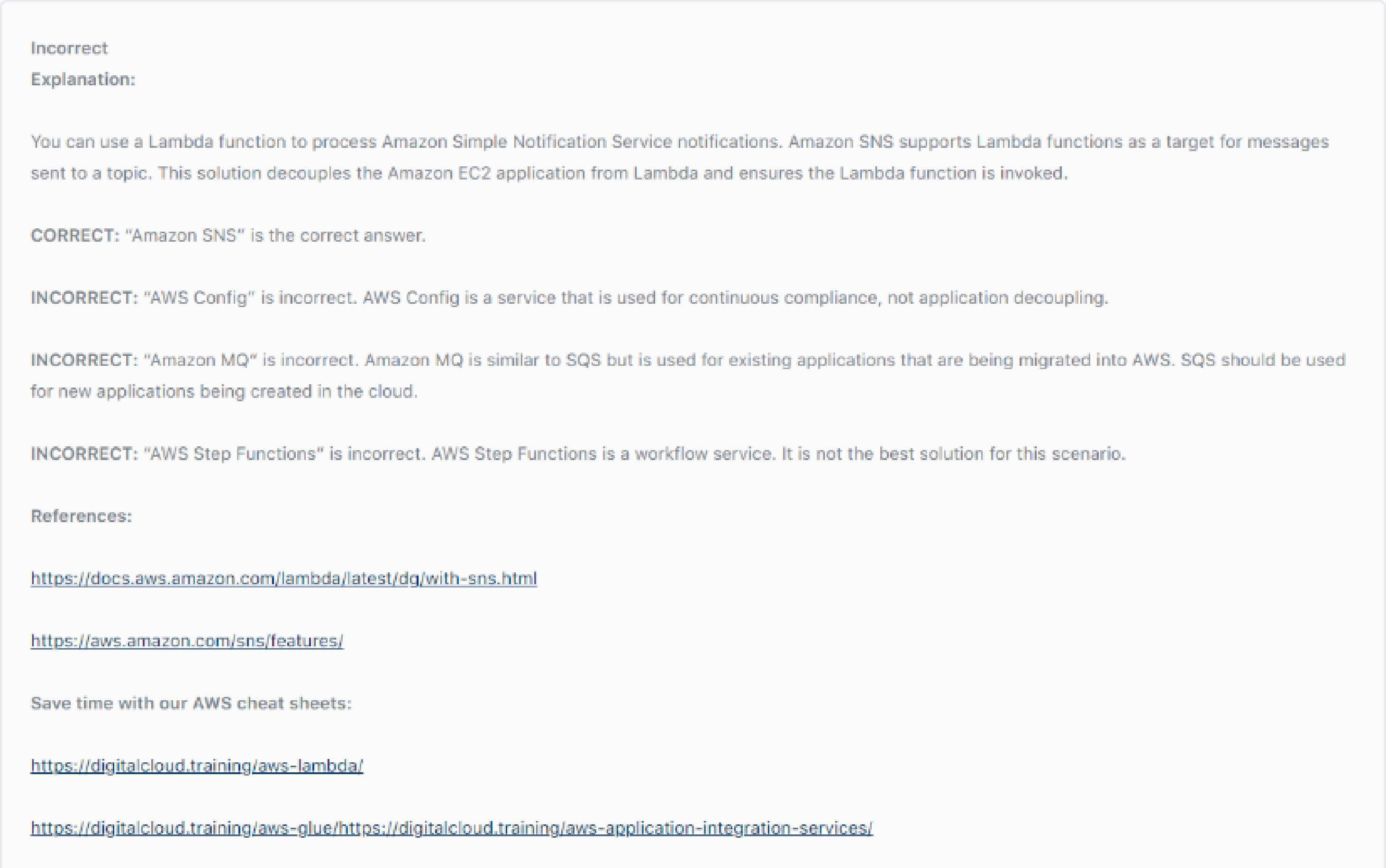
Which service can be used to decouple the compute services?

○ AWS Config

○ Amazon SNS

○ Amazon MQ

● Amazon Step Functions



Incorrect

Explanation:

You can use a Lambda function to process Amazon Simple Notification Service notifications. Amazon SNS supports Lambda functions as a target for messages sent to a topic. This solution decouples the Amazon EC2 application from Lambda and ensures the Lambda function is invoked.

CORRECT: "Amazon SNS" is the correct answer.

INCORRECT: "AWS Config" is incorrect. AWS Config is a service that is used for continuous compliance, not application decoupling.

INCORRECT: "Amazon MQ" is incorrect. Amazon MQ is similar to SQS but is used for existing applications that are being migrated into AWS. SQS should be used for new applications being created in the cloud.

INCORRECT: "AWS Step Functions" is incorrect. AWS Step Functions is a workflow service. It is not the best solution for this scenario.

References:

https://docs.aws.amazon.com/lambda/latest/dg/with-sns.html

https://aws.amazon.com/sns/features/

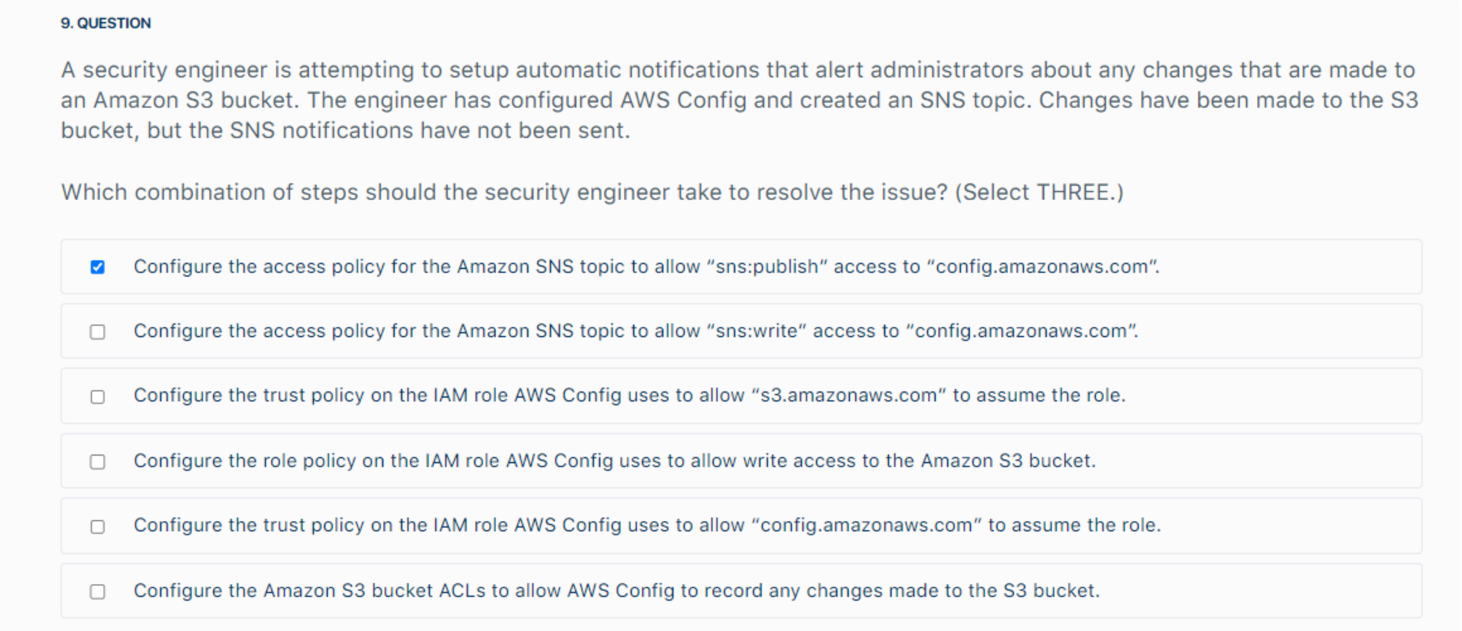
Save time with our AWS cheat sheets:

https://digitalcloud.training/aws-lambda/

https://digitalcloud.training/aws-glue/https://digitalcloud.training/aws-application-integration-services/

I really don't like this question. I understand what synchronous replication is, and how it differs from asynchronously replication. But how do we "asynchronously invoke" a lambda function? What does it even mean to "asynchronously invoke" something.

The fact is that Neal Davis, nor anyone else, ever divulges what a workflow actually is. No one explains what the necessary and sufficient conditions of a workflow are. And yet, that is what this question seems to hang on: the concept of a **workflow**.

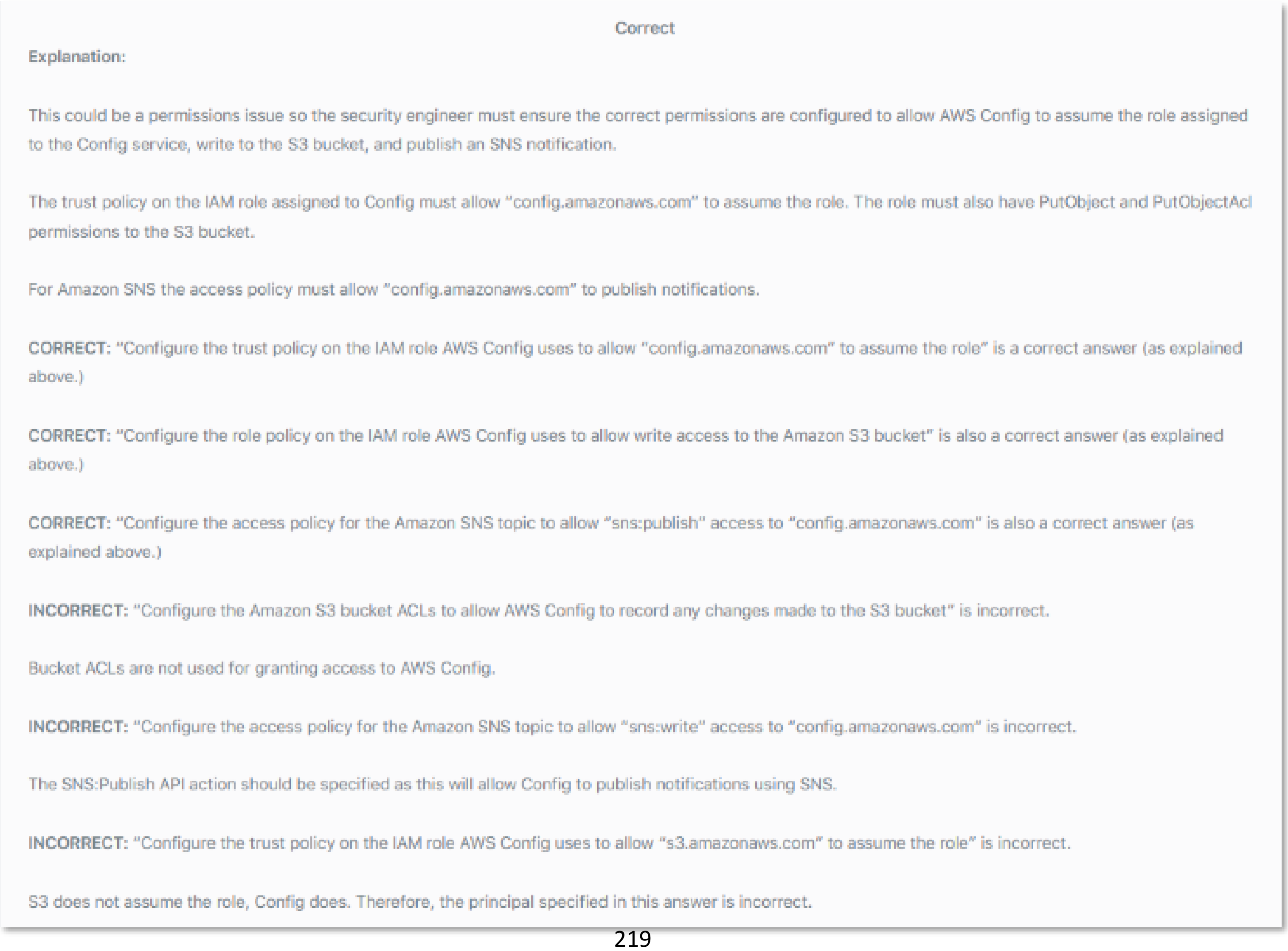


9. QUESTION

A security engineer is attempting to setup automatic notifications that alert administrators about any changes that are made to an Amazon S3 bucket. The engineer has configured AWS Config and created an SNS topic. Changes have been made to the S3 bucket, but the SNS notifications have not been sent.

Which combination of steps should the security engineer take to resolve the issue? (Select THREE.)

- ☑ Configure the access policy for the Amazon SNS topic to allow "sns:publish" access to "config.amazonaws.com".
- ☐ Configure the access policy for the Amazon SNS topic to allow "sns:write" access to "config.amazonaws.com".
- ☐ Configure the trust policy on the IAM role AWS Config uses to allow "s3.amazonaws.com" to assume the role.
- ☐ Configure the role policy on the IAM role AWS Config uses to allow write access to the Amazon S3 bucket.
- ☐ Configure the trust policy on the IAM role AWS Config uses to allow "config.amazonaws.com" to assume the role.
- ☐ Configure the Amazon S3 bucket ACLs to allow AWS Config to record any changes made to the S3 bucket.



Correct

Explanation:

This could be a permissions issue so the security engineer must ensure the correct permissions are configured to allow AWS Config to assume the role assigned to the Config service, write to the S3 bucket, and publish an SNS notification.

The trust policy on the IAM role assigned to Config must allow "config.amazonaws.com" to assume the role. The role must also have PutObject and PutObjectAcl permissions to the S3 bucket.

For Amazon SNS the access policy must allow "config.amazonaws.com" to publish notifications.

CORRECT: "Configure the trust policy on the IAM role AWS Config uses to allow "config.amazonaws.com" to assume the role" is a correct answer (as explained above.)

CORRECT: "Configure the role policy on the IAM role AWS Config uses to allow write access to the Amazon S3 bucket" is also a correct answer (as explained above.)

CORRECT: "Configure the access policy for the Amazon SNS topic to allow "sns:publish" access to "config.amazonaws.com" is also a correct answer (as explained above.)

INCORRECT: "Configure the Amazon S3 bucket ACLs to allow AWS Config to record any changes made to the S3 bucket" is incorrect.

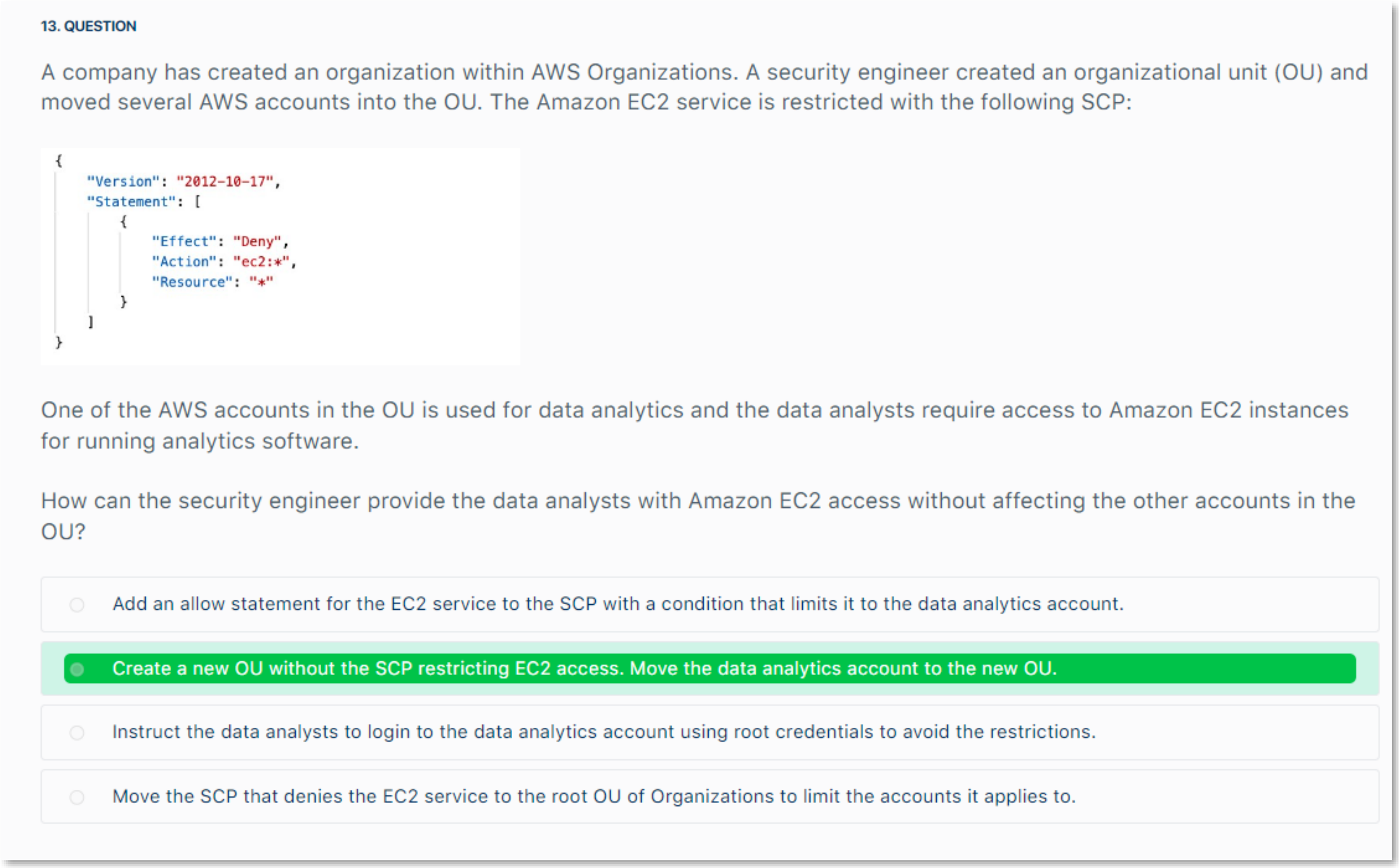
Bucket ACLs are not used for granting access to AWS Config.

INCORRECT: "Configure the access policy for the Amazon SNS topic to allow "sns:write" access to "config.amazonaws.com" is incorrect.

The SNS:Publish API action should be specified as this will allow Config to publish notifications using SNS.

INCORRECT: "Configure the trust policy on the IAM role AWS Config uses to allow "s3.amazonaws.com" to assume the role" is incorrect.

S3 does not assume the role, Config does. Therefore, the principal specified in this answer is incorrect.



13. QUESTION

A company has created an organization within AWS Organizations. A security engineer created an organizational unit (OU) and moved several AWS accounts into the OU. The Amazon EC2 service is restricted with the following SCP:

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Deny",
            "Action": "ec2:*",
            "Resource": "*"
        }
    ]
}
```

One of the AWS accounts in the OU is used for data analytics and the data analysts require access to Amazon EC2 instances for running analytics software.

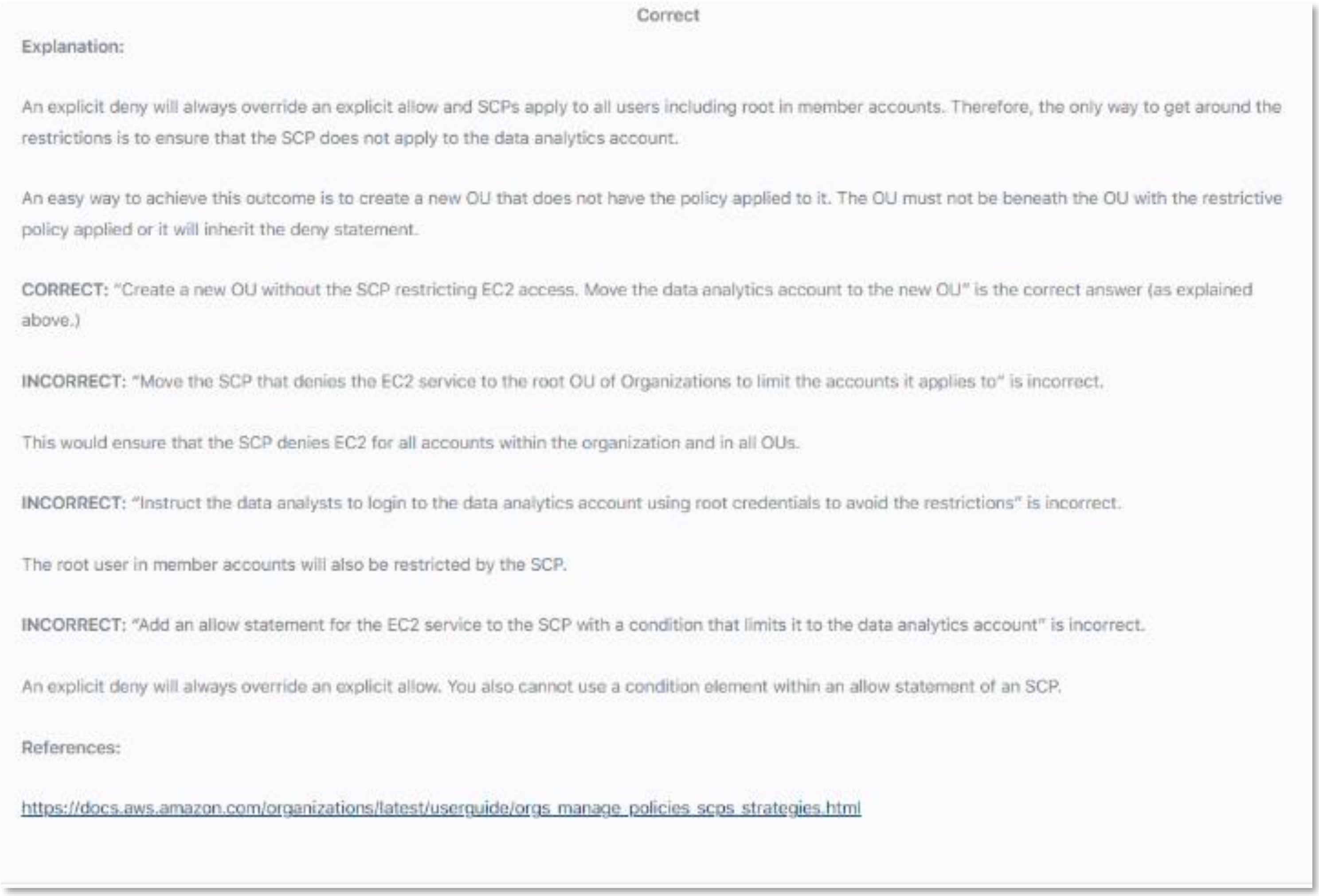
How can the security engineer provide the data analysts with Amazon EC2 access without affecting the other accounts in the OU?

○ Add an allow statement for the EC2 service to the SCP with a condition that limits it to the data analytics account.

● Create a new OU without the SCP restricting EC2 access. Move the data analytics account to the new OU.

○ Instruct the data analysts to login to the data analytics account using root credentials to avoid the restrictions.

○ Move the SCP that denies the EC2 service to the root OU of Organizations to limit the accounts it applies to.



Correct

Explanation:

An explicit deny will always override an explicit allow and SCPs apply to all users including root in member accounts. Therefore, the only way to get around the restrictions is to ensure that the SCP does not apply to the data analytics account.

An easy way to achieve this outcome is to create a new OU that does not have the policy applied to it. The OU must not be beneath the OU with the restrictive policy applied or it will inherit the deny statement.

CORRECT: "Create a new OU without the SCP restricting EC2 access. Move the data analytics account to the new OU" is the correct answer (as explained above.)

INCORRECT: "Move the SCP that denies the EC2 service to the root OU of Organizations to limit the accounts it applies to" is incorrect.

This would ensure that the SCP denies EC2 for all accounts within the organization and in all OUs.

INCORRECT: "Instruct the data analysts to login to the data analytics account using root credentials to avoid the restrictions" is incorrect.

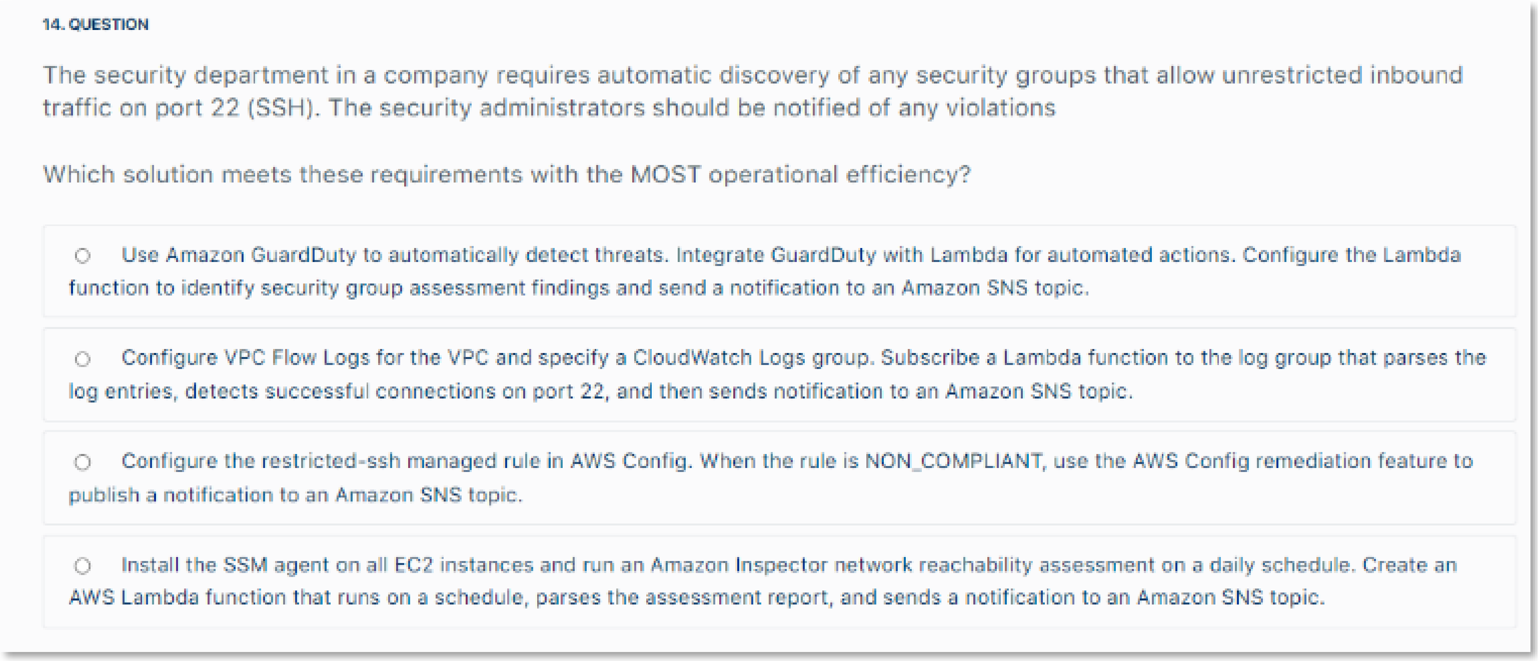
The root user in member accounts will also be restricted by the SCP.

INCORRECT: "Add an allow statement for the EC2 service to the SCP with a condition that limits it to the data analytics account" is incorrect.

An explicit deny will always override an explicit allow. You also cannot use a condition element within an allow statement of an SCP.

References:

https://docs.aws.amazon.com/organizations/latest/userguide/orgs_manage_policies_scps_strategies.html



14. QUESTION

The security department in a company requires automatic discovery of any security groups that allow unrestricted inbound traffic on port 22 (SSH). The security administrators should be notified of any violations

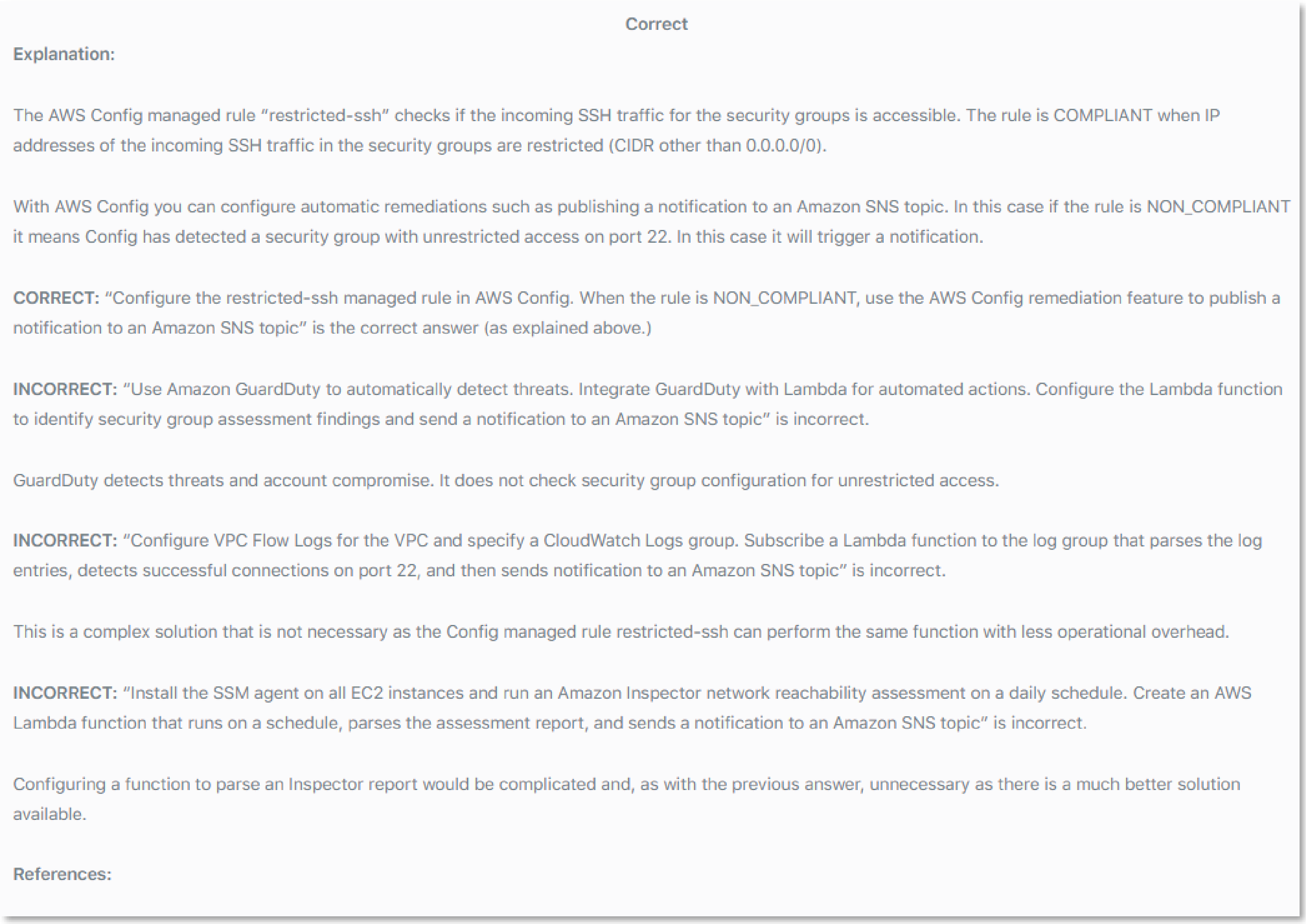
Which solution meets these requirements with the MOST operational efficiency?

○ Use Amazon GuardDuty to automatically detect threats. Integrate GuardDuty with Lambda for automated actions. Configure the Lambda function to identify security group assessment findings and send a notification to an Amazon SNS topic.

○ Configure VPC Flow Logs for the VPC and specify a CloudWatch Logs group. Subscribe a Lambda function to the log group that parses the log entries, detects successful connections on port 22, and then sends notification to an Amazon SNS topic.

○ Configure the restricted-ssh managed rule in AWS Config. When the rule is NON_COMPLIANT, use the AWS Config remediation feature to publish a notification to an Amazon SNS topic.

○ Install the SSM agent on all EC2 instances and run an Amazon Inspector network reachability assessment on a daily schedule. Create an AWS Lambda function that runs on a schedule, parses the assessment report, and sends a notification to an Amazon SNS topic.



Correct

Explanation:

The AWS Config managed rule "restricted-ssh" checks if the incoming SSH traffic for the security groups is accessible. The rule is COMPLIANT when IP addresses of the incoming SSH traffic in the security groups are restricted (CIDR other than 0.0.0.0/0).

With AWS Config you can configure automatic remediations such as publishing a notification to an Amazon SNS topic. In this case if the rule is NON_COMPLIANT it means Config has detected a security group with unrestricted access on port 22. In this case it will trigger a notification.

CORRECT: "Configure the restricted-ssh managed rule in AWS Config. When the rule is NON_COMPLIANT, use the AWS Config remediation feature to publish a notification to an Amazon SNS topic" is the correct answer (as explained above.)

INCORRECT: "Use Amazon GuardDuty to automatically detect threats. Integrate GuardDuty with Lambda for automated actions. Configure the Lambda function to identify security group assessment findings and send a notification to an Amazon SNS topic" is incorrect.

GuardDuty detects threats and account compromise. It does not check security group configuration for unrestricted access.

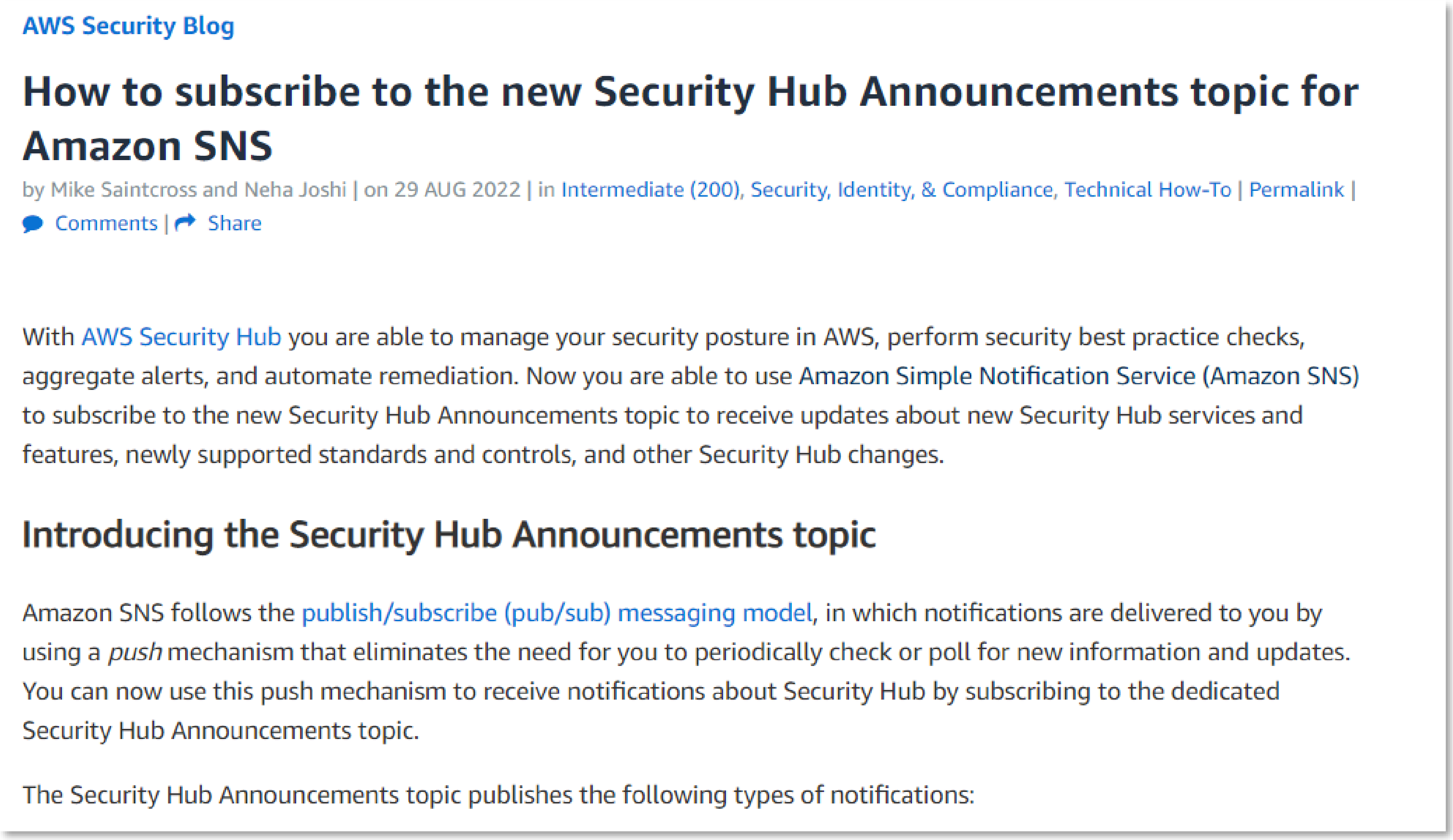
INCORRECT: "Configure VPC Flow Logs for the VPC and specify a CloudWatch Logs group. Subscribe a Lambda function to the log group that parses the log entries, detects successful connections on port 22, and then sends notification to an Amazon SNS topic" is incorrect.

This is a complex solution that is not necessary as the Config managed rule restricted-ssh can perform the same function with less operational overhead.

INCORRECT: "Install the SSM agent on all EC2 instances and run an Amazon Inspector network reachability assessment on a daily schedule. Create an AWS Lambda function that runs on a schedule, parses the assessment report, and sends a notification to an Amazon SNS topic" is incorrect.

Configuring a function to parse an Inspector report would be complicated and, as with the previous answer, unnecessary as there is a much better solution available.

References:



AWS Security Blog

How to subscribe to the new Security Hub Announcements topic for Amazon SNS

by Mike Saintcross and Neha Joshi | on 29 AUG 2022 | in Intermediate (200), Security, Identity, & Compliance, Technical How-To | Permalink | Comments | Share

With AWS Security Hub you are able to manage your security posture in AWS, perform security best practice checks, aggregate alerts, and automate remediation. Now you are able to use Amazon Simple Notification Service (Amazon SNS) to subscribe to the new Security Hub Announcements topic to receive updates about new Security Hub services and features, newly supported standards and controls, and other Security Hub changes.

Introducing the Security Hub Announcements topic

Amazon SNS follows the publish/subscribe (pub/sub) messaging model, in which notifications are delivered to you by using a *push* mechanism that eliminates the need for you to periodically check or poll for new information and updates. You can now use this push mechanism to receive notifications about Security Hub by subscribing to the dedicated Security Hub Announcements topic.

The Security Hub Announcements topic publishes the following types of notifications:

TPN

91. ***Phenomenon1*** – the tendency of X to Y.
92. ***Phen2*** – the tendency of X to Y.
93. ***Phen3*** – the tendency of X to Y.
94. ***Phen4*** – the tendency of X to Y.
95. ***Phen5*** – the tendency of X to Y.
96. ***Phen6*** – the tendency of X to Y.
97. ***Phen7*** – the tendency of X to Y.
98. ***Phen8*** – the tendency of X to Y.
99. ***Phen9*** – the tendency of X to Y.
100. ***Phen10*** – the tendency of X to Y.

Glossary

Term1

Description of what term means here.

Term2

Description of what term means here.

Term3

Description of what term means here.

Bibliography

I. Official

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.
https://aws.amazon.com/blogs/aws/subscribe-to-aws-daily-feature-updates-via-amazon-sns/

II. Unofficial

https://medium.com/awesome-cloud/aws-difference-between-multi-az-and-read-replicas-in-amazon-rds-60fe848ef53a

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

IV. General

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

IAM

←

EPARC

UGR

User

Group

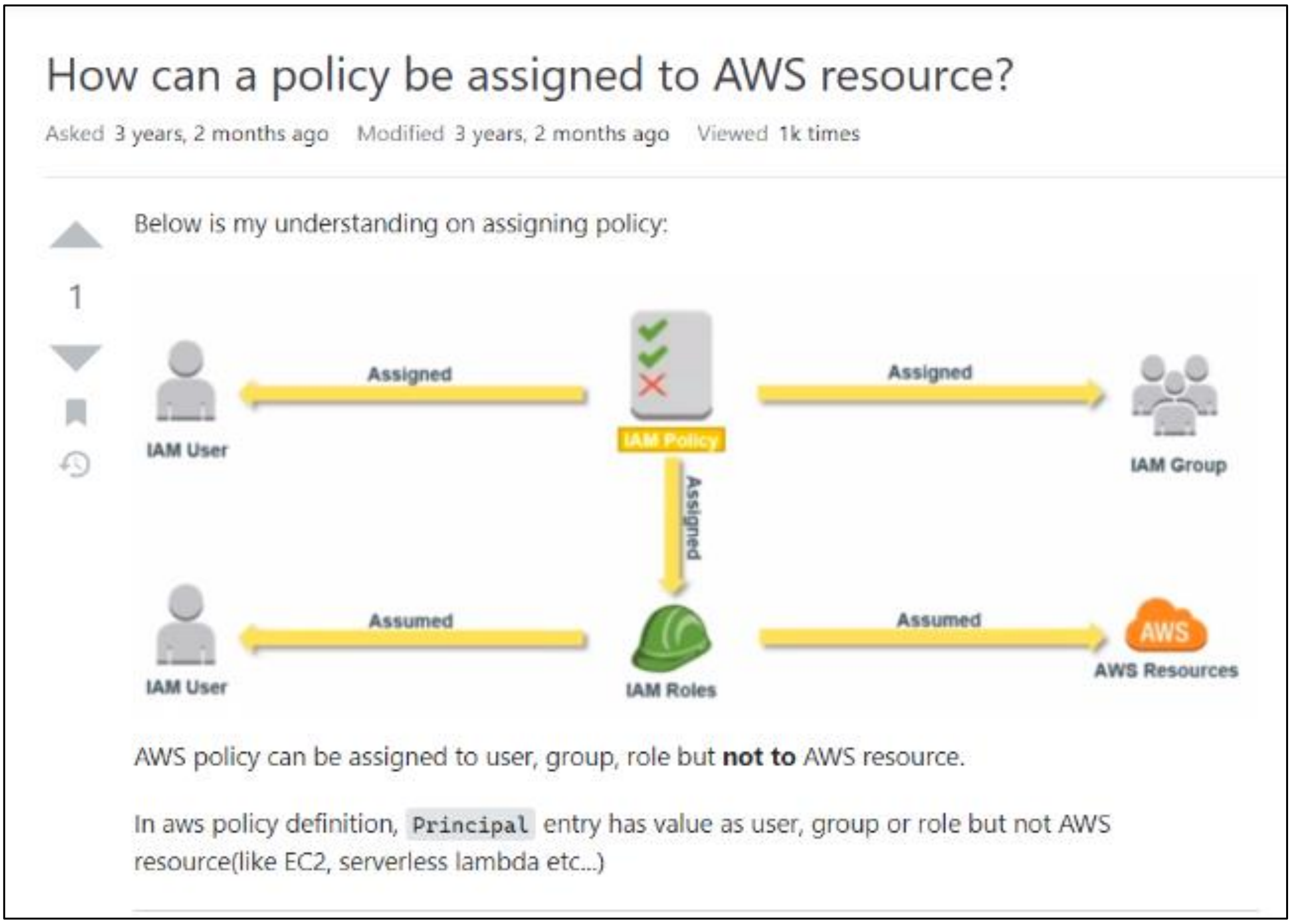
Role

SiD

Jeff Wierer explaining how IAM policies work in 2015.

Becky Weiss giving a presentation in 2022 at the Reinvent conference, entitled "AWS Identity and Access Management (IAM) deep dive"

This question on Stack Exchange has a number of good sources on it:



How can a policy be assigned to AWS resource?

Asked 3 years, 2 months ago Modified 3 years, 2 months ago Viewed 1k times

Below is my understanding on assigning policy:

AWS policy can be assigned to user, group, role but **not to** AWS resource.

In aws policy definition, `Principal` entry has value as user, group or role but not AWS resource(like EC2, serverless lambda etc...)

TABLE OF CONTENTS

About this guide

1. Control access to any resource

2. Why AWS IAM is so hard to use

Effective IAM for Amazon Web Services

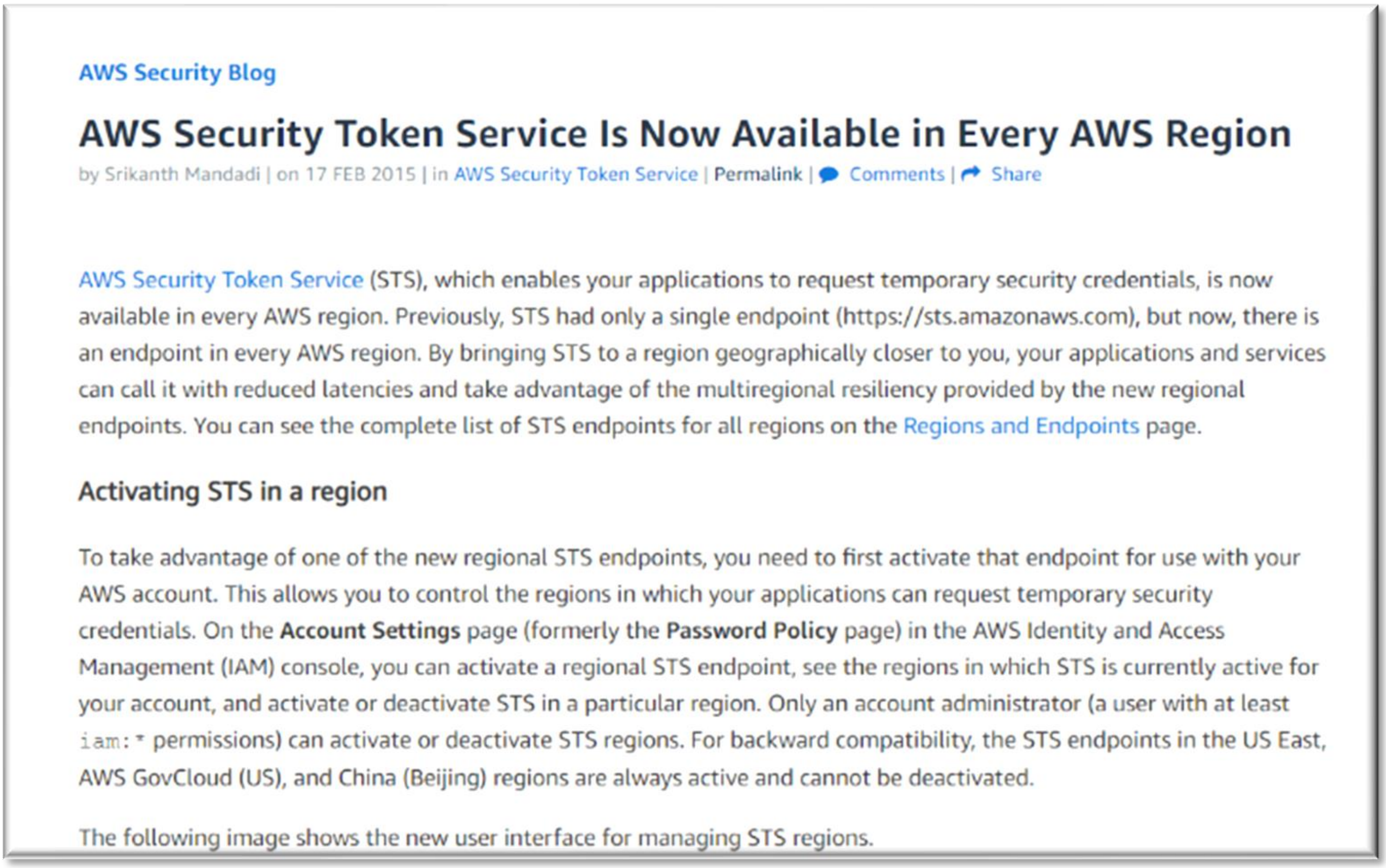
By Stephen Kuenzli

Effective IAM for Amazon Web Services is for Cloud engineers who design, develop, and review AWS IAM security policies in their daily work.

If you're struggling to deliver effective AWS security policies, this guide will help you understand why it's hard and how both you and your organization can use IAM well.

A free guide produced by Stephen Kuenzli (click here)

STS



AWS Security Blog

AWS Security Token Service Is Now Available in Every AWS Region

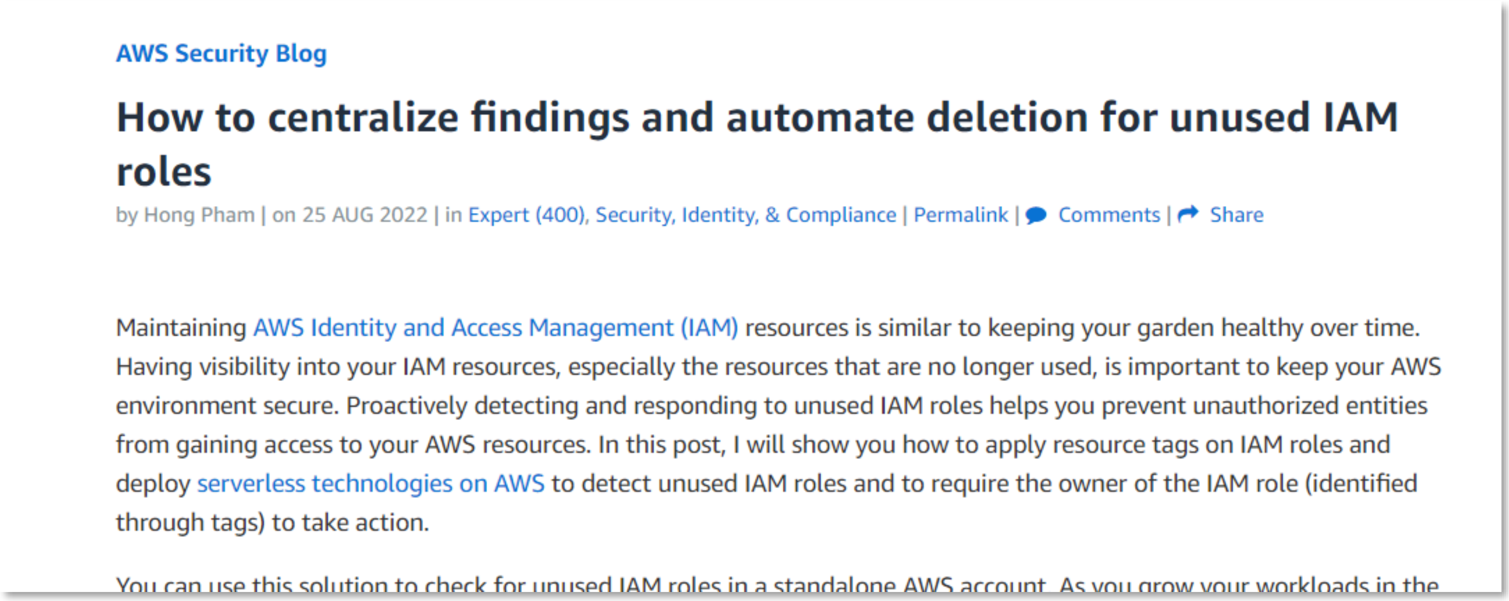
by Srikanth Mandadi | on 17 FEB 2015 | in AWS Security Token Service | Permalink | Comments | Share

AWS Security Token Service (STS), which enables your applications to request temporary security credentials, is now available in every AWS region. Previously, STS had only a single endpoint (https://sts.amazonaws.com), but now, there is an endpoint in every AWS region. By bringing STS to a region geographically closer to you, your applications and services can call it with reduced latencies and take advantage of the multiregional resiliency provided by the new regional endpoints. You can see the complete list of STS endpoints for all regions on the Regions and Endpoints page.

Activating STS in a region

To take advantage of one of the new regional STS endpoints, you need to first activate that endpoint for use with your AWS account. This allows you to control the regions in which your applications can request temporary security credentials. On the **Account Settings** page (formerly the **Password Policy** page) in the AWS Identity and Access Management (IAM) console, you can activate a regional STS endpoint, see the regions in which STS is currently active for your account, and activate or deactivate STS in a particular region. Only an account administrator (a user with at least `iam:*` permissions) can activate or deactivate STS regions. For backward compatibility, the STS endpoints in the US East, AWS GovCloud (US), and China (Beijing) regions are always active and cannot be deactivated.

The following image shows the new user interface for managing STS regions.



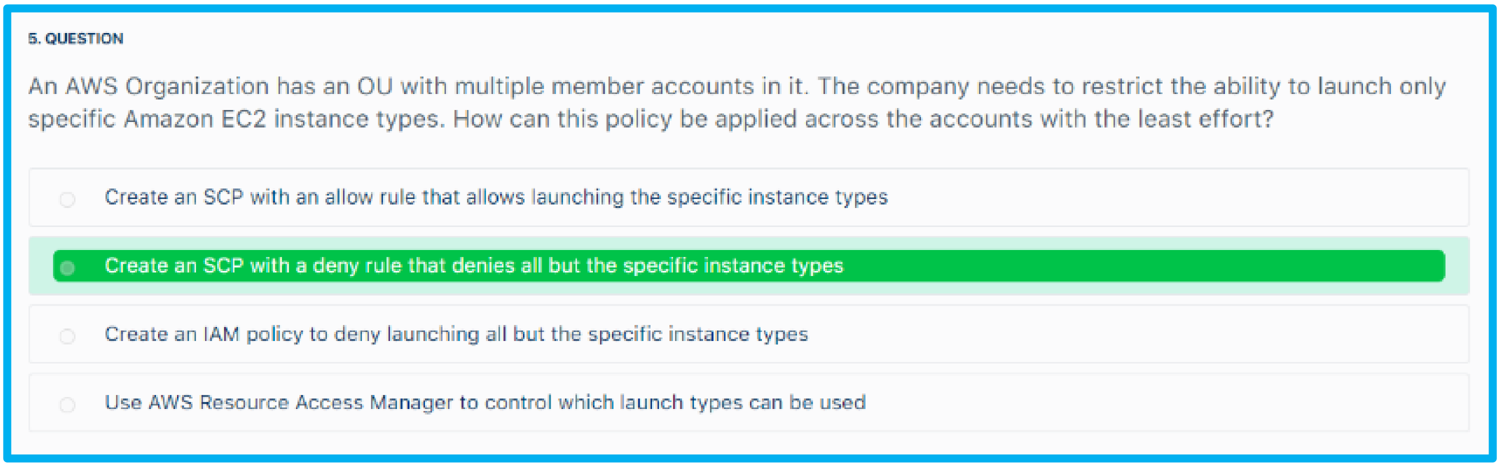
AWS Security Blog

How to centralize findings and automate deletion for unused IAM roles

by Hong Pham | on 25 AUG 2022 | in Expert (400), Security, Identity, & Compliance | Permalink | Comments | Share

Maintaining AWS Identity and Access Management (IAM) resources is similar to keeping your garden healthy over time. Having visibility into your IAM resources, especially the resources that are no longer used, is important to keep your AWS environment secure. Proactively detecting and responding to unused IAM roles helps you prevent unauthorized entities from gaining access to your AWS resources. In this post, I will show you how to apply resource tags on IAM roles and deploy serverless technologies on AWS to detect unused IAM roles and to require the owner of the IAM role (identified through tags) to take action.

You can use this solution to check for unused IAM roles in a standalone AWS account. As you grow your workloads in the



5. QUESTION

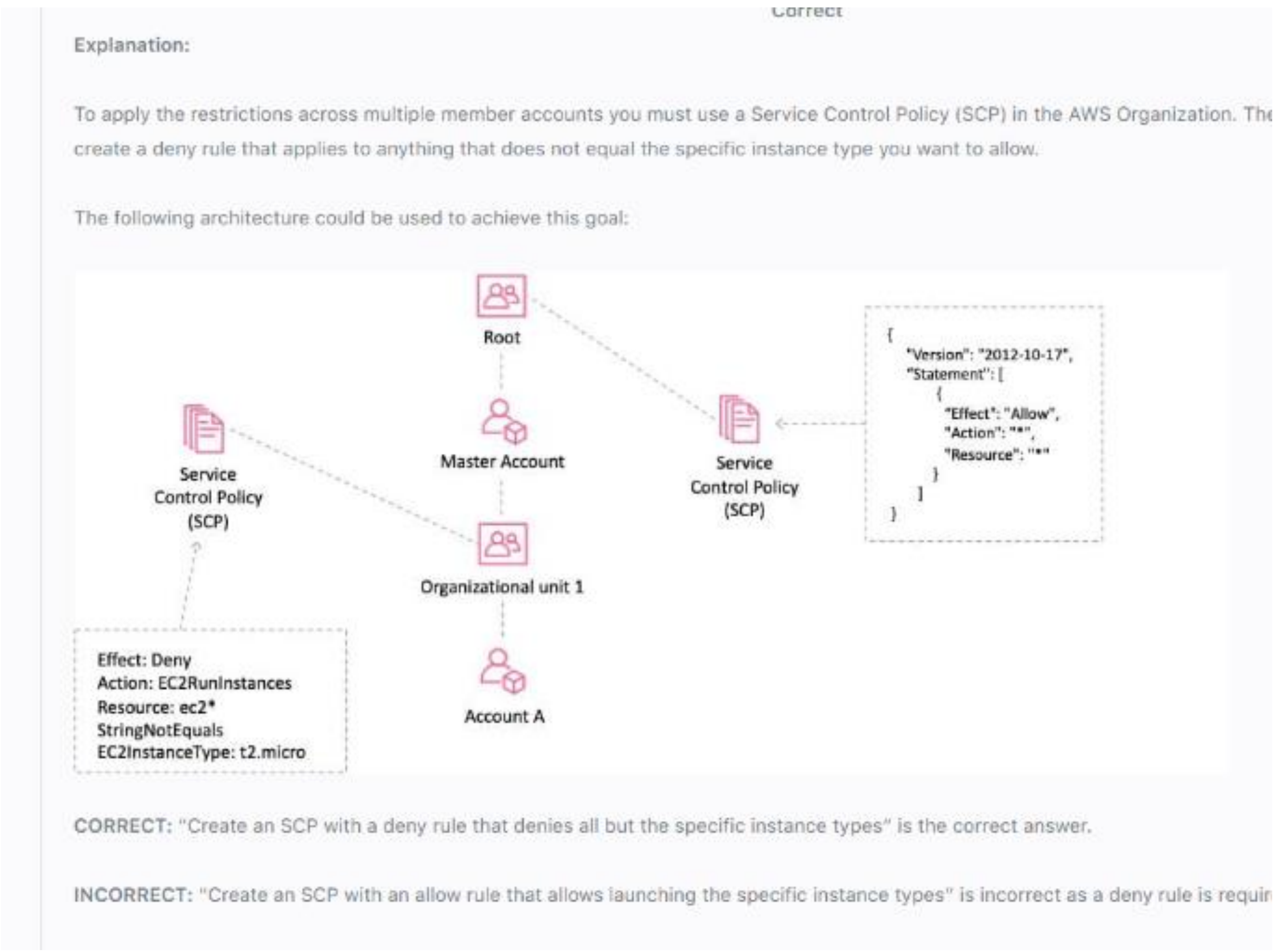
An AWS Organization has an OU with multiple member accounts in it. The company needs to restrict the ability to launch only specific Amazon EC2 instance types. How can this policy be applied across the accounts with the least effort?

○ Create an SCP with an allow rule that allows launching the specific instance types

● Create an SCP with a deny rule that denies all but the specific instance types

○ Create an IAM policy to deny launching all but the specific instance types

○ Use AWS Resource Access Manager to control which launch types can be used



Correct

Explanation:

To apply the restrictions across multiple member accounts you must use a Service Control Policy (SCP) in the AWS Organization. The create a deny rule that applies to anything that does not equal the specific instance type you want to allow.

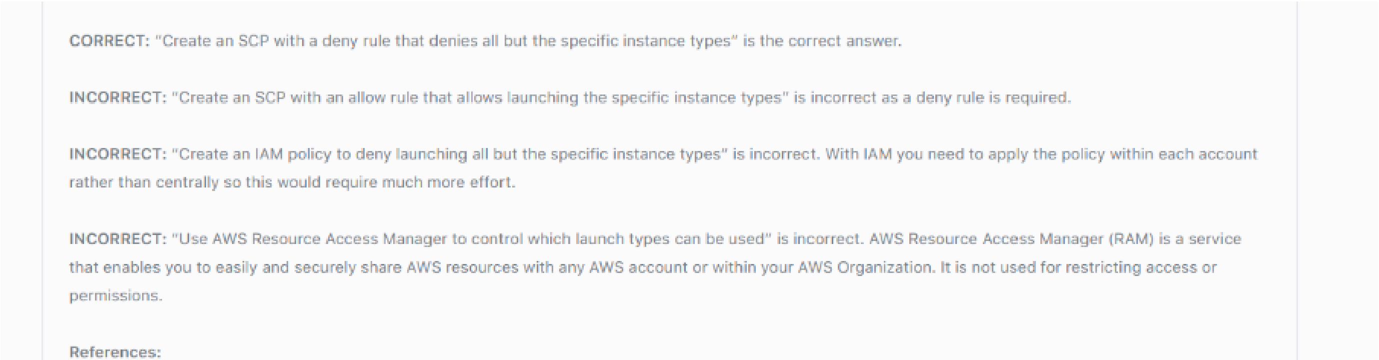
The following architecture could be used to achieve this goal:

```
{
  "Version": "2012-10-17",
  "Statement": [
      {
        "Effect": "Allow",
        "Action": "*",
        "Resource": "*"
      }
  ]
}
```

Effect: Deny
Action: EC2RunInstances
Resource: ec2*
StringNotEquals
EC2InstanceType: t2.micro

CORRECT: "Create an SCP with a deny rule that denies all but the specific instance types" is the correct answer.

INCORRECT: "Create an SCP with an allow rule that allows launching the specific instance types" is incorrect as a deny rule is requir



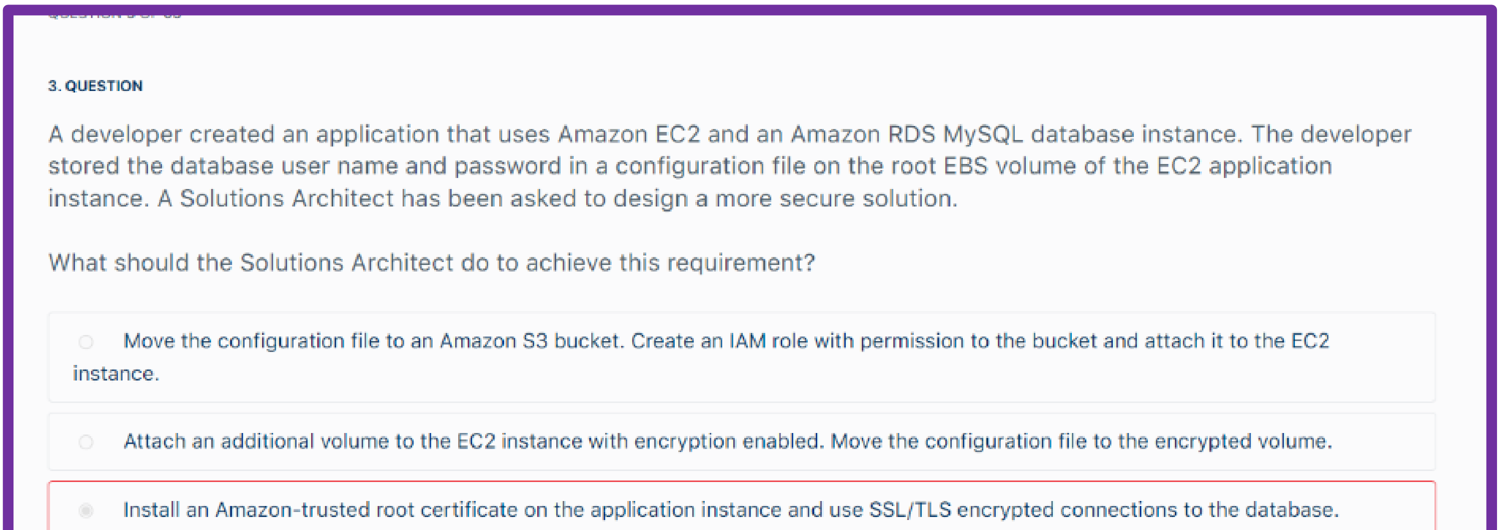
CORRECT: "Create an SCP with a deny rule that denies all but the specific instance types" is the correct answer.

INCORRECT: "Create an SCP with an allow rule that allows launching the specific instance types" is incorrect as a deny rule is required.

INCORRECT: "Create an IAM policy to deny launching all but the specific instance types" is incorrect. With IAM you need to apply the policy within each account rather than centrally so this would require much more effort.

INCORRECT: "Use AWS Resource Access Manager to control which launch types can be used" is incorrect. AWS Resource Access Manager (RAM) is a service that enables you to easily and securely share AWS resources with any AWS account or within your AWS Organization. It is not used for restricting access or permissions.

References:



3. QUESTION

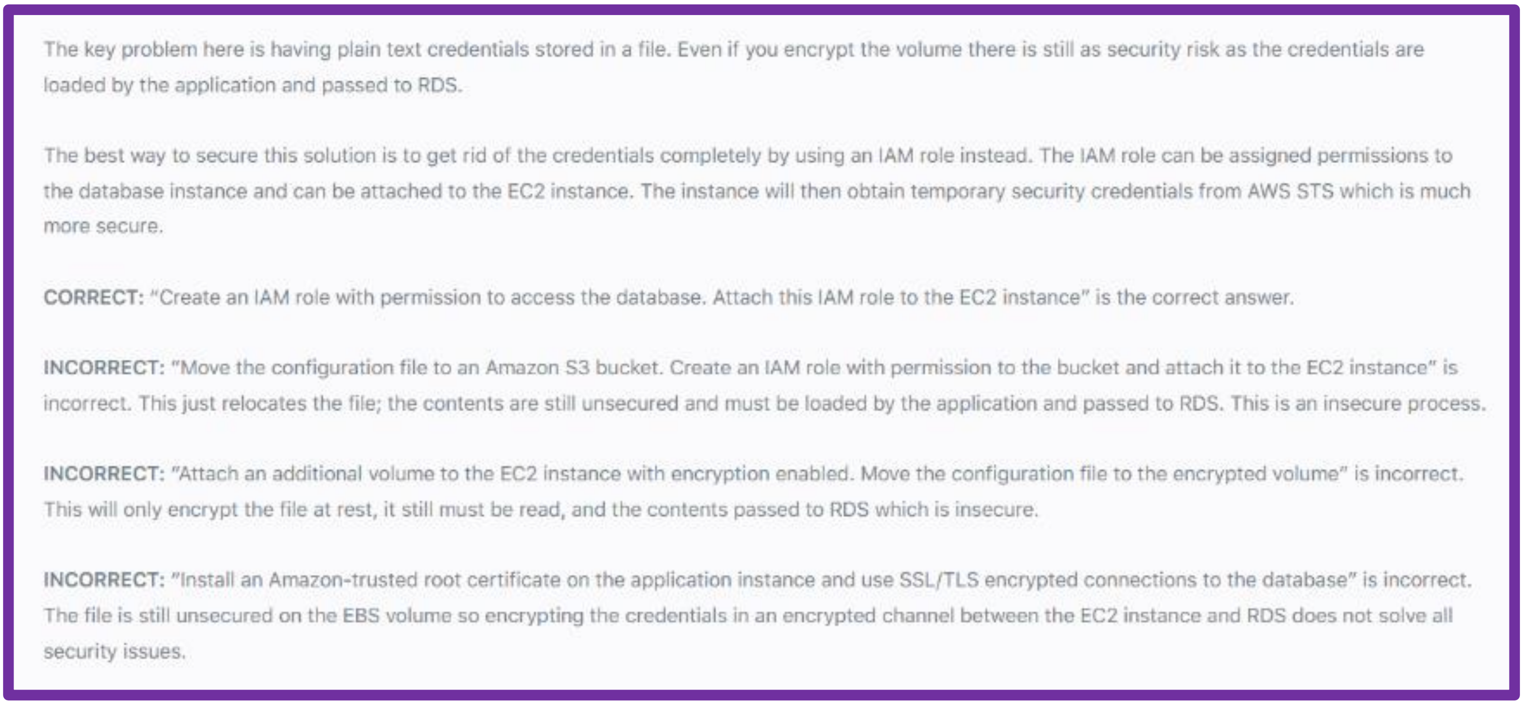
A developer created an application that uses Amazon EC2 and an Amazon RDS MySQL database instance. The developer stored the database user name and password in a configuration file on the root EBS volume of the EC2 application instance. A Solutions Architect has been asked to design a more secure solution.

What should the Solutions Architect do to achieve this requirement?

○ Move the configuration file to an Amazon S3 bucket. Create an IAM role with permission to the bucket and attach it to the EC2 instance.

○ Attach an additional volume to the EC2 instance with encryption enabled. Move the configuration file to the encrypted volume.

● Install an Amazon-trusted root certificate on the application instance and use SSL/TLS encrypted connections to the database.



The key problem here is having plain text credentials stored in a file. Even if you encrypt the volume there is still as security risk as the credentials are loaded by the application and passed to RDS.

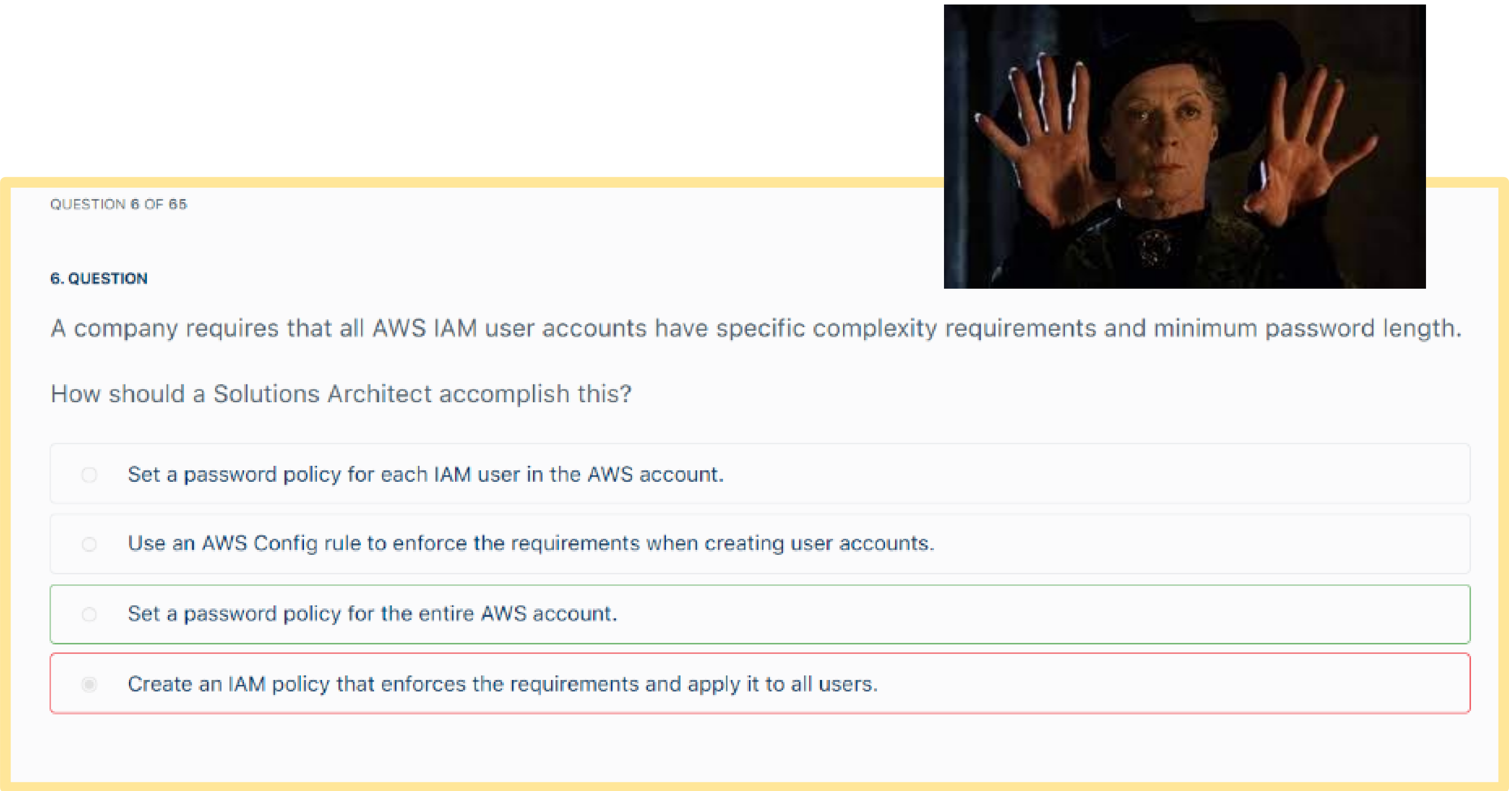
The best way to secure this solution is to get rid of the credentials completely by using an IAM role instead. The IAM role can be assigned permissions to the database instance and can be attached to the EC2 instance. The instance will then obtain temporary security credentials from AWS STS which is much more secure.

CORRECT: "Create an IAM role with permission to access the database. Attach this IAM role to the EC2 instance" is the correct answer.

INCORRECT: "Move the configuration file to an Amazon S3 bucket. Create an IAM role with permission to the bucket and attach it to the EC2 instance" is incorrect. This just relocates the file; the contents are still unsecured and must be loaded by the application and passed to RDS. This is an insecure process.

INCORRECT: "Attach an additional volume to the EC2 instance with encryption enabled. Move the configuration file to the encrypted volume" is incorrect. This will only encrypt the file at rest, it still must be read, and the contents passed to RDS which is insecure.

INCORRECT: "Install an Amazon-trusted root certificate on the application instance and use SSL/TLS encrypted connections to the database" is incorrect. The file is still unsecured on the EBS volume so encrypting the credentials in an encrypted channel between the EC2 instance and RDS does not solve all security issues.



QUESTION 6 OF 65

6. QUESTION

A company requires that all AWS IAM user accounts have specific complexity requirements and minimum password length.

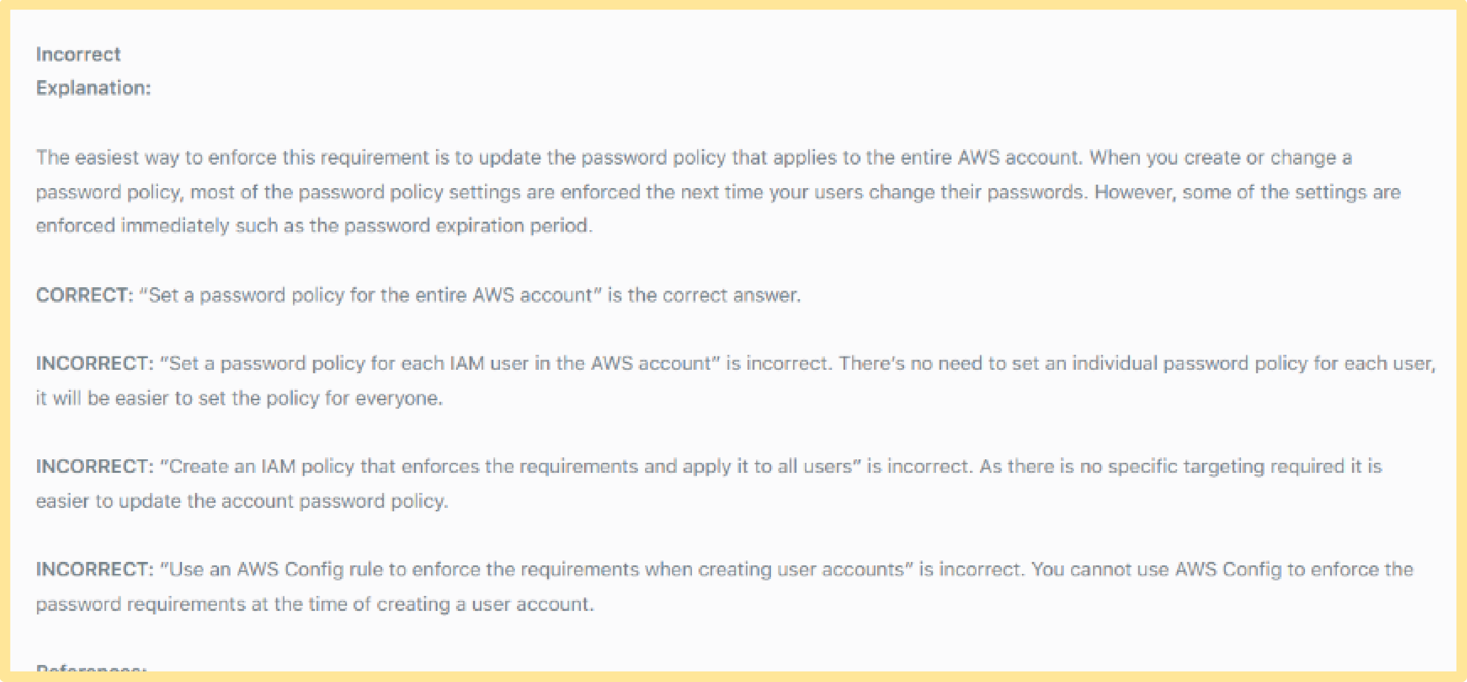
How should a Solutions Architect accomplish this?

○ Set a password policy for each IAM user in the AWS account.

○ Use an AWS Config rule to enforce the requirements when creating user accounts.

○ Set a password policy for the entire AWS account.

● Create an IAM policy that enforces the requirements and apply it to all users.



Incorrect
Explanation:

The easiest way to enforce this requirement is to update the password policy that applies to the entire AWS account. When you create or change a password policy, most of the password policy settings are enforced the next time your users change their passwords. However, some of the settings are enforced immediately such as the password expiration period.

CORRECT: "Set a password policy for the entire AWS account" is the correct answer.

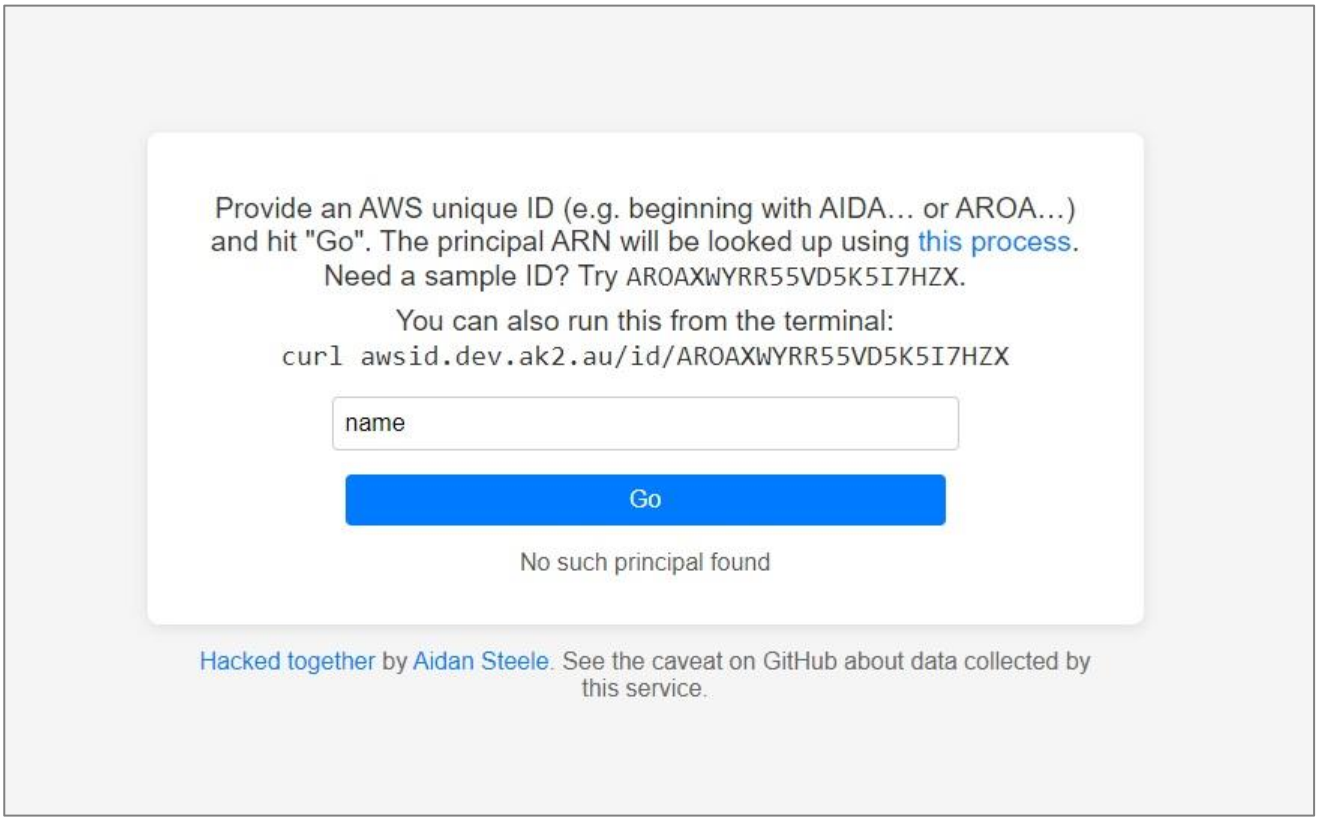
INCORRECT: "Set a password policy for each IAM user in the AWS account" is incorrect. There's no need to set an individual password policy for each user, it will be easier to set the policy for everyone.

INCORRECT: "Create an IAM policy that enforces the requirements and apply it to all users" is incorrect. As there is no specific targeting required it is easier to update the account password policy.

INCORRECT: "Use an AWS Config rule to enforce the requirements when creating user accounts" is incorrect. You cannot use AWS Config to enforce the password requirements at the time of creating a user account.

References:

I found this question to be quite difficult. The reason I selected the option involving an IAM policy is that the question *explicitly* stated that the thing we want to govern is IAM users. The explanation states "there is no specific targeting required". But this is not a reason for not using a service. Services frequently require a plethora of capabilities—my not needing a capability does not mean that I should not use the service.



Provide an AWS unique ID (e.g. beginning with AIDA… or AROA…) and hit "Go". The principal ARN will be looked up using this process. Need a sample ID? Try AROAXWYRR55VD5K5I7HZX.

You can also run this from the terminal:

```
curl awsid.dev.ak2.au/id/AROAXWYRR55VD5K5I7HZX
```

name

Go

No such principal found

Hacked together by Aidan Steele. See the caveat on GitHub about data collected by this service.

[Steele 2024]

TPN

101. ***Phenomenon1*** – the tendency of X to Y.
102. ***Phen2*** – the tendency of X to Y.
103. ***Phen3*** – the tendency of X to Y.
104. ***Phen4*** – the tendency of X to Y.
105. ***Phen5*** – the tendency of X to Y.
106. ***Phen6*** – the tendency of X to Y.
107. ***Phen7*** – the tendency of X to Y.
108. ***Phen8*** – the tendency of X to Y.
109. ***Phen9*** – the tendency of X to Y.
110. ***Phen10*** – the tendency of X to Y.

Glossary

Term1

Description of what term means here.

Term2

Description of what term means here.

Term3

Description of what term means here.

Bibliography

I. Official
II. Unofficial
III. Critical
IV. General

I. Official

[Weiss 2022]

Weiss, Becky (2022). AWS Identity and Access Management (IAM) deep dive. July 2022. Re:Inforce Conference (Boston, MA). Available at: <https://www.youtube.com/watch?v=YMj33ToS8cI&t=4s&ab_channel=AWSEvents>

[Amit 2021]

Amit (2021). How does IAM Evaluation logic work using an explicit Deny policy with multiple condition keys?. 15th Feb 2021. YouTube Channel: Amazon Web Services. Available at: <https://www.youtube.com/watch?v=8KLh1idp-1M&ab_channel=AmazonWebServices>.

[Wierer 2015]

Wierer, Jeff (2015). How to Become an IAM Policy Ninja in 60 minutes or less. 12th Oct 2015. Reinvent conference (SEC305) 2015. Available at: <https://www.youtube.com/watch?v=Du478i9O_mc&ab_channel=AmazonWebServices>.

[Amit 2021]

How does IAM evaluation logic work using an explicit Deny policy with multiple condition keys. YouTube Channel: Amazon Web Services. Available at: <https://www.youtube.com/watch?v=8KLh1idp-1M&ab_channel=AmazonWebServices>

[Mandadi 2015]

Mandadi, Srikanth (2015). AWS Security Token Service is Now Available in Every AWS Region. Feb 17th 2015. *AWS Security Blog*. Available at: <https://aws.amazon.com/blogs/security/aws-security-token-service-is-now-available-in-every-aws-region/>

II. Unofficial

[Be'ery 2024]

Be'ery, Tal (2024). Revealing the Inner Structure of AWS Session Tokens. *Medium*. Available at: <https://medium.com/@TalBeerySec/revealing-the-inner-structure-of-aws-session-tokens-a6c76469cba7>

[Morrison 2017]

Smith, David (year). AWS IAM Policies in a Nutshell. 23rd March 2017. Available at:

< https://start.jcolemorrison.com/aws-iam-policies-in-a-nutshell/>

[Kozliner 2020]

Kozliner, Evan (2020). AWS IAM Introduction. 23rd Oct 2020. Medium. Available at:
< https://towardsdatascience.com/aws-iam-introduction-20c1f017c43>

[Abdulla 2020]

Abdulla, Hamzah (2020). AWS IAM: Policies 101. 15th July 2020. *Medium*. Available at:
<https://hamzahabdulla1.medium.com/aws-iam-policies-101-426efaa849f4>.

[Kuenzli 2019]

Kuenzli, Stephen (2019). Why is AWS IAM so hard?. 25th Nov 2019. Available at:
<https://nodramadevops.com/2019/11/why-is-aws-iam-so-hard/>

[Forum 2019]

How can a policy be assigned to AWS resource? StackExchange. Question asked on 14th July 2019. Available at:
<https://stackoverflow.com/questions/57023953/how-can-a-policy-be-assigned-to-aws-resource>

[Steele 2024]

https://awsid.dev.ak2.au/

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

[Surname1]

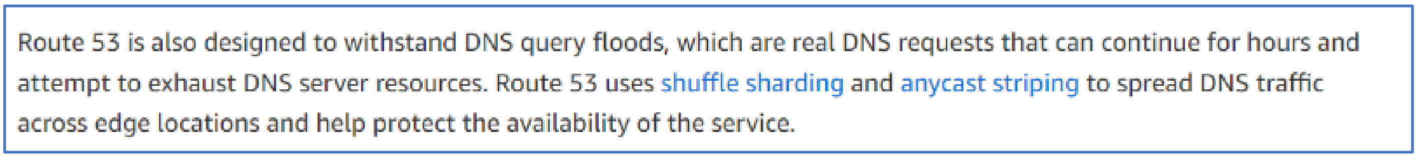
Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

IV. General

[Surname1]

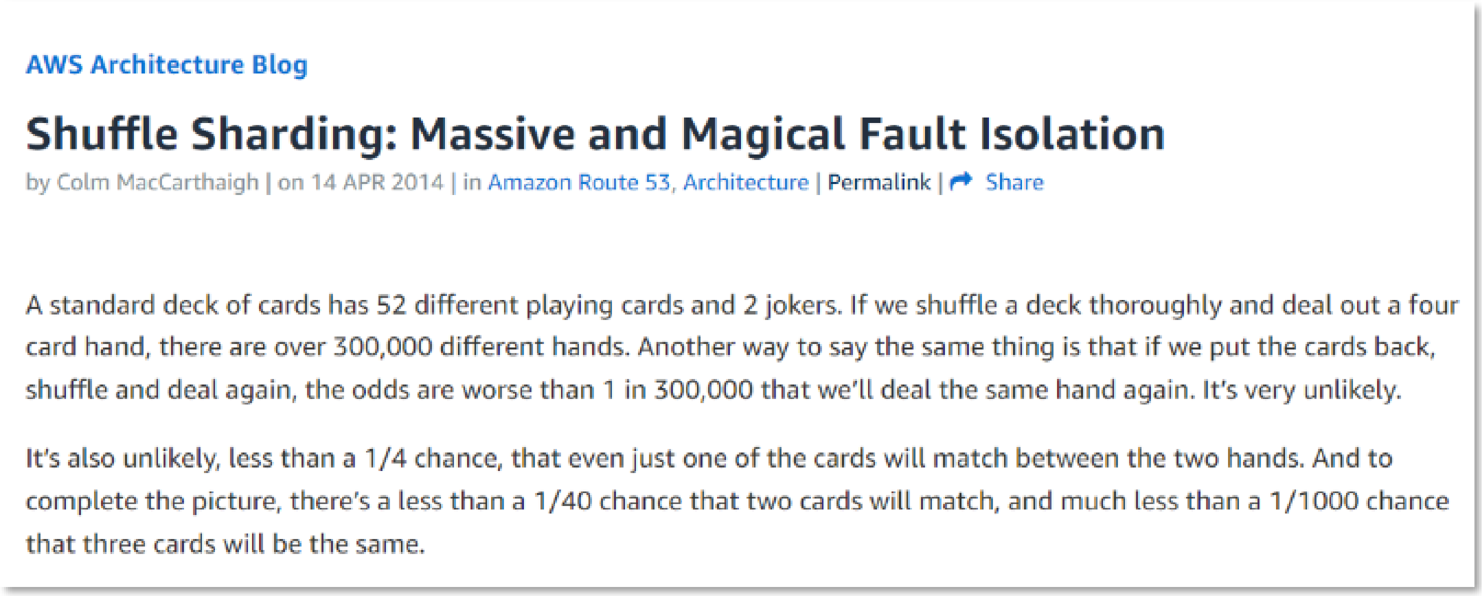
Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

Route 53



Route 53 is also designed to withstand DNS query floods, which are real DNS requests that can continue for hours and attempt to exhaust DNS server resources. Route 53 uses shuffle sharding and anycast striping to spread DNS traffic across edge locations and help protect the availability of the service.

Holly Willey in her March 2017 article on DDoS attacks, commenting on Route 53 capabilities



AWS Architecture Blog

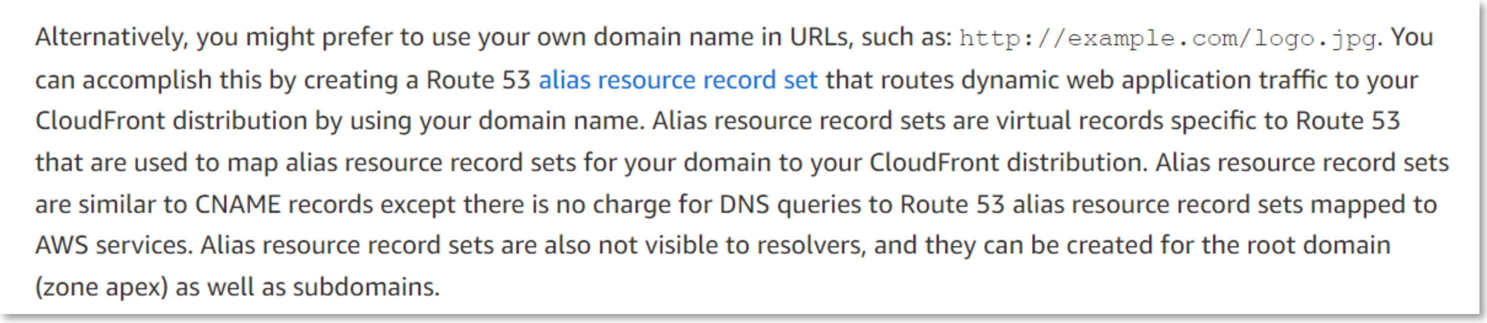
Shuffle Sharding: Massive and Magical Fault Isolation

by Colm MacCarthaigh | on 14 APR 2014 | in Amazon Route 53, Architecture | Permalink | Share

A standard deck of cards has 52 different playing cards and 2 jokers. If we shuffle a deck thoroughly and deal out a four card hand, there are over 300,000 different hands. Another way to say the same thing is that if we put the cards back, shuffle and deal again, the odds are worse than 1 in 300,000 that we'll deal the same hand again. It's very unlikely.

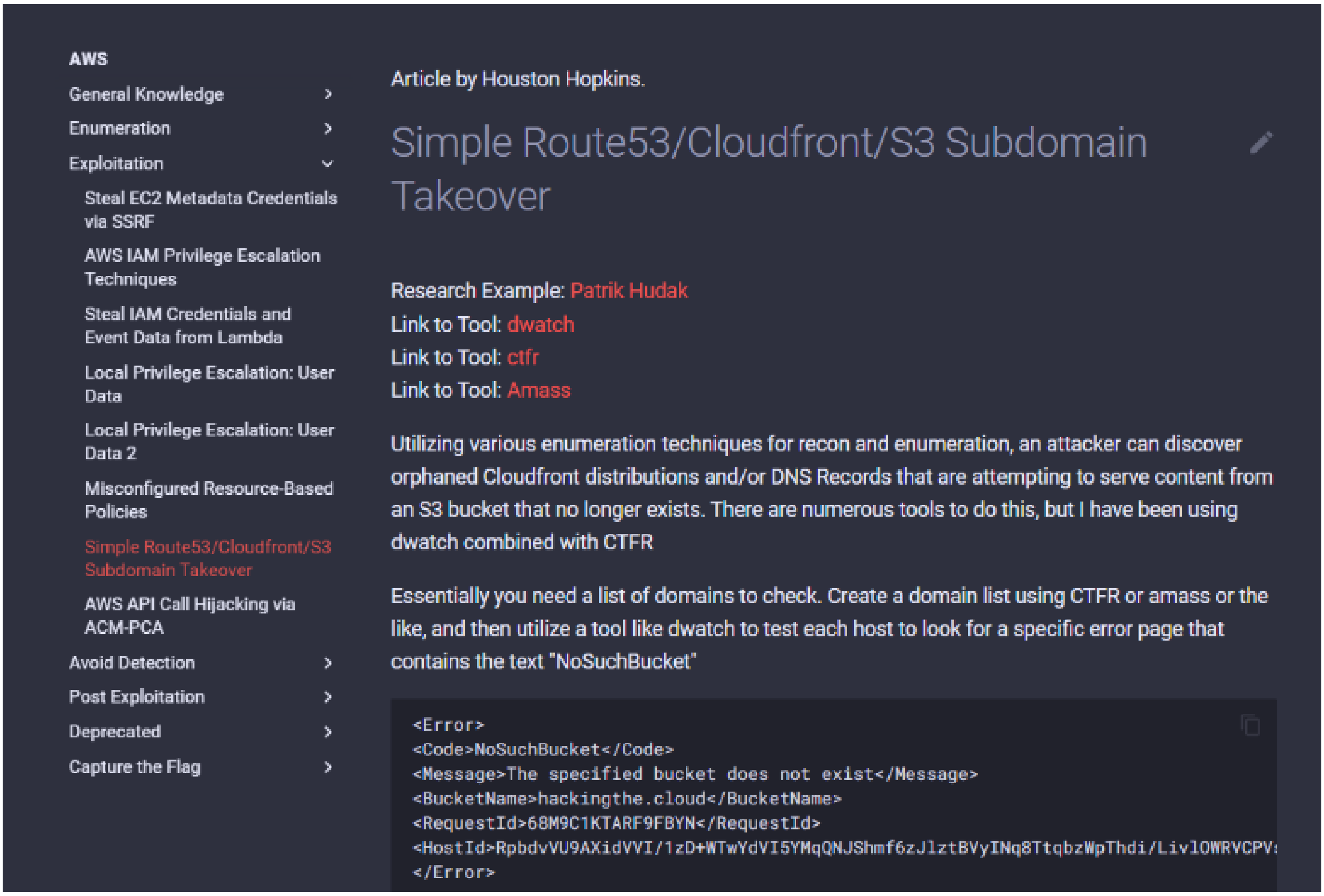
It's also unlikely, less than a 1/4 chance, that even just one of the cards will match between the two hands. And to complete the picture, there's a less than a 1/40 chance that two cards will match, and much less than a 1/1000 chance that three cards will be the same.

Colm MacCarthaigh explaining how shuffle sharding works



Alternatively, you might prefer to use your own domain name in URLs, such as: `http://example.com/logo.jpg`. You can accomplish this by creating a Route 53 alias resource record set that routes dynamic web application traffic to your CloudFront distribution by using your domain name. Alias resource record sets are virtual records specific to Route 53 that are used to map alias resource record sets for your domain to your CloudFront distribution. Alias resource record sets are similar to CNAME records except there is no charge for DNS queries to Route 53 alias resource record sets mapped to AWS services. Alias resource record sets are also not visible to resolvers, and they can be created for the root domain (zone apex) as well as subdomains.

Holly Willey (in March 2017) on alias records in Route 53


AWS
General Knowledge
Enumeration
Exploitation
Steal EC2 Metadata Credentials via SSRF
AWS IAM Privilege Escalation Techniques
Steal IAM Credentials and Event Data from Lambda
Local Privilege Escalation: User Data
Local Privilege Escalation: User Data 2
Misconfigured Resource-Based Policies
Simple Route53/Cloudfront/S3 Subdomain Takeover
AWS API Call Hijacking via ACM-PCA
Avoid Detection
Post Exploitation
Deprecated
Capture the Flag

Article by Houston Hopkins.

Simple Route53/Cloudfront/S3 Subdomain Takeover

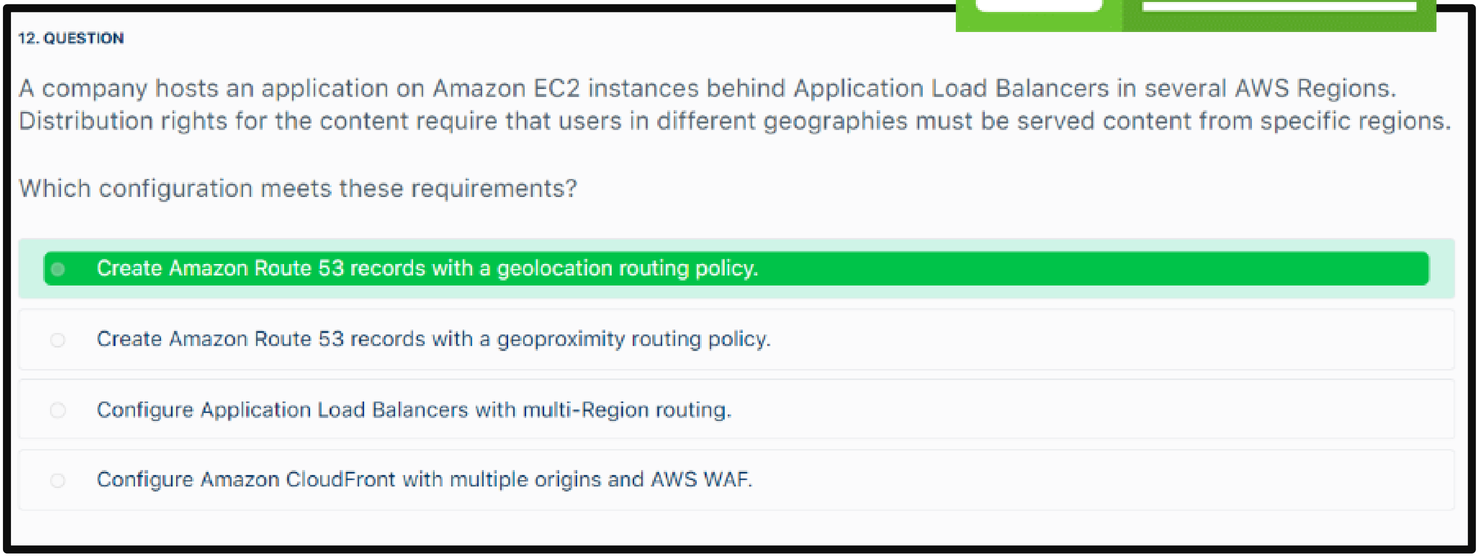
Research Example: Patrik Hudak
Link to Tool: dwatch
Link to Tool: ctfr
Link to Tool: Amass

Utilizing various enumeration techniques for recon and enumeration, an attacker can discover orphaned Cloudfront distributions and/or DNS Records that are attempting to serve content from an S3 bucket that no longer exists. There are numerous tools to do this, but I have been using dwatch combined with CTFR

Essentially you need a list of domains to check. Create a domain list using CTFR or amass or the like, and then utilize a tool like dwatch to test each host to look for a specific error page that contains the text "NoSuchBucket"

```
<Error>
<Code>NoSuchBucket</Code>
<Message>The specified bucket does not exist</Message>
<BucketName>hackingthe.cloud</BucketName>
<RequestId>68M9C1KTARF9FBYN</RequestId>
<HostId>RpbdvVU9AXidVVI/1zD+WTwYdVI5YMqQNJShmf6zJlztBVyINq8TtqbzWpThdi/Livl0WRVCPV
</Error>
```

https://hackingthe.cloud/aws/exploitation/orphaned_%20cloudfront_or_dns_takeover_via_s3/?ck_subscriber_id=1560524742



12. QUESTION

A company hosts an application on Amazon EC2 instances behind Application Load Balancers in several AWS Regions. Distribution rights for the content require that users in different geographies must be served content from specific regions.

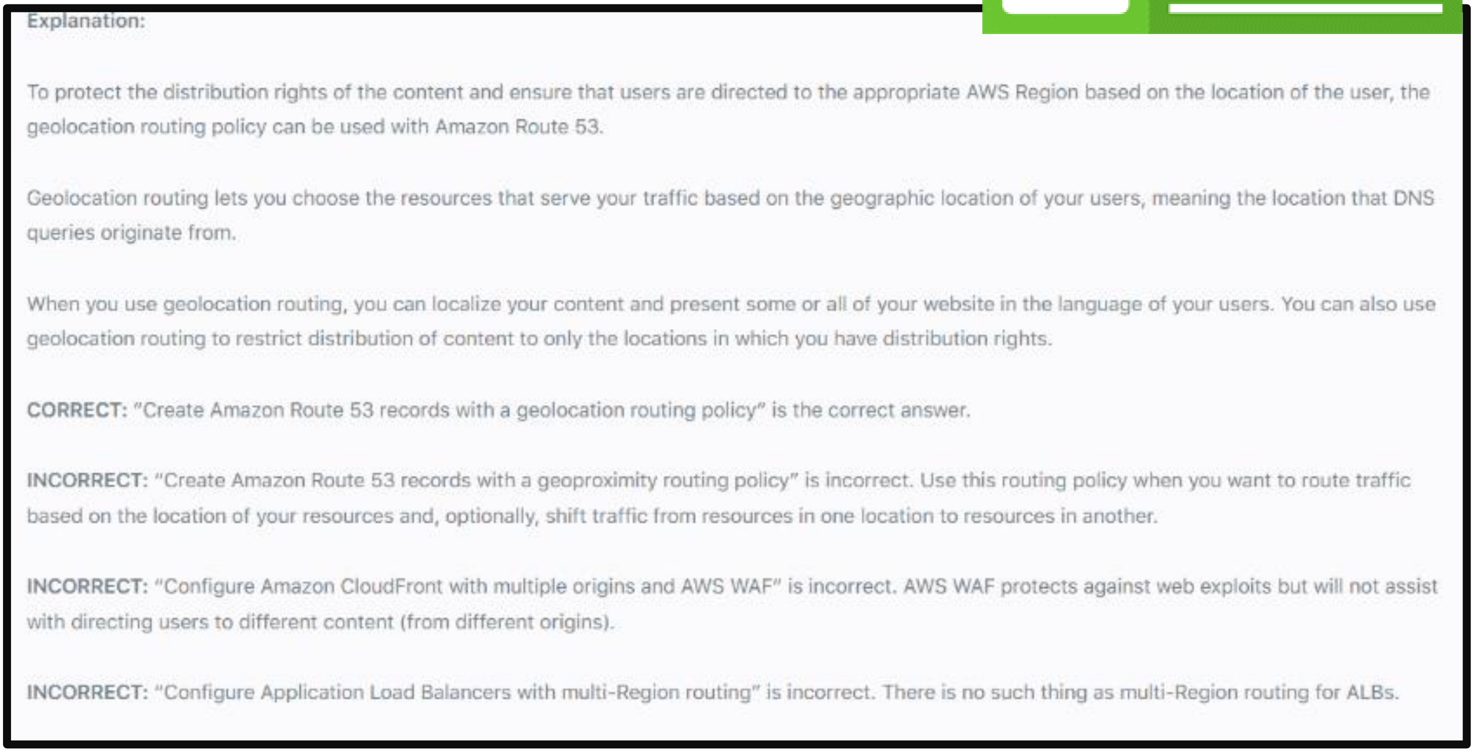
Which configuration meets these requirements?

● Create Amazon Route 53 records with a geolocation routing policy.

○ Create Amazon Route 53 records with a geoproximity routing policy.

○ Configure Application Load Balancers with multi-Region routing.

○ Configure Amazon CloudFront with multiple origins and AWS WAF.



Explanation:

To protect the distribution rights of the content and ensure that users are directed to the appropriate AWS Region based on the location of the user, the geolocation routing policy can be used with Amazon Route 53.

Geolocation routing lets you choose the resources that serve your traffic based on the geographic location of your users, meaning the location that DNS queries originate from.

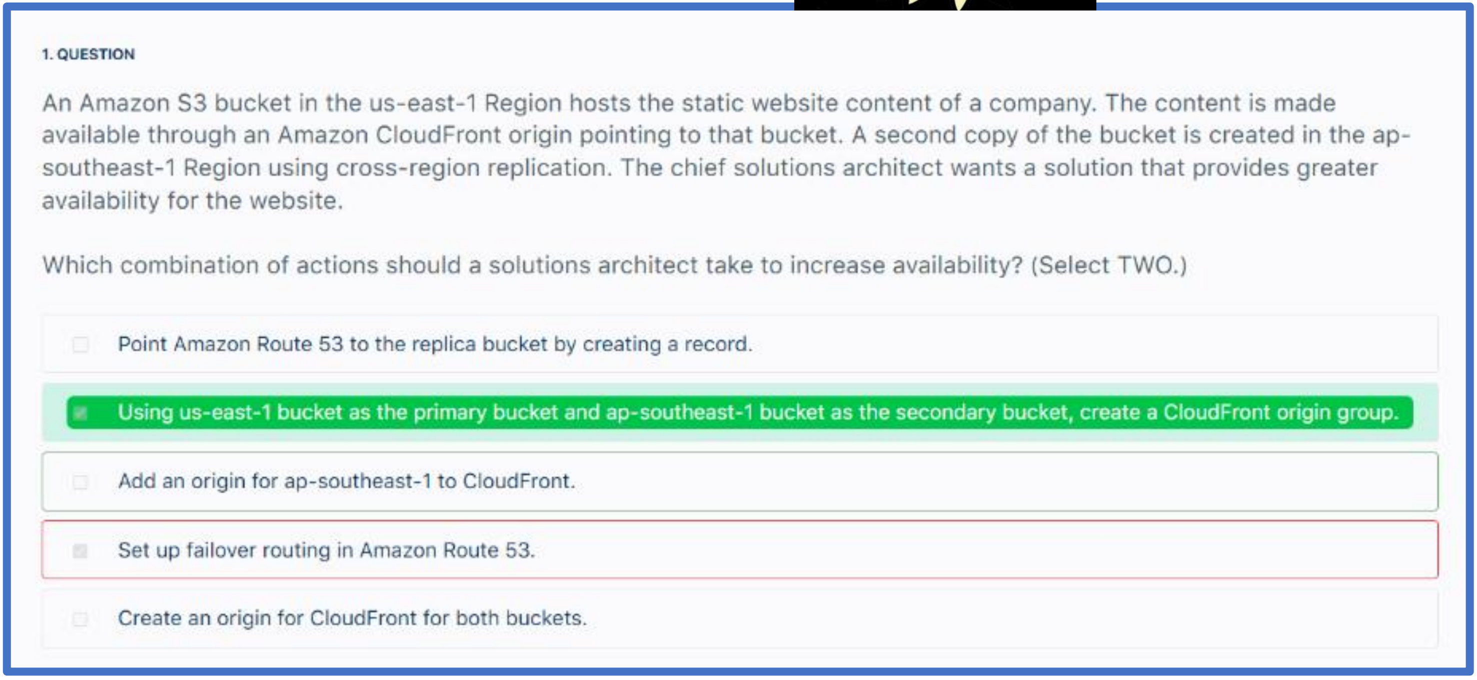
When you use geolocation routing, you can localize your content and present some or all of your website in the language of your users. You can also use geolocation routing to restrict distribution of content to only the locations in which you have distribution rights.

CORRECT: "Create Amazon Route 53 records with a geolocation routing policy" is the correct answer.

INCORRECT: "Create Amazon Route 53 records with a geoproximity routing policy" is incorrect. Use this routing policy when you want to route traffic based on the location of your resources and, optionally, shift traffic from resources in one location to resources in another.

INCORRECT: "Configure Amazon CloudFront with multiple origins and AWS WAF" is incorrect. AWS WAF protects against web exploits but will not assist with directing users to different content (from different origins).

INCORRECT: "Configure Application Load Balancers with multi-Region routing" is incorrect. There is no such thing as multi-Region routing for ALBs.

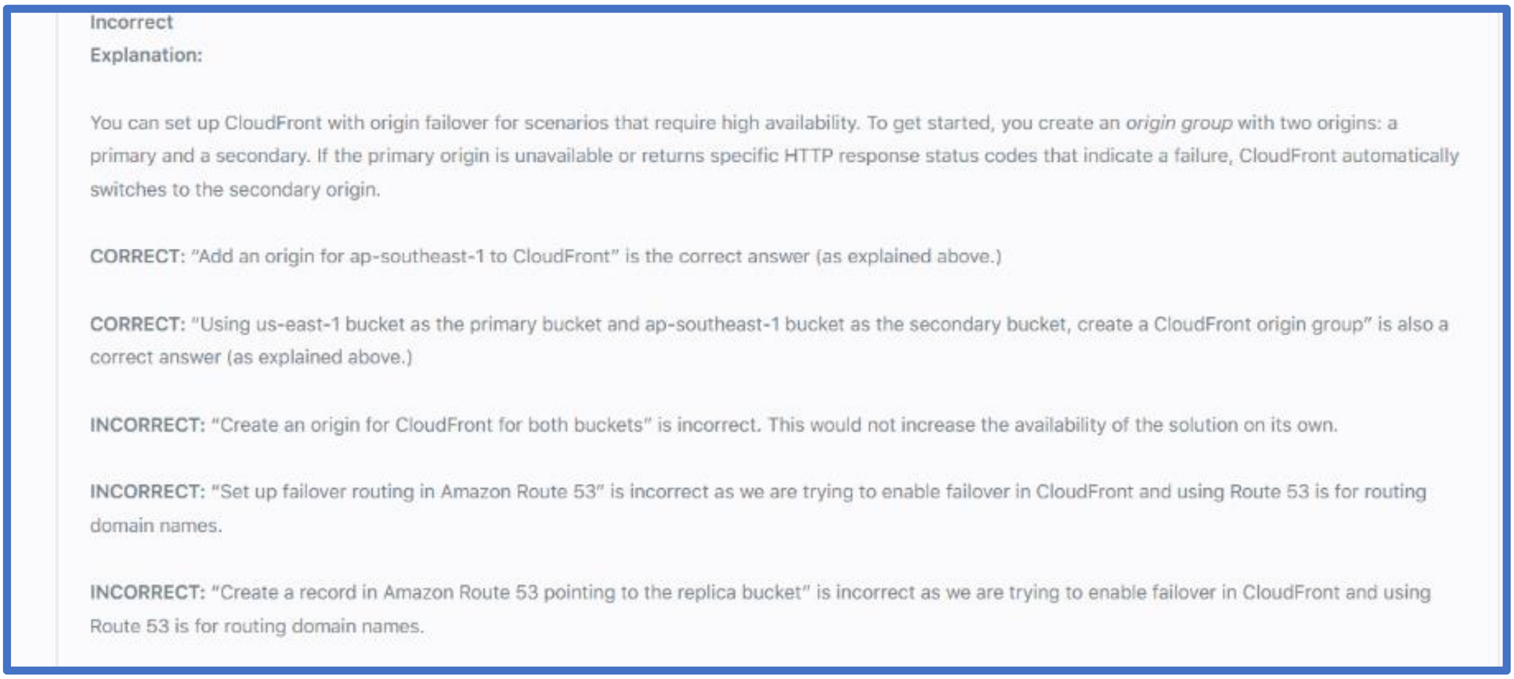


1. QUESTION

An Amazon S3 bucket in the us-east-1 Region hosts the static website content of a company. The content is made available through an Amazon CloudFront origin pointing to that bucket. A second copy of the bucket is created in the ap-southeast-1 Region using cross-region replication. The chief solutions architect wants a solution that provides greater availability for the website.

Which combination of actions should a solutions architect take to increase availability? (Select TWO.)

- □ Point Amazon Route 53 to the replica bucket by creating a record.
- ☑ Using us-east-1 bucket as the primary bucket and ap-southeast-1 bucket as the secondary bucket, create a CloudFront origin group.
- □ Add an origin for ap-southeast-1 to CloudFront.
- ☑ Set up failover routing in Amazon Route 53.
- □ Create an origin for CloudFront for both buckets.



Incorrect

Explanation:

You can set up CloudFront with origin failover for scenarios that require high availability. To get started, you create an *origin group* with two origins: a primary and a secondary. If the primary origin is unavailable or returns specific HTTP response status codes that indicate a failure, CloudFront automatically switches to the secondary origin.

CORRECT: "Add an origin for ap-southeast-1 to CloudFront" is the correct answer (as explained above.)

CORRECT: "Using us-east-1 bucket as the primary bucket and ap-southeast-1 bucket as the secondary bucket, create a CloudFront origin group" is also a correct answer (as explained above.)

INCORRECT: "Create an origin for CloudFront for both buckets" is incorrect. This would not increase the availability of the solution on its own.

INCORRECT: "Set up failover routing in Amazon Route 53" is incorrect as we are trying to enable failover in CloudFront and using Route 53 is for routing domain names.

INCORRECT: "Create a record in Amazon Route 53 pointing to the replica bucket" is incorrect as we are trying to enable failover in CloudFront and using Route 53 is for routing domain names.

The above question really annoyed me. I selected an option involving using Route 53. This DNS service has routing policies. One such policy is called the failover policy—this undoubtedly helps to increase availability. So, it is simply incorrect to say that using a failover policy will not help with our availability.

The explanation simply says "we are trying to enable failover in CloudFront". But this is begging the question. We didn't (as a student answering the question) know that there was some strange prescription to stay with CloudFront exclusively. Why not use Route 53?

Bibliography

I. Official

[Haken 2024]

Haken, Michael and John Formento and Saurabh Kumar (2024). Creating an organizational multi-Region failover strategy. Available at: <https://aws.amazon.com/blogs/architecture/creating-an-organizational-multi-region-failover-strategy/?ck_subscriber_id=1560524742>

II. Unofficial

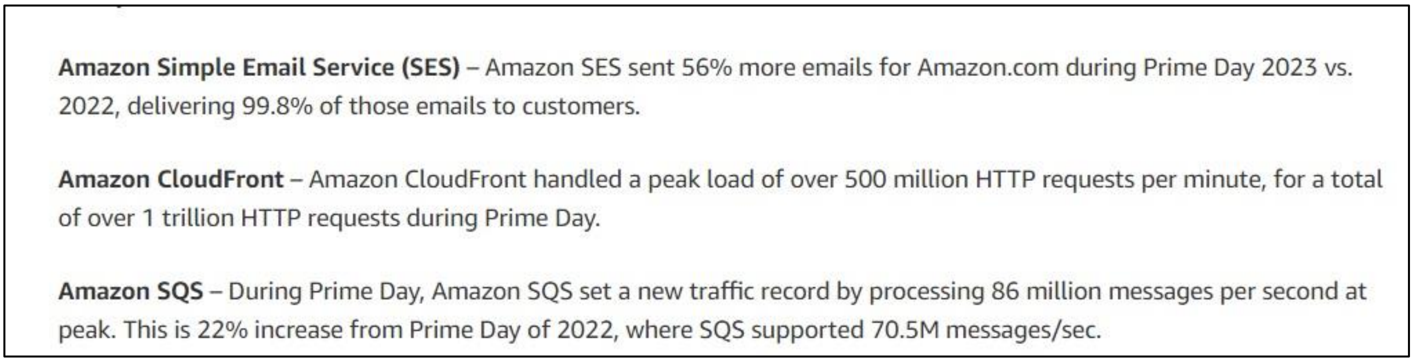
https://git.jon-e.net/jonny/DNSocial?ck_subscriber_id=1560524742

Elastic Beanstalk

Elastic Beanstalk

Simple Email Service

Chris Wheeler explaining SES to Jeff Barr in 2012

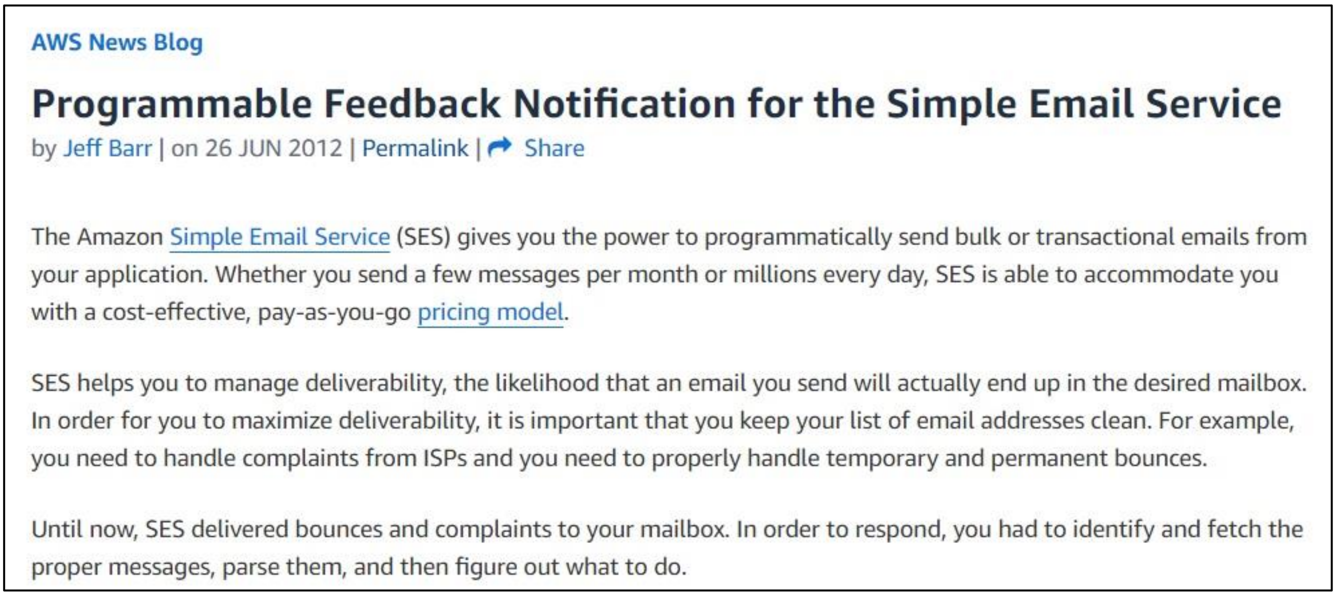


Amazon Simple Email Service (SES) – Amazon SES sent 56% more emails for Amazon.com during Prime Day 2023 vs. 2022, delivering 99.8% of those emails to customers.

Amazon CloudFront – Amazon CloudFront handled a peak load of over 500 million HTTP requests per minute, for a total of over 1 trillion HTTP requests during Prime Day.

Amazon SQS – During Prime Day, Amazon SQS set a new traffic record by processing 86 million messages per second at peak. This is 22% increase from Prime Day of 2022, where SQS supported 70.5M messages/sec.

From an article by Jeff [Barr 2023] about the vast AWS resources involved in Prime Day 2023



AWS News Blog

Programmable Feedback Notification for the Simple Email Service

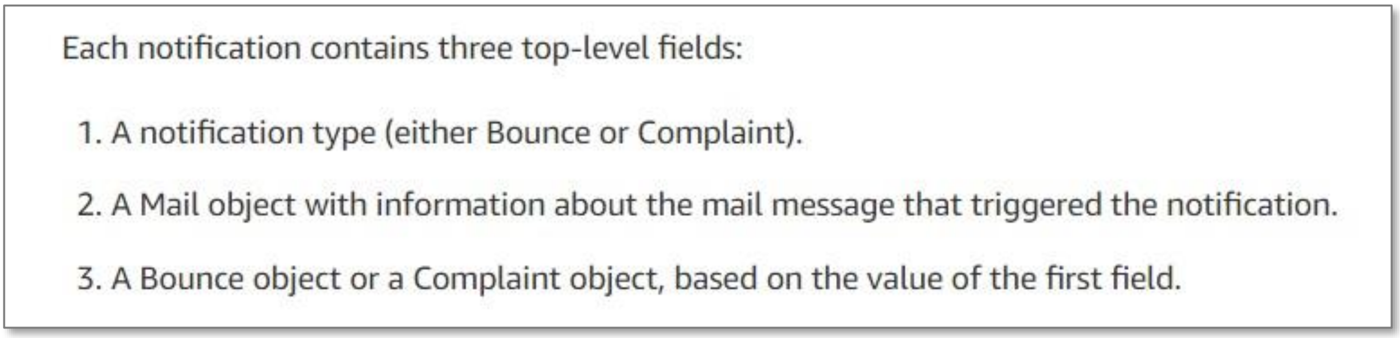
by Jeff Barr | on 26 JUN 2012 | Permalink | ➦ Share

The Amazon Simple Email Service (SES) gives you the power to programmatically send bulk or transactional emails from your application. Whether you send a few messages per month or millions every day, SES is able to accommodate you with a cost-effective, pay-as-you-go pricing model.

SES helps you to manage deliverability, the likelihood that an email you send will actually end up in the desired mailbox. In order for you to maximize deliverability, it is important that you keep your list of email addresses clean. For example, you need to handle complaints from ISPs and you need to properly handle temporary and permanent bounces.

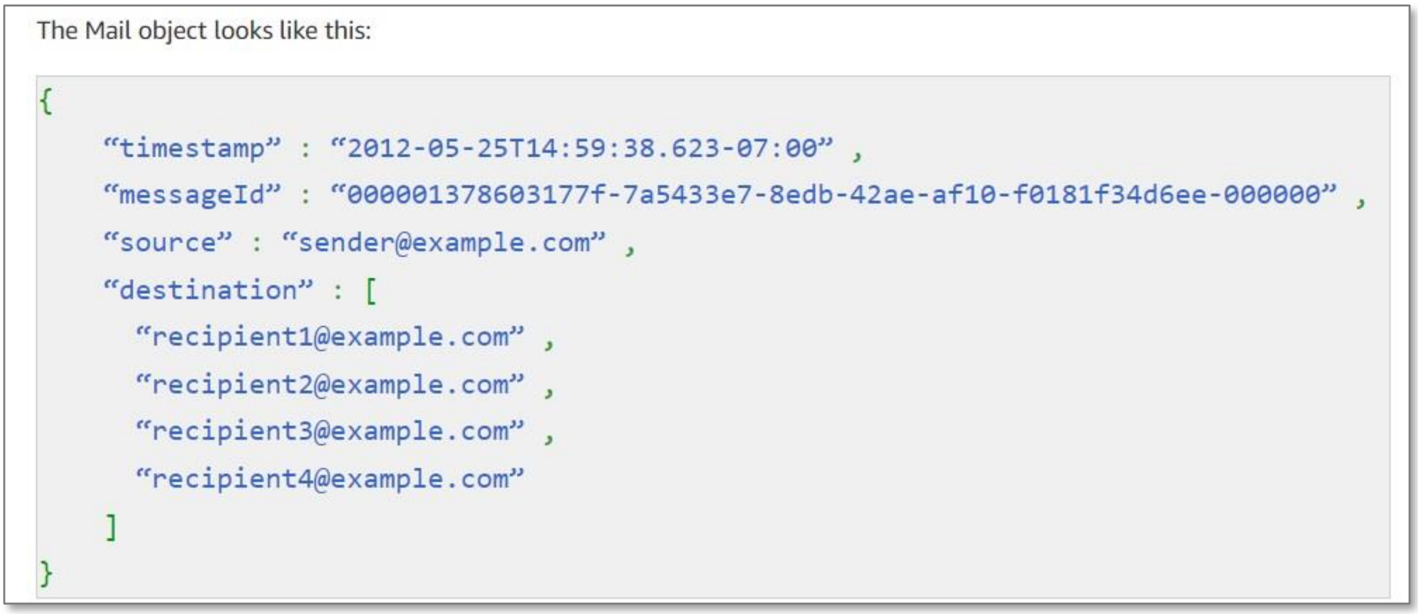
Until now, SES delivered bounces and complaints to your mailbox. In order to respond, you had to identify and fetch the proper messages, parse them, and then figure out what to do.

Article [Barr 2012] explaining that SES can help deal with bounces. Sometimes, you will get a complaint from an ISP (Internet Service Provider)



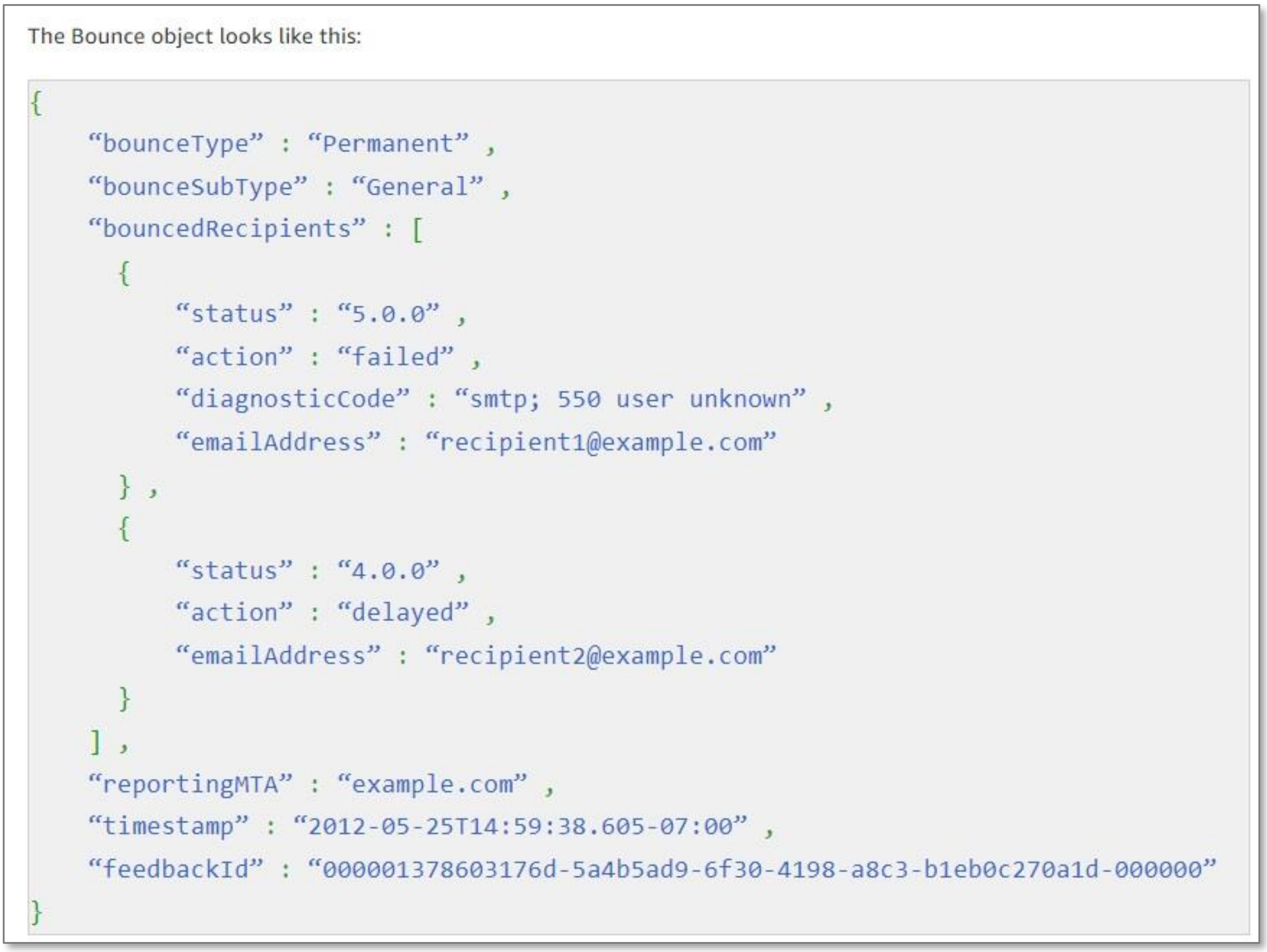
Each notification contains three top-level fields:

1. A notification type (either Bounce or Complaint).
2. A Mail object with information about the mail message that triggered the notification.
3. A Bounce object or a Complaint object, based on the value of the first field.



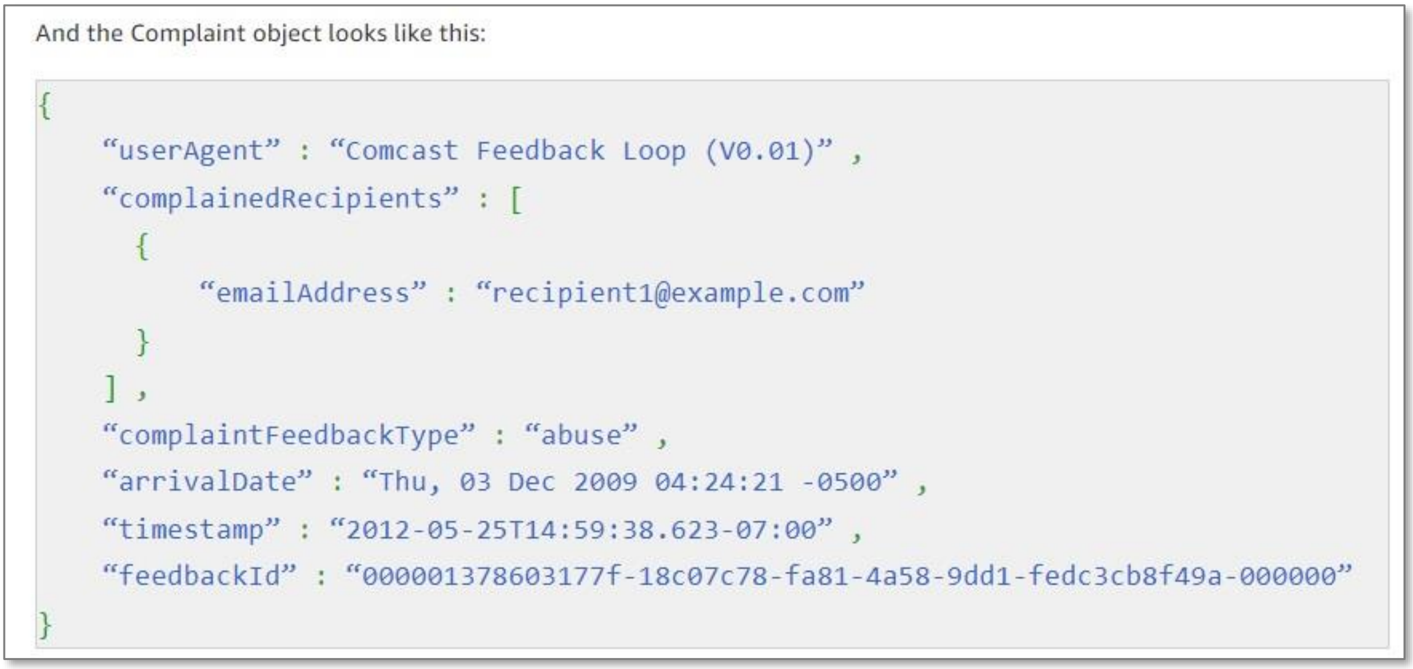
The Mail object looks like this:

```
{
    “timestamp” : “2012-05-25T14:59:38.623-07:00” ,
    “messageId” : “000001378603177f-7a5433e7-8edb-42ae-af10-f0181f34d6ee-000000” ,
    “source” : “sender@example.com” ,
    “destination” : [
      “recipient1@example.com” ,
      “recipient2@example.com” ,
      “recipient3@example.com” ,
      “recipient4@example.com”
    ]
}
```



The Bounce object looks like this:

```
{
    “bounceType” : “Permanent” ,
    “bounceSubType” : “General” ,
    “bouncedRecipients” : [
      {
          “status” : “5.0.0” ,
          “action” : “failed” ,
          “diagnosticCode” : “smtp; 550 user unknown” ,
          “emailAddress” : “recipient1@example.com”
      } ,
      {
          “status” : “4.0.0” ,
          “action” : “delayed” ,
          “emailAddress” : “recipient2@example.com”
      }
    ] ,
    “reportingMTA” : “example.com” ,
    “timestamp” : “2012-05-25T14:59:38.605-07:00” ,
    “feedbackId” : “000001378603176d-5a4b5ad9-6f30-4198-a8c3-b1eb0c270a1d-000000”
}
```



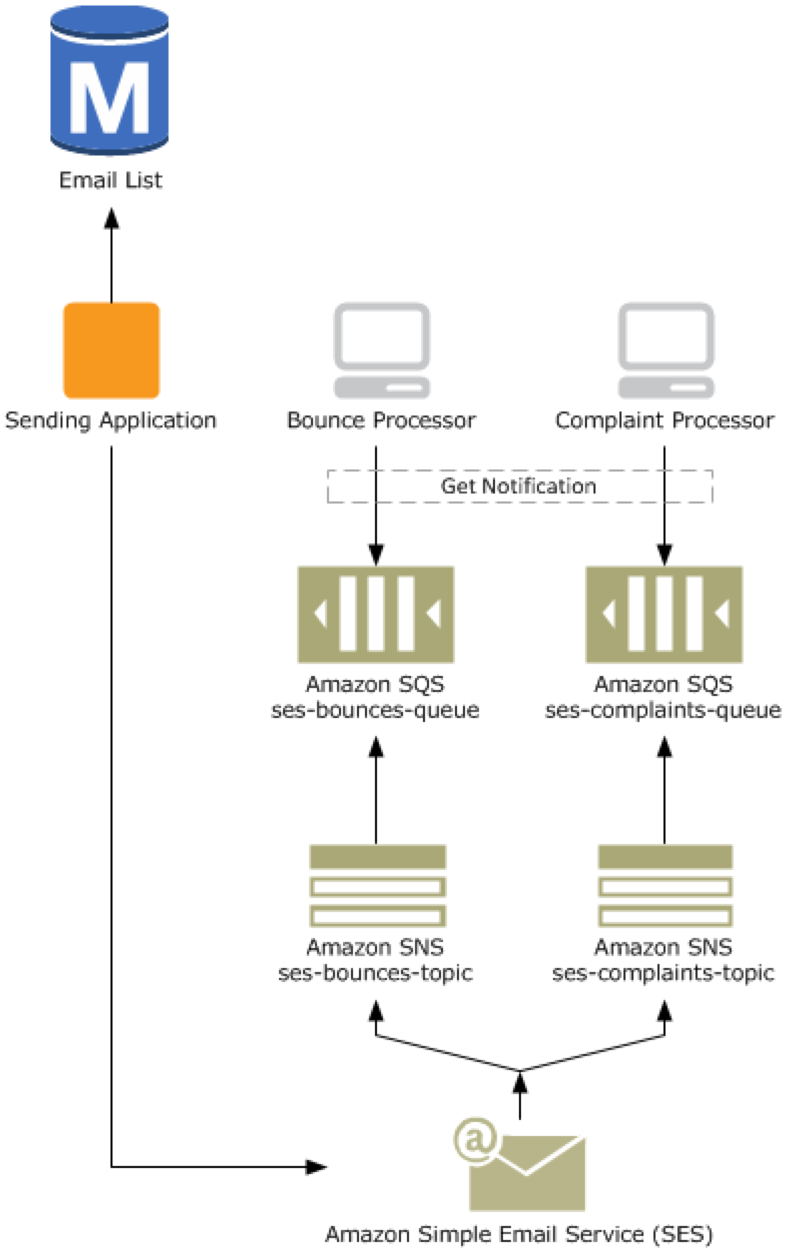
And the Complaint object looks like this:

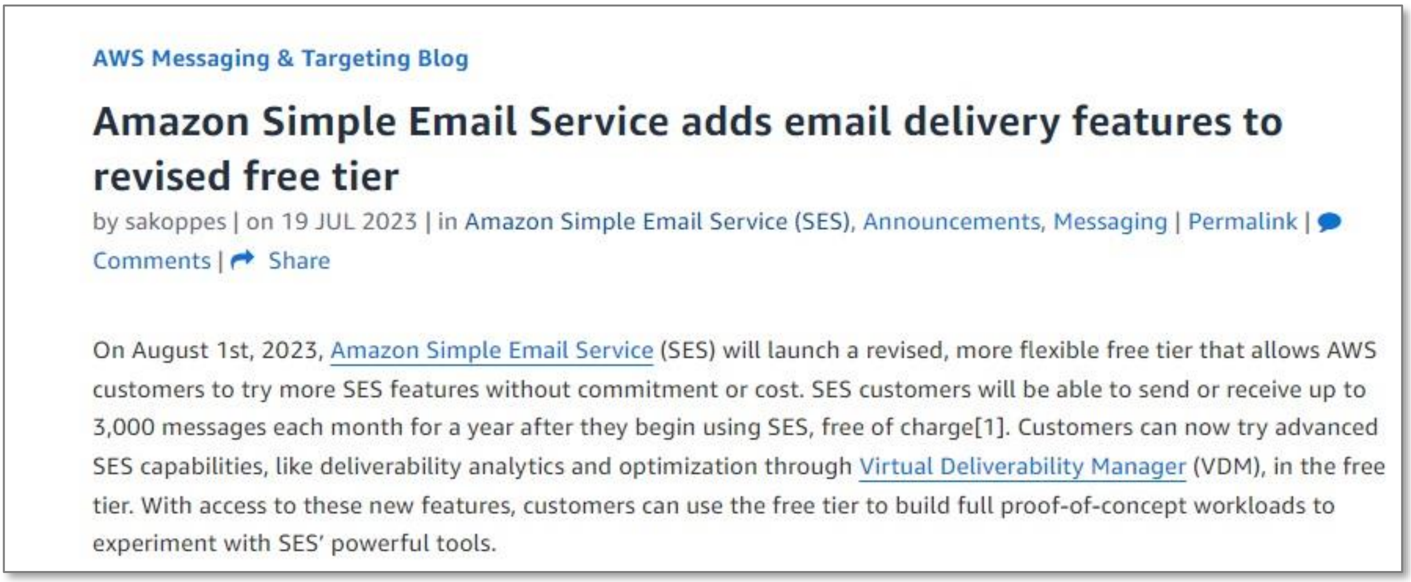
```
{
    “userAgent” : “Comcast Feedback Loop (V0.01)” ,
    “complainedRecipients” : [
      {
          “emailAddress” : “recipient1@example.com”
      }
    ] ,
    “complaintFeedbackType” : “abuse” ,
    “arrivalDate” : “Thu, 03 Dec 2009 04:24:21 -0500” ,
    “timestamp” : “2012-05-25T14:59:38.623-07:00” ,
    “feedbackId” : “000001378603177f-18c07c78-fa81-4a58-9dd1-fedc3cb8f49a-000000”
}
```

Following this article from Barr, we got a very interesting article from Rohan Deshpande. It explains how to deal with these bounces and complaints. We can automate the removal of an address from our sender's list. The code bit uses might look intimidating at first. Give it a minute to become clear. Deshpande chooses to use the .NET framework.

Interestingly, the architecture of the solution involves combing SQS queues and SNS topics. The topic receives the bounces and complaints. (In fact, there is one topic for each.) Recall that SNS is a "push" model. It dumps messages into the world when it needs to. It is not a "pull" model – it is not a reservoir of messages waiting for you to draw upon them.

For this reason, we need SQS. A queue is almost like a reservoir of messages. The messages stay there, stagnant and waiting. When *we* are ready, we can deal with the messages. Deshpande explains:


Email List
Sending Application
Bounce Processor
Complaint Processor
Get Notification
Amazon SQS
ses-bounces-queue
Amazon SQS
ses-complaints-queue
Amazon SNS
ses-bounces-topic
Amazon SNS
ses-complaints-topic
Amazon Simple Email Service (SES)

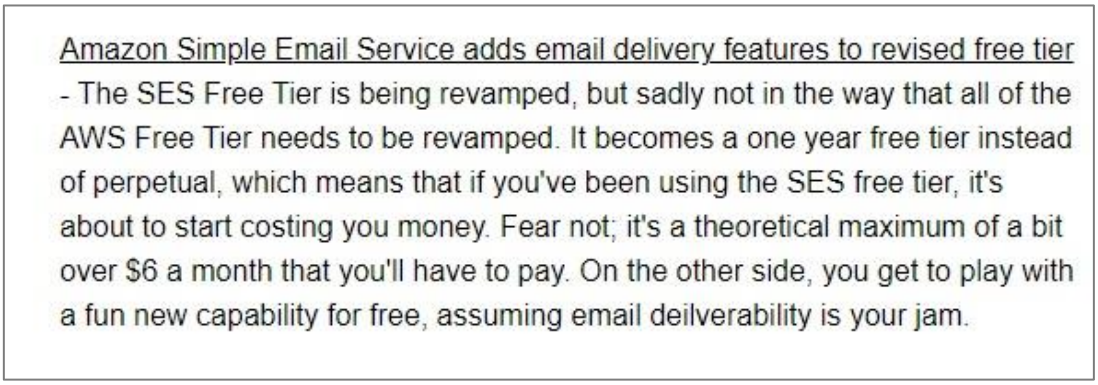


AWS Messaging & Targeting Blog

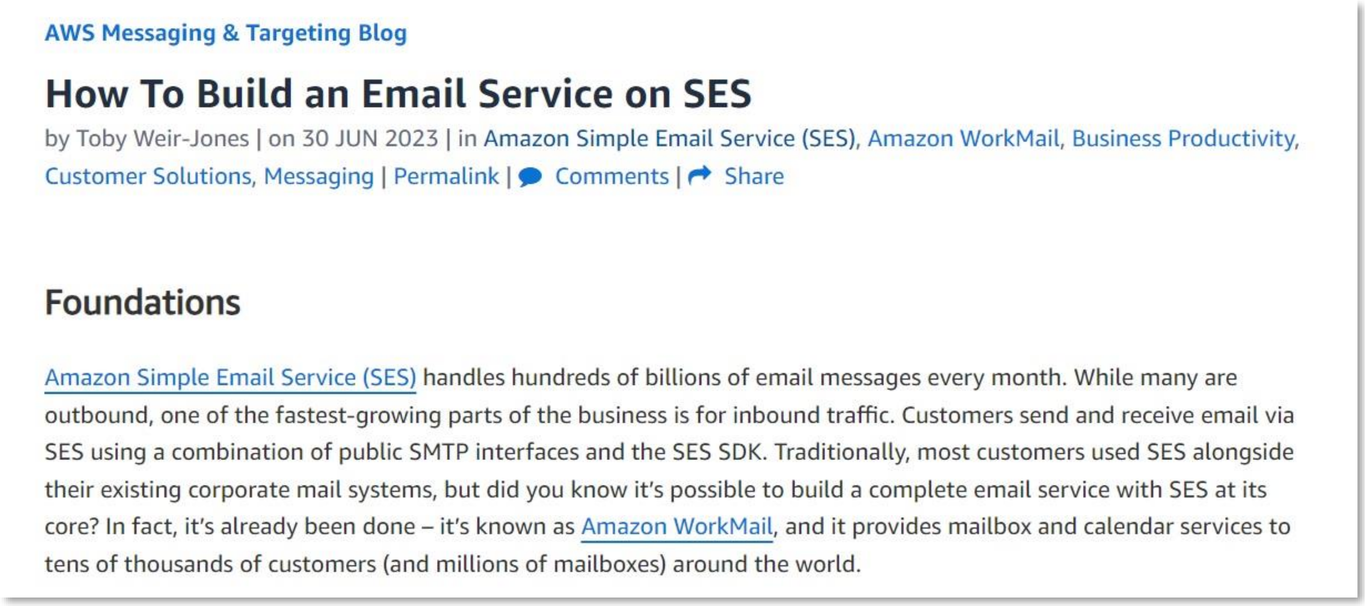
Amazon Simple Email Service adds email delivery features to revised free tier

by sakoppes | on 19 JUL 2023 | in Amazon Simple Email Service (SES), Announcements, Messaging | Permalink | 💬 Comments | ➦ Share

On August 1st, 2023, Amazon Simple Email Service (SES) will launch a revised, more flexible free tier that allows AWS customers to try more SES features without commitment or cost. SES customers will be able to send or receive up to 3,000 messages each month for a year after they begin using SES, free of charge[1]. Customers can now try advanced SES capabilities, like deliverability analytics and optimization through Virtual Deliverability Manager (VDM), in the free tier. With access to these new features, customers can use the free tier to build full proof-of-concept workloads to experiment with SES' powerful tools.



Amazon Simple Email Service adds email delivery features to revised free tier - The SES Free Tier is being revamped, but sadly not in the way that all of the AWS Free Tier needs to be revamped. It becomes a one year free tier instead of perpetual, which means that if you've been using the SES free tier, it's about to start costing you money. Fear not; it's a theoretical maximum of a bit over $6 a month that you'll have to pay. On the other side, you get to play with a fun new capability for free, assuming email deilverability is your jam.



AWS Messaging & Targeting Blog

How To Build an Email Service on SES

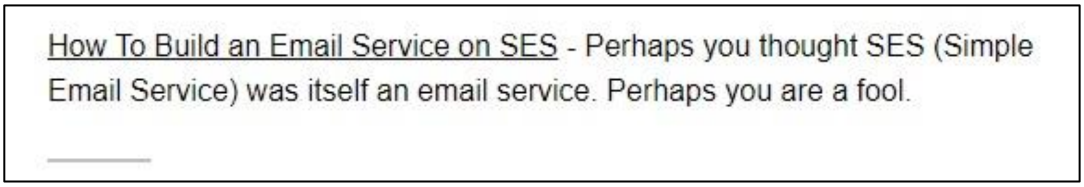
by Toby Weir-Jones | on 30 JUN 2023 | in Amazon Simple Email Service (SES), Amazon WorkMail, Business Productivity, Customer Solutions, Messaging | Permalink | Comments | Share

Foundations

Amazon Simple Email Service (SES) handles hundreds of billions of email messages every month. While many are outbound, one of the fastest-growing parts of the business is for inbound traffic. Customers send and receive email via SES using a combination of public SMTP interfaces and the SES SDK. Traditionally, most customers used SES alongside their existing corporate mail systems, but did you know it's possible to build a complete email service with SES at its core? In fact, it's already been done – it's known as Amazon WorkMail, and it provides mailbox and calendar services to tens of thousands of customers (and millions of mailboxes) around the world.

What are we to make of the above blogpost? This blogpost is helpful because it describes the relationship which SES has to Amazon WorkMail. SES is a sort of intelligent software that maintains everything needed for delivering email. We need to maintain a list of reputable IPs, we need to have policies for attempting to re-send failed emails, we might want to add banners to emails which come into an organization from outside etc. WorkMail, in contrast, "provides mailbox and calendar services". WorkMail allows a complete solutions, providing a mailbox. SES doesn't necessarily provide a mailbox – or rather, SES can be used with a range of different mailboxes. Indeed, Weir-Jones explains that:

> Traditionally, most customers used SES alongside their existing corporate mail systems, but did you know it's possible to build a complete email service with SES at its core? In fact, it's already been done – it's known as Amazon WorkMail
>
> [Weir-Jones 2023]



How To Build an Email Service on SES - Perhaps you thought SES (Simple Email Service) was itself an email service. Perhaps you are a fool.

Email from Corey Quinn on 10th July 2023

SES manages the protocols that are involved in email. It is not WorkMail that directly receives information transmitted in the Simple Mail Transfer Protocol (SMTP). Rather, it is SES which maintains a list of destinations, which information can be sent to via SMTP:

> When a message arrives via SMTP, SES first interrogates a back-end directory to confirm that the message is destined for an SES customer.
>
> [Weir-Jones 2023]

SES is then response for conducting some kind of query behind the scenes. SES needs to determine where to forward the email to, and so will consider the domain with which the email address is associated:

> it looks up how the customer's domain is configured, or if it is a WorkMail customer domain.
>
> From there the message passes through the SES message scanner, where its content is evaluated for spam or malware, and a scoring indicator is added to the message headers.
>
> That score may result in the message being dropped altogether, or it may result in the message ultimately being delivered to a Junk Mail folder in a WorkMail mailbox.
>
> [Weir-Jones 2023]

So, we can see that SES assesses the email according to several criteria. If these criteria are met, then the email may end up being send to a WorkMail mailbox.

What actually is a mailbox, in the context of Amazon WorkMail? We are told that:

> In practice a mailbox is a structured object format also within S3, but without raw S3 access because the storage is managed as a system resource within WorkMail instead of being owned by an end customer.
>
> [Weir-Jones 2023]

The main contribution of this article is to emphasise that SES can be used along. SES can be used without WorkMail, and this can allow creative solutions. AWS Lambda can be used:

> There are a number of options which may take place while that message is in transit, however, and the SES framework supports those with its flexible routing options.
>
> For example, a very popular choice is for customers to trigger a transport rule powered by AWS Lambda for inbound and/or outbound messages.
>
> [Weir-Jones 2023]

We are told that the "ingredients for success" in an email system are maintaining a good reputation so that you can send mail out and ensuring you only receive mail from those with decent reputations. We are told that the benefit of SES is that it can handle huge spikes in growth. We are told that:

> The classic on-premise enterprise use case, however, still runs the risk of overwhelming the capacity of the (single) mail server, either due to a malicious action by a sender or a huge increase in usage over a very short period of time.

> SES absorbs those spikes automatically and has orders of magnitude more capacity than any typical on-premise deployment
>
> [Weir-Jones 2023]

Crucially, the point made by Weir-Jones is that the measures built into the email protocols and standards are not sufficient to deal with these issues. Email protocols look like they can deal with the issue of being bombarded by emails – they allow for re-tries, for example. But this is not enough. You cannot rely on one server. You should use SES.

Glossary

MTA – Mail Transfer Agent

SPF – Sender Policy Framework

DKIM – Domain Keys Identified Mail

DMARC – Domain-based Message Authentication, Reporting, and Conformance

Bibliography

I. Official
II. Unofficial
III. Critical
IV. General

I. Official

[Barr 2012]

Barr, Jeff (2012). Programmable Feedback Notification for the Simple Email Service. June 26th 2012. Available at: <https://aws.amazon.com/blogs/aws/programmable-feedback-notification-for-the-simple-email-service/>

[Deshpande 2012]

Deshpande, Rohan (2012). Handling Bounces and Complaints. *AWS Messaging and Targeting Blog*. June 26th 2012. Available at: <https://aws.amazon.com/blogs/messaging-and-targeting/handling-bounces-and-complaints/>

[Weir-Jones 2023]

Weir-Jones, Toby (2023). How to Build an Email Service on SES. *AWS Messaging and Targeting Blog*. Available at: <https://aws.amazon.com/blogs/messaging-and-targeting/how-to-build-an-email-service-on-ses/?ck_subscriber_id=1560524742>

[Nguyen 2023]

Nguyen, Tristan (2023). A Guide to Maintaining a Healthy Email Database. Available at: <https://aws.amazon.com/blogs/messaging-and-targeting/guide-to-maintaining-healthy-email-database/?ck_subscriber_id=1560524742>

CloudFormation

NOTE

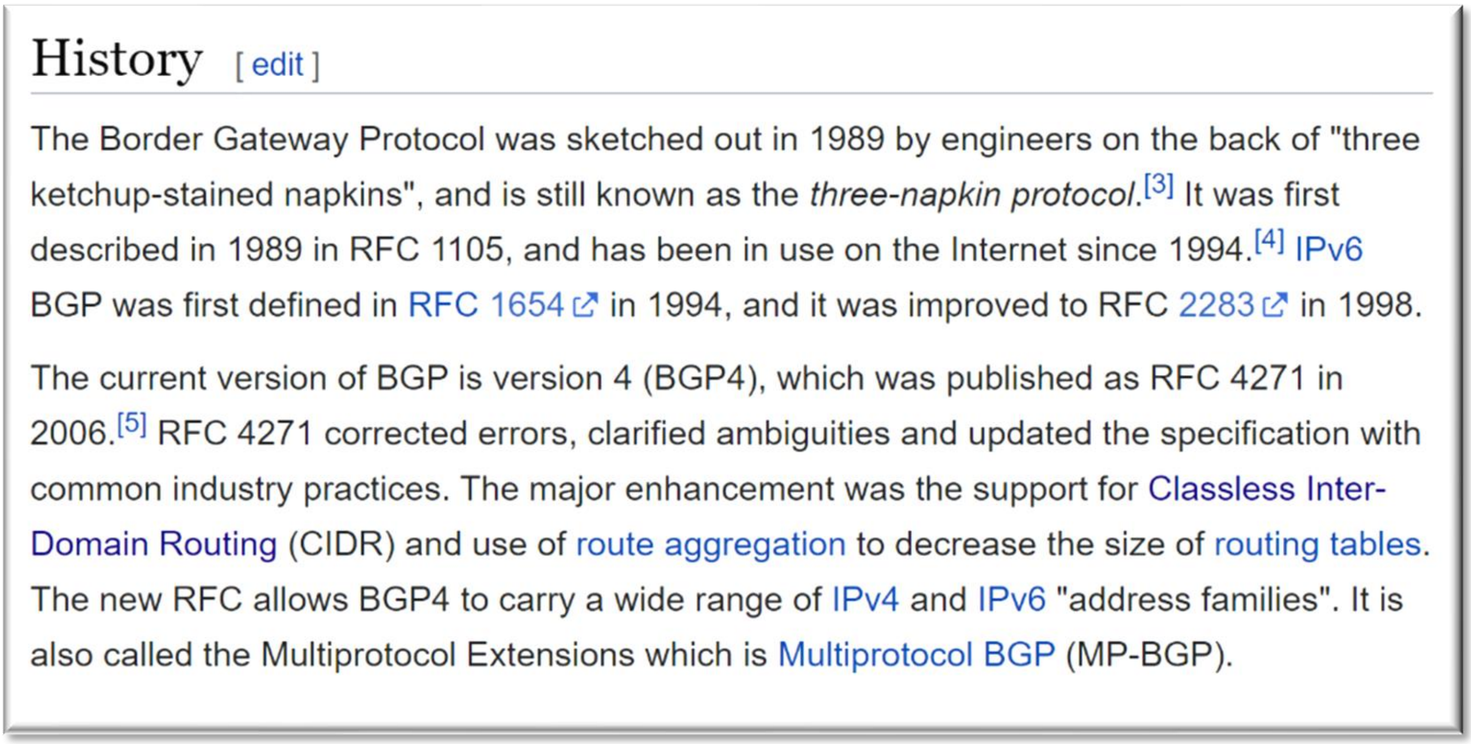
CloudFront or CloudFormation?

Some people get confused between the terms “CloudFormation” and “CloudFront”. Cloudfor-**mation** is about auto-**mation**.

Cloud*Front* bears the *brunt* of those incoming requests (using caching and so on). CloudFront’s edge locations get out there and do the dirty work of brining content close to customers. CloudFormation –to its elation—is well away from the front line.

Direct Connect

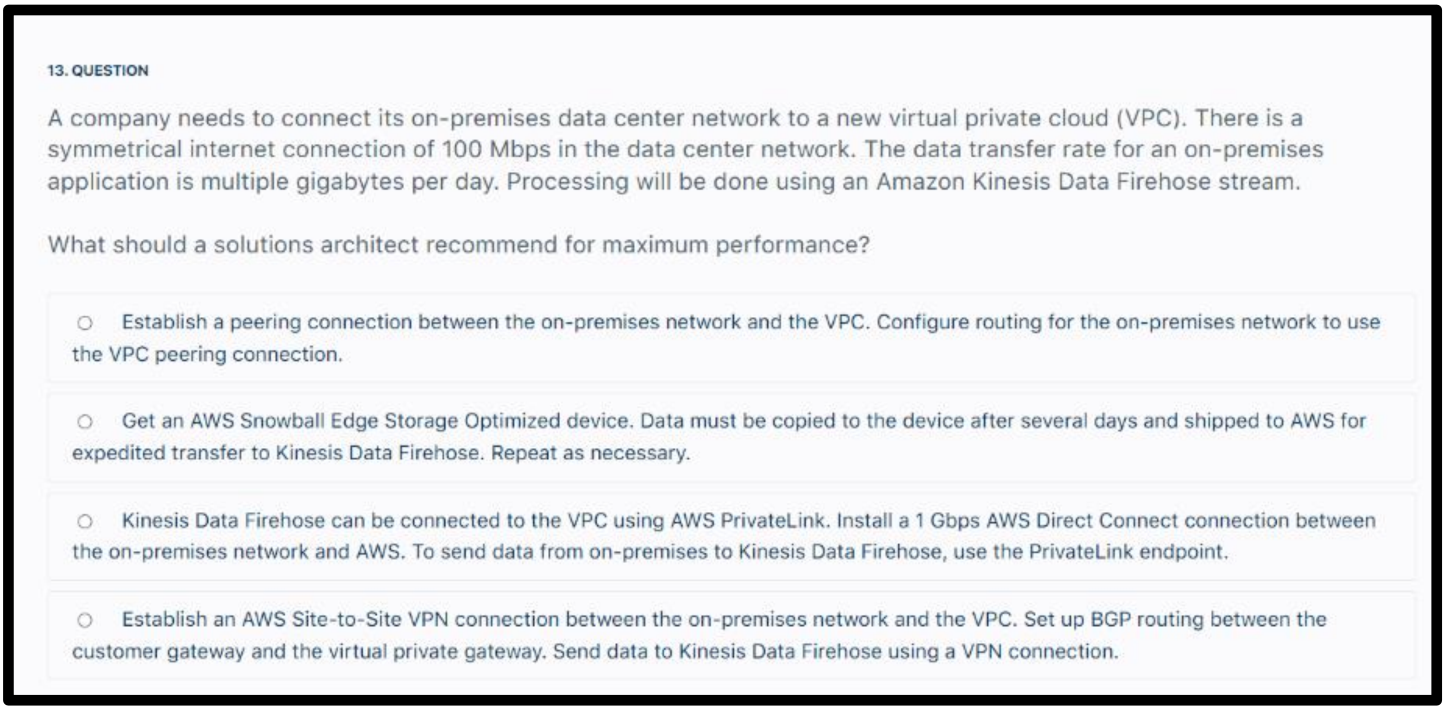
What on earth is BGP?



History [edit]

The Border Gateway Protocol was sketched out in 1989 by engineers on the back of "three ketchup-stained napkins", and is still known as the *three-napkin protocol*.[3] It was first described in 1989 in RFC 1105, and has been in use on the Internet since 1994.[4] IPv6 BGP was first defined in RFC 1654 in 1994, and it was improved to RFC 2283 in 1998.

The current version of BGP is version 4 (BGP4), which was published as RFC 4271 in 2006.[5] RFC 4271 corrected errors, clarified ambiguities and updated the specification with common industry practices. The major enhancement was the support for Classless Inter-Domain Routing (CIDR) and use of route aggregation to decrease the size of routing tables. The new RFC allows BGP4 to carry a wide range of IPv4 and IPv6 "address families". It is also called the Multiprotocol Extensions which is Multiprotocol BGP (MP-BGP).



13. QUESTION

A company needs to connect its on-premises data center network to a new virtual private cloud (VPC). There is a symmetrical internet connection of 100 Mbps in the data center network. The data transfer rate for an on-premises application is multiple gigabytes per day. Processing will be done using an Amazon Kinesis Data Firehose stream.

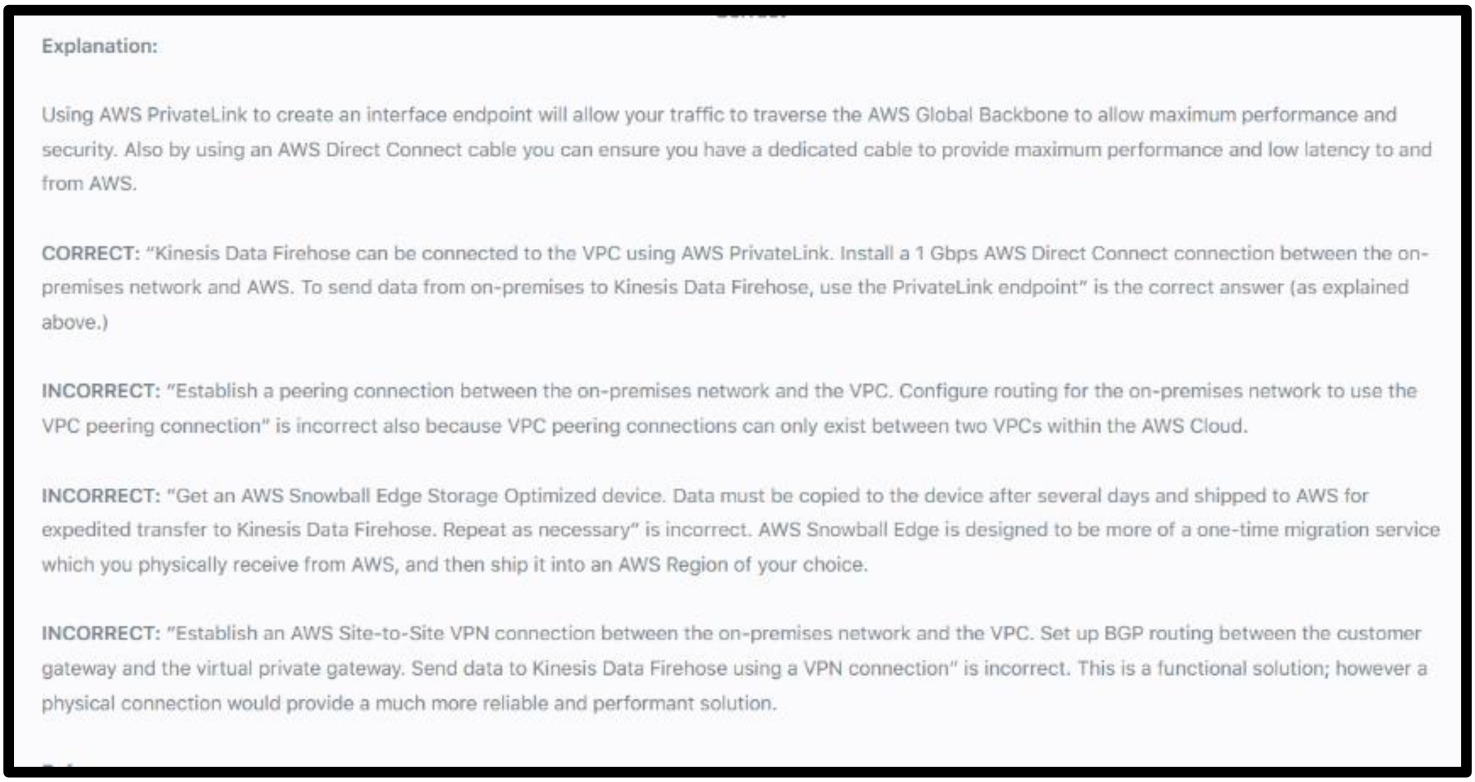
What should a solutions architect recommend for maximum performance?

○ Establish a peering connection between the on-premises network and the VPC. Configure routing for the on-premises network to use the VPC peering connection.

○ Get an AWS Snowball Edge Storage Optimized device. Data must be copied to the device after several days and shipped to AWS for expedited transfer to Kinesis Data Firehose. Repeat as necessary.

○ Kinesis Data Firehose can be connected to the VPC using AWS PrivateLink. Install a 1 Gbps AWS Direct Connect connection between the on-premises network and AWS. To send data from on-premises to Kinesis Data Firehose, use the PrivateLink endpoint.

○ Establish an AWS Site-to-Site VPN connection between the on-premises network and the VPC. Set up BGP routing between the customer gateway and the virtual private gateway. Send data to Kinesis Data Firehose using a VPN connection.



Explanation:

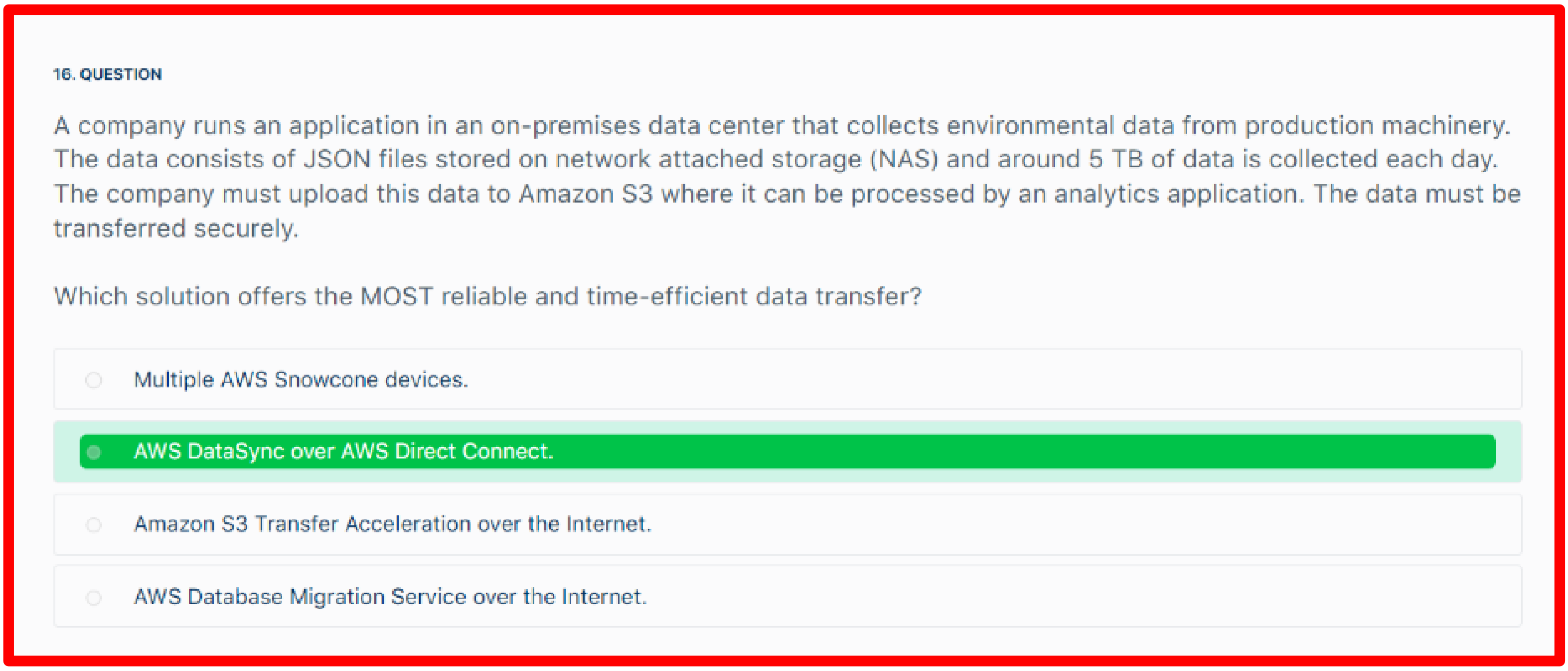
Using AWS PrivateLink to create an interface endpoint will allow your traffic to traverse the AWS Global Backbone to allow maximum performance and security. Also by using an AWS Direct Connect cable you can ensure you have a dedicated cable to provide maximum performance and low latency to and from AWS.

CORRECT: "Kinesis Data Firehose can be connected to the VPC using AWS PrivateLink. Install a 1 Gbps AWS Direct Connect connection between the on-premises network and AWS. To send data from on-premises to Kinesis Data Firehose, use the PrivateLink endpoint" is the correct answer (as explained above.)

INCORRECT: "Establish a peering connection between the on-premises network and the VPC. Configure routing for the on-premises network to use the VPC peering connection" is incorrect also because VPC peering connections can only exist between two VPCs within the AWS Cloud.

INCORRECT: "Get an AWS Snowball Edge Storage Optimized device. Data must be copied to the device after several days and shipped to AWS for expedited transfer to Kinesis Data Firehose. Repeat as necessary" is incorrect. AWS Snowball Edge is designed to be more of a one-time migration service which you physically receive from AWS, and then ship it into an AWS Region of your choice.

INCORRECT: "Establish an AWS Site-to-Site VPN connection between the on-premises network and the VPC. Set up BGP routing between the customer gateway and the virtual private gateway. Send data to Kinesis Data Firehose using a VPN connection" is incorrect. This is a functional solution; however a physical connection would provide a much more reliable and performant solution.



16. QUESTION

A company runs an application in an on-premises data center that collects environmental data from production machinery. The data consists of JSON files stored on network attached storage (NAS) and around 5 TB of data is collected each day. The company must upload this data to Amazon S3 where it can be processed by an analytics application. The data must be transferred securely.

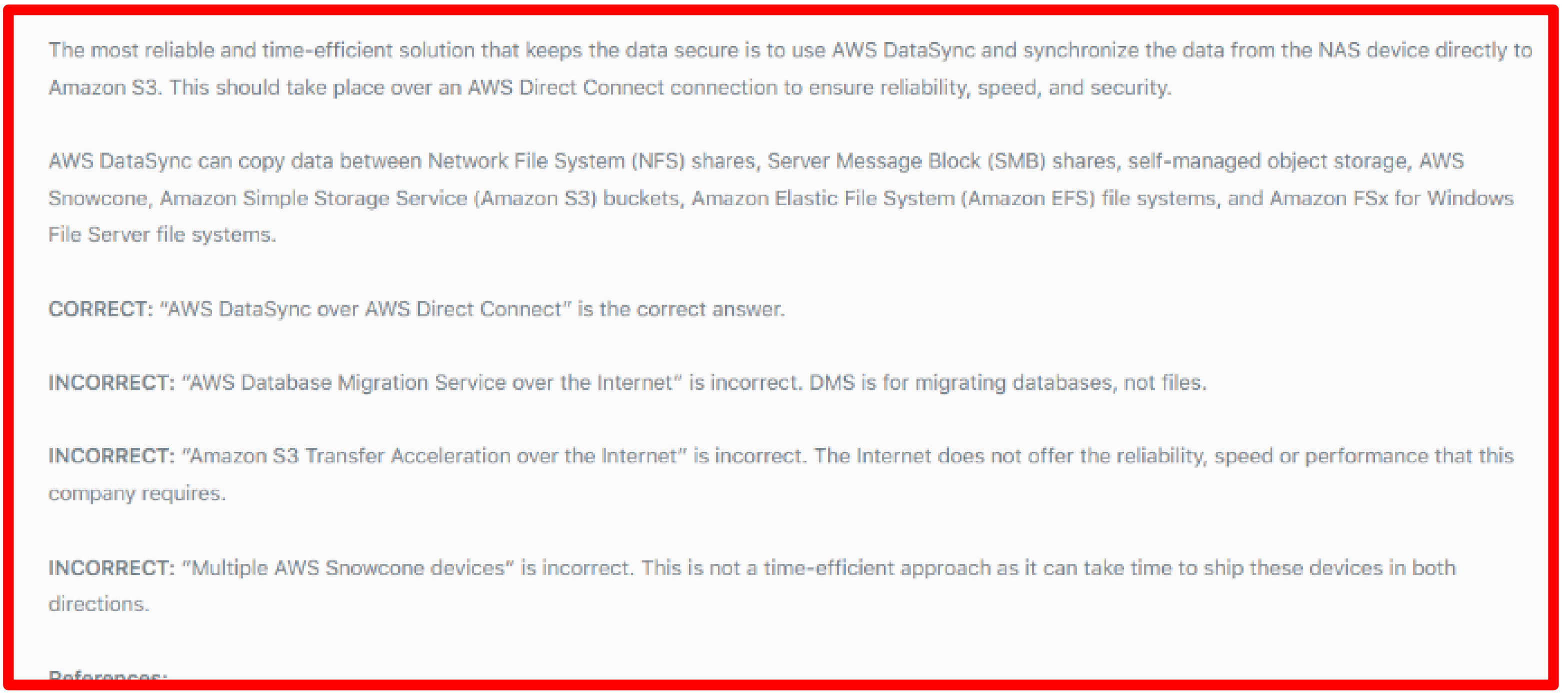
Which solution offers the MOST reliable and time-efficient data transfer?

○ Multiple AWS Snowcone devices.

● AWS DataSync over AWS Direct Connect.

○ Amazon S3 Transfer Acceleration over the Internet.

○ AWS Database Migration Service over the Internet.



The most reliable and time-efficient solution that keeps the data secure is to use AWS DataSync and synchronize the data from the NAS device directly to Amazon S3. This should take place over an AWS Direct Connect connection to ensure reliability, speed, and security.

AWS DataSync can copy data between Network File System (NFS) shares, Server Message Block (SMB) shares, self-managed object storage, AWS Snowcone, Amazon Simple Storage Service (Amazon S3) buckets, Amazon Elastic File System (Amazon EFS) file systems, and Amazon FSx for Windows File Server file systems.

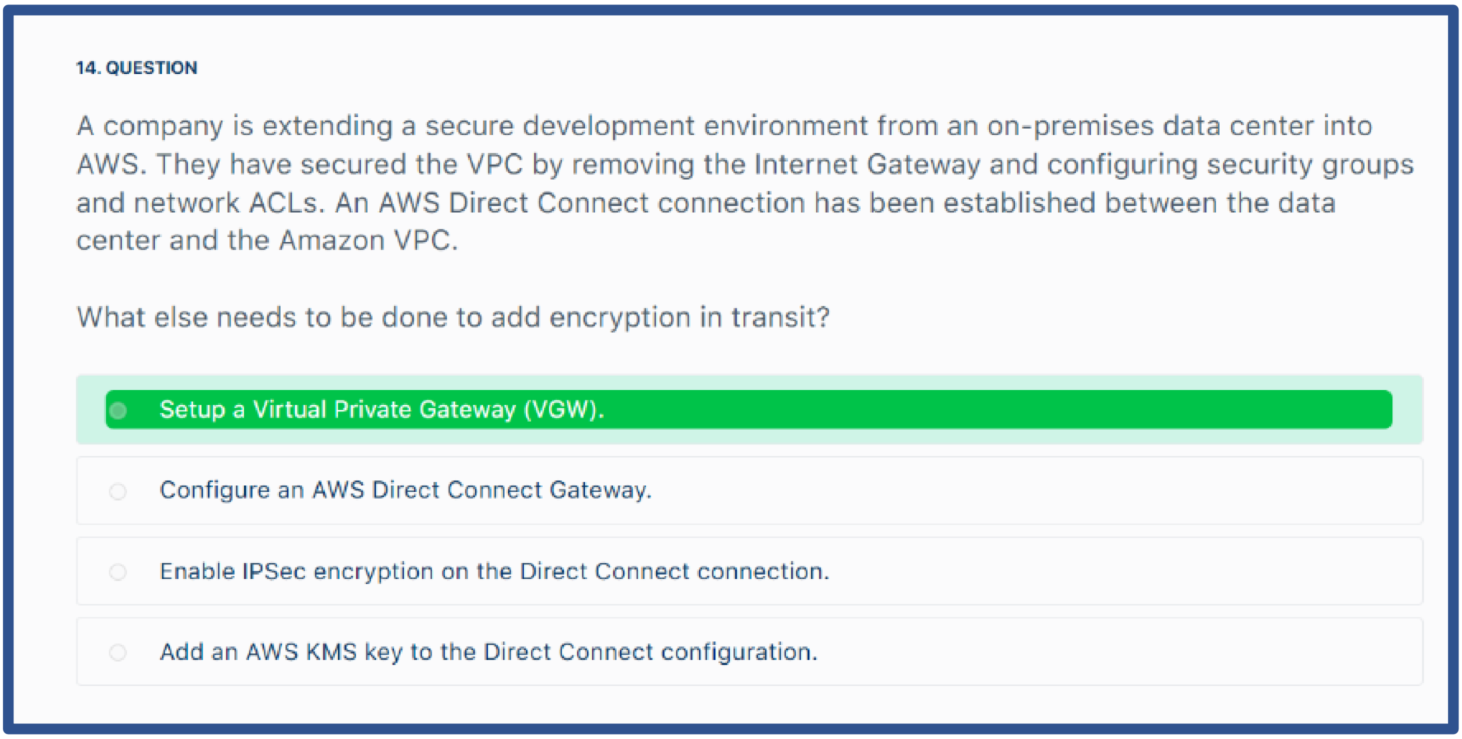
CORRECT: "AWS DataSync over AWS Direct Connect" is the correct answer.

INCORRECT: "AWS Database Migration Service over the Internet" is incorrect. DMS is for migrating databases, not files.

INCORRECT: "Amazon S3 Transfer Acceleration over the Internet" is incorrect. The Internet does not offer the reliability, speed or performance that this company requires.

INCORRECT: "Multiple AWS Snowcone devices" is incorrect. This is not a time-efficient approach as it can take time to ship these devices in both directions.

References:

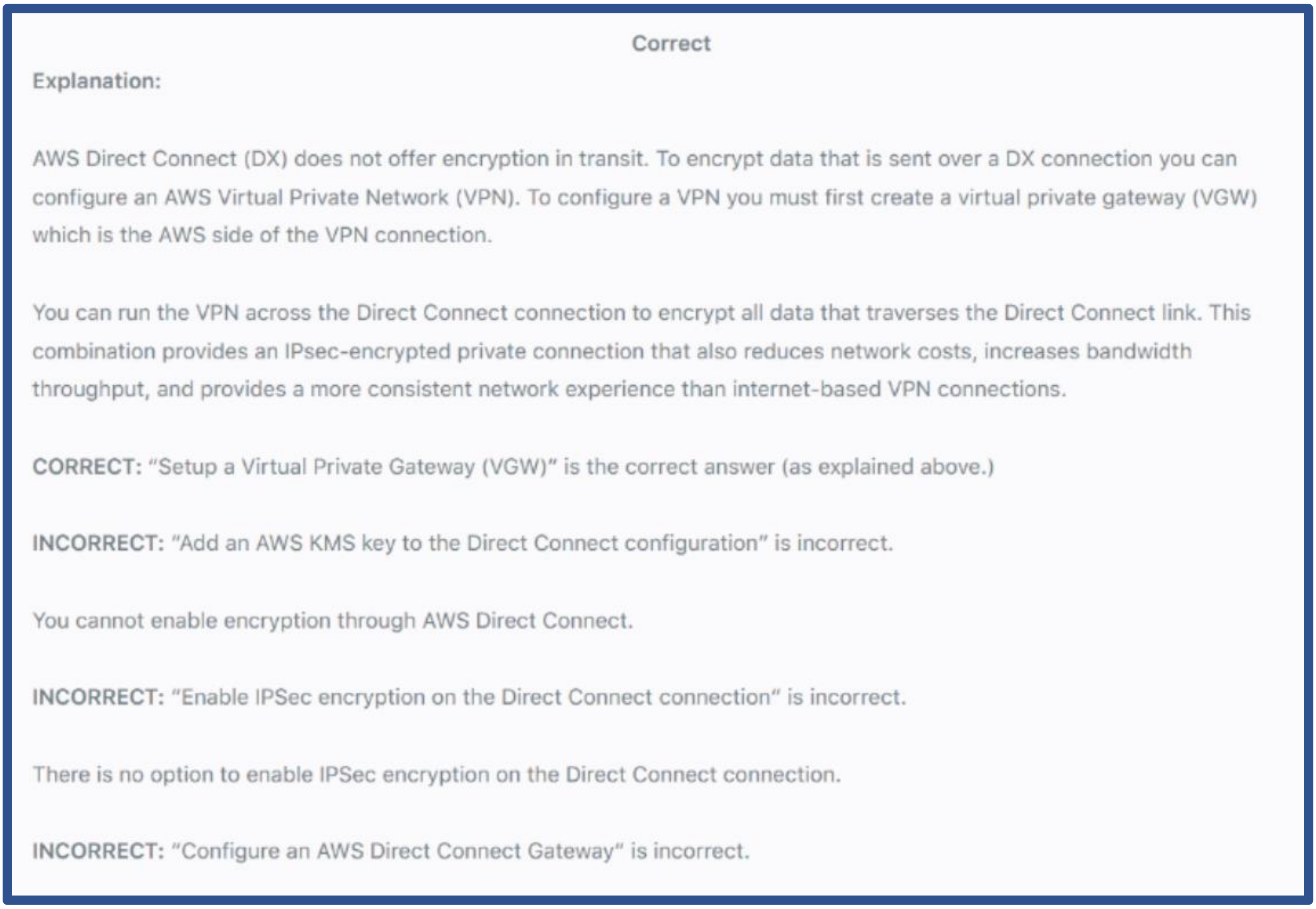


14. QUESTION

A company is extending a secure development environment from an on-premises data center into AWS. They have secured the VPC by removing the Internet Gateway and configuring security groups and network ACLs. An AWS Direct Connect connection has been established between the data center and the Amazon VPC.

What else needs to be done to add encryption in transit?

- ● Setup a Virtual Private Gateway (VGW).
- ○ Configure an AWS Direct Connect Gateway.
- ○ Enable IPSec encryption on the Direct Connect connection.
- ○ Add an AWS KMS key to the Direct Connect configuration.



Correct

Explanation:

AWS Direct Connect (DX) does not offer encryption in transit. To encrypt data that is sent over a DX connection you can configure an AWS Virtual Private Network (VPN). To configure a VPN you must first create a virtual private gateway (VGW) which is the AWS side of the VPN connection.

You can run the VPN across the Direct Connect connection to encrypt all data that traverses the Direct Connect link. This combination provides an IPsec-encrypted private connection that also reduces network costs, increases bandwidth throughput, and provides a more consistent network experience than internet-based VPN connections.

CORRECT: "Setup a Virtual Private Gateway (VGW)" is the correct answer (as explained above.)

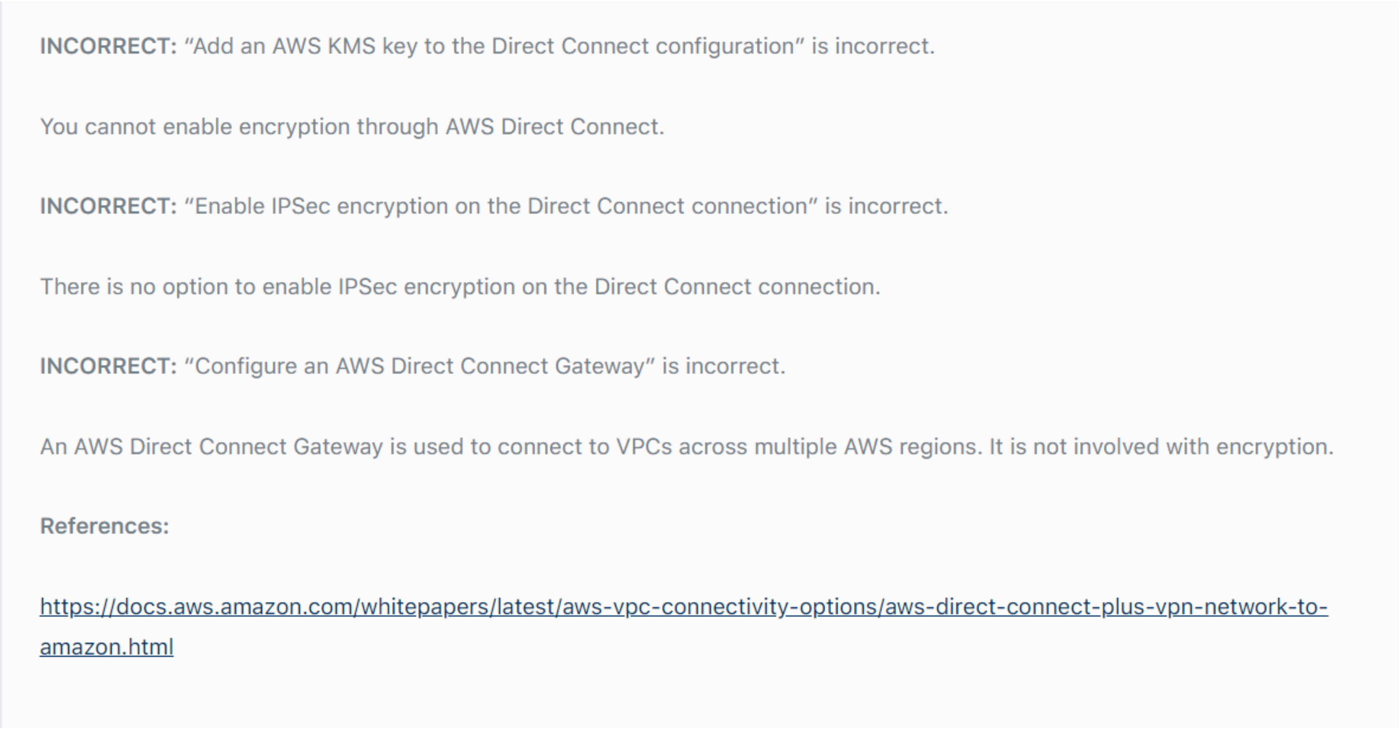
INCORRECT: "Add an AWS KMS key to the Direct Connect configuration" is incorrect.

You cannot enable encryption through AWS Direct Connect.

INCORRECT: "Enable IPSec encryption on the Direct Connect connection" is incorrect.

There is no option to enable IPSec encryption on the Direct Connect connection.

INCORRECT: "Configure an AWS Direct Connect Gateway" is incorrect.



INCORRECT: "Add an AWS KMS key to the Direct Connect configuration" is incorrect.

You cannot enable encryption through AWS Direct Connect.

INCORRECT: "Enable IPSec encryption on the Direct Connect connection" is incorrect.

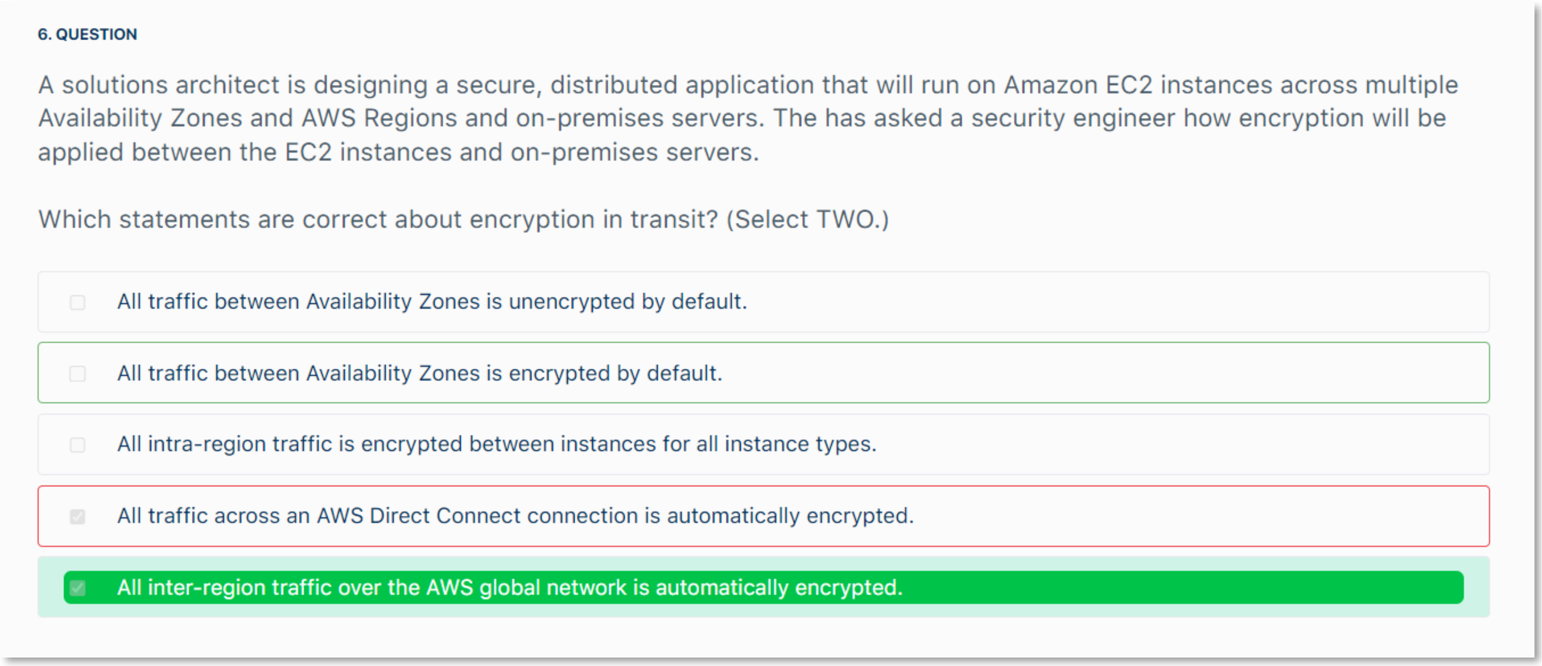
There is no option to enable IPSec encryption on the Direct Connect connection.

INCORRECT: "Configure an AWS Direct Connect Gateway" is incorrect.

An AWS Direct Connect Gateway is used to connect to VPCs across multiple AWS regions. It is not involved with encryption.

References:

https://docs.aws.amazon.com/whitepapers/latest/aws-vpc-connectivity-options/aws-direct-connect-plus-vpn-network-to-amazon.html



6. QUESTION

A solutions architect is designing a secure, distributed application that will run on Amazon EC2 instances across multiple Availability Zones and AWS Regions and on-premises servers. The has asked a security engineer how encryption will be applied between the EC2 instances and on-premises servers.

Which statements are correct about encryption in transit? (Select TWO.)

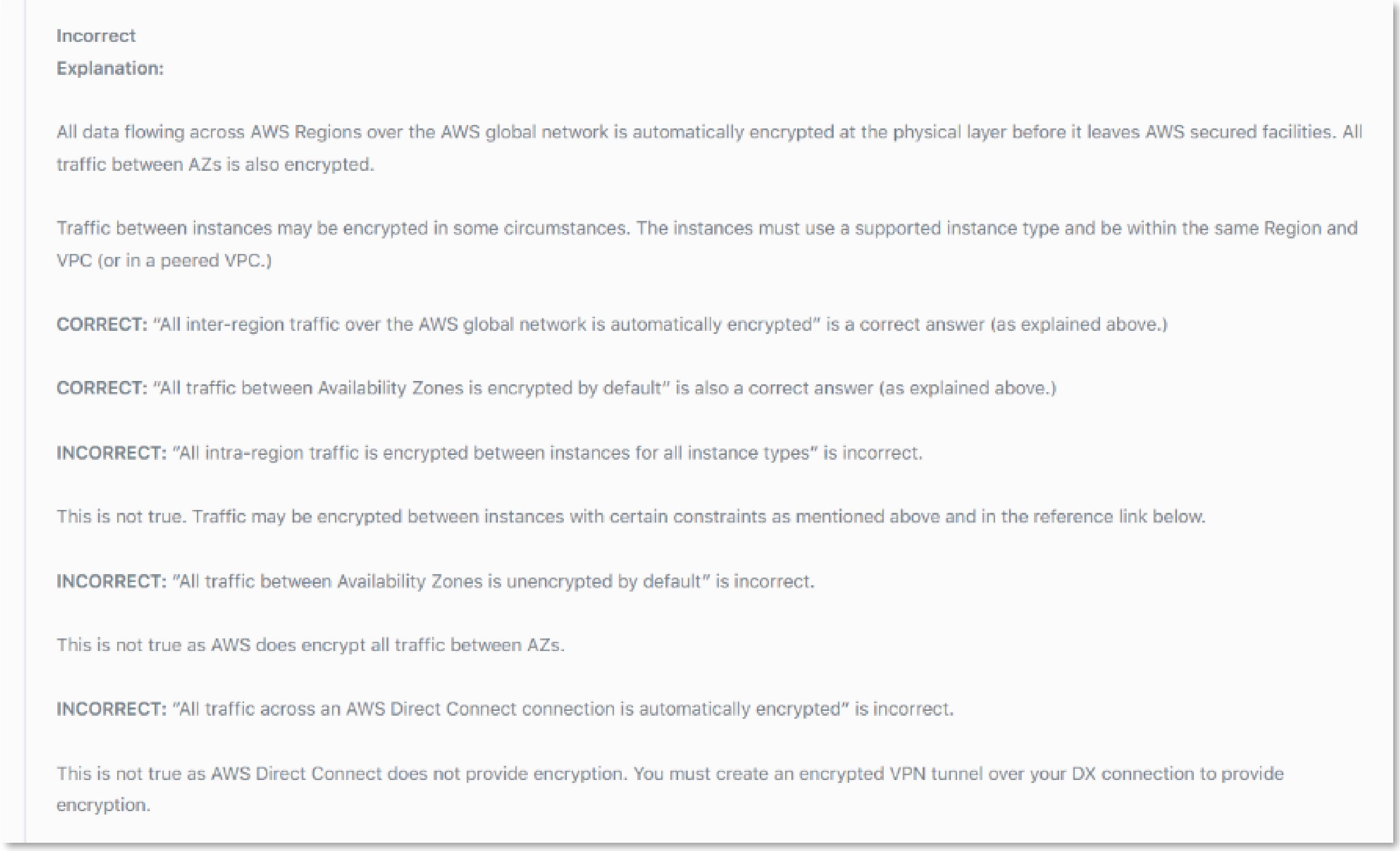
☐ All traffic between Availability Zones is unencrypted by default.

☐ All traffic between Availability Zones is encrypted by default.

☐ All intra-region traffic is encrypted between instances for all instance types.

☑ All traffic across an AWS Direct Connect connection is automatically encrypted.

☑ All inter-region traffic over the AWS global network is automatically encrypted.



Incorrect

Explanation:

All data flowing across AWS Regions over the AWS global network is automatically encrypted at the physical layer before it leaves AWS secured facilities. All traffic between AZs is also encrypted.

Traffic between instances may be encrypted in some circumstances. The instances must use a supported instance type and be within the same Region and VPC (or in a peered VPC.)

CORRECT: "All inter-region traffic over the AWS global network is automatically encrypted" is a correct answer (as explained above.)

CORRECT: "All traffic between Availability Zones is encrypted by default" is also a correct answer (as explained above.)

INCORRECT: "All intra-region traffic is encrypted between instances for all instance types" is incorrect.

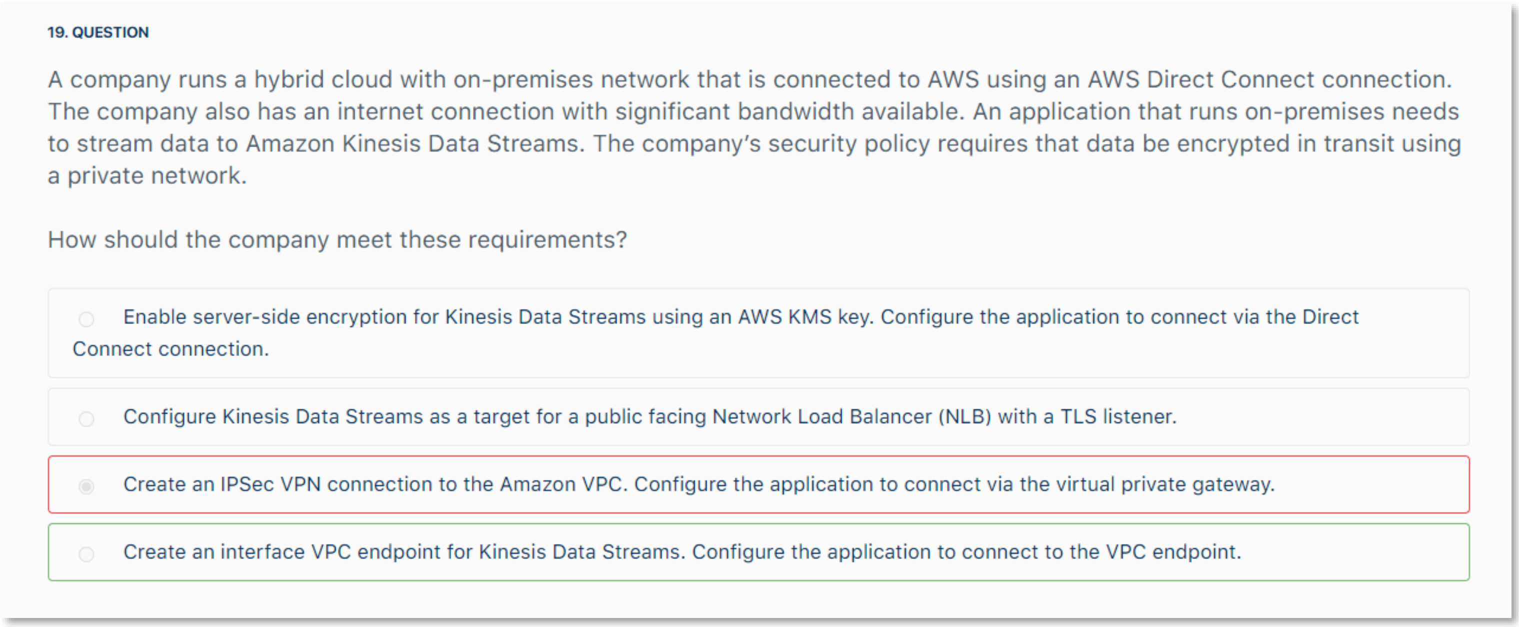
This is not true. Traffic may be encrypted between instances with certain constraints as mentioned above and in the reference link below.

INCORRECT: "All traffic between Availability Zones is unencrypted by default" is incorrect.

This is not true as AWS does encrypt all traffic between AZs.

INCORRECT: "All traffic across an AWS Direct Connect connection is automatically encrypted" is incorrect.

This is not true as AWS Direct Connect does not provide encryption. You must create an encrypted VPN tunnel over your DX connection to provide encryption.



19. QUESTION

A company runs a hybrid cloud with on-premises network that is connected to AWS using an AWS Direct Connect connection. The company also has an internet connection with significant bandwidth available. An application that runs on-premises needs to stream data to Amazon Kinesis Data Streams. The company's security policy requires that data be encrypted in transit using a private network.

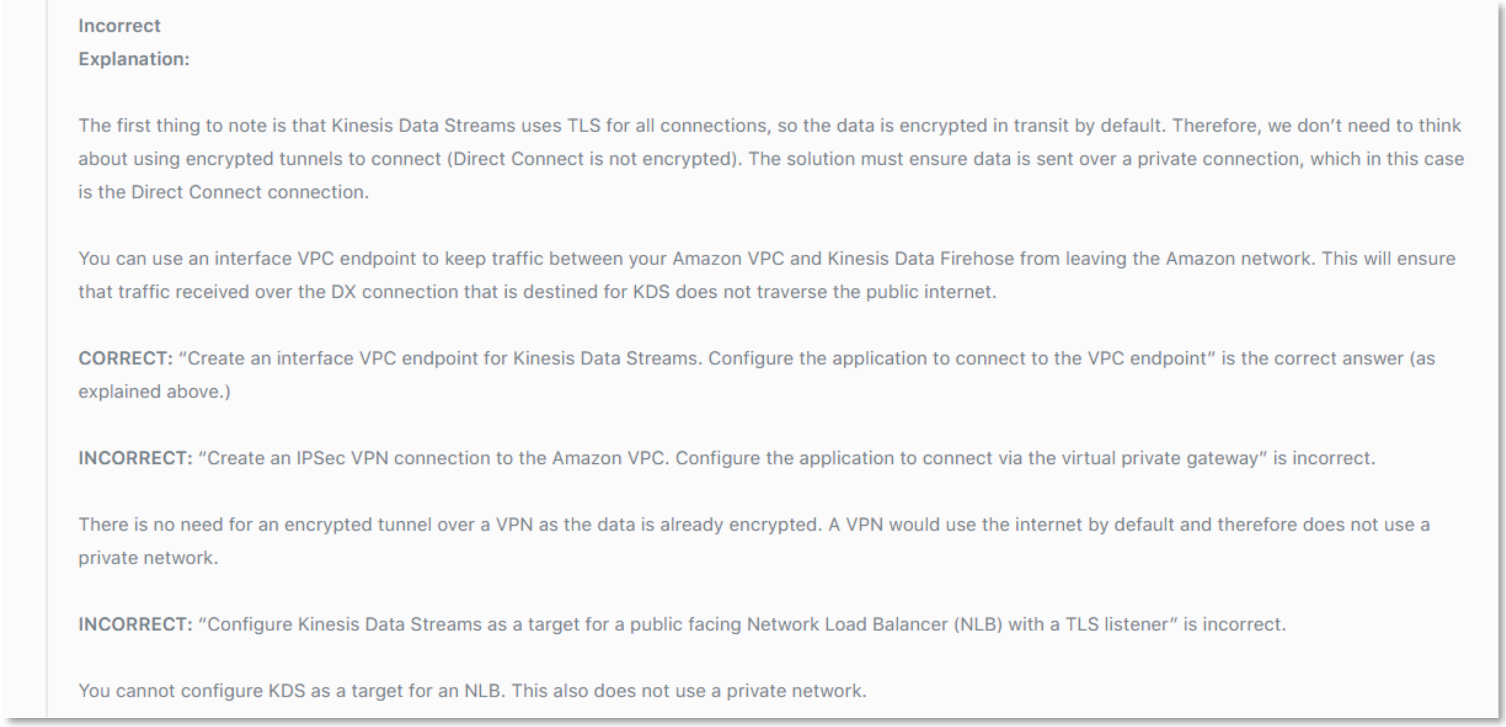
How should the company meet these requirements?

○ Enable server-side encryption for Kinesis Data Streams using an AWS KMS key. Configure the application to connect via the Direct Connect connection.

○ Configure Kinesis Data Streams as a target for a public facing Network Load Balancer (NLB) with a TLS listener.

● Create an IPSec VPN connection to the Amazon VPC. Configure the application to connect via the virtual private gateway.

○ Create an interface VPC endpoint for Kinesis Data Streams. Configure the application to connect to the VPC endpoint.



Incorrect

Explanation:

The first thing to note is that Kinesis Data Streams uses TLS for all connections, so the data is encrypted in transit by default. Therefore, we don't need to think about using encrypted tunnels to connect (Direct Connect is not encrypted). The solution must ensure data is sent over a private connection, which in this case is the Direct Connect connection.

You can use an interface VPC endpoint to keep traffic between your Amazon VPC and Kinesis Data Firehose from leaving the Amazon network. This will ensure that traffic received over the DX connection that is destined for KDS does not traverse the public internet.

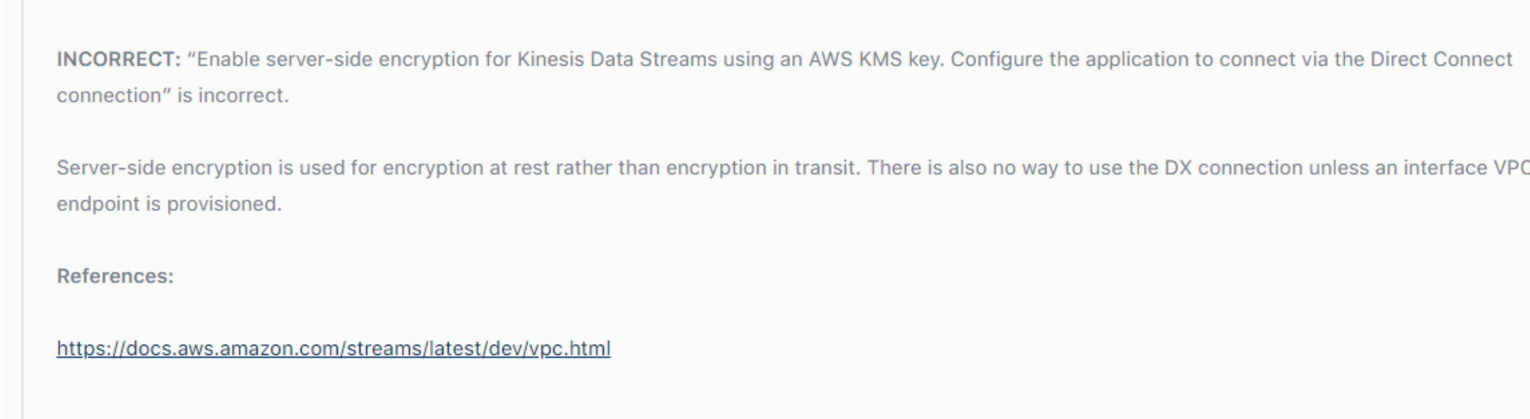
CORRECT: "Create an interface VPC endpoint for Kinesis Data Streams. Configure the application to connect to the VPC endpoint" is the correct answer (as explained above.)

INCORRECT: "Create an IPSec VPN connection to the Amazon VPC. Configure the application to connect via the virtual private gateway" is incorrect.

There is no need for an encrypted tunnel over a VPN as the data is already encrypted. A VPN would use the internet by default and therefore does not use a private network.

INCORRECT: "Configure Kinesis Data Streams as a target for a public facing Network Load Balancer (NLB) with a TLS listener" is incorrect.

You cannot configure KDS as a target for an NLB. This also does not use a private network.



INCORRECT: "Enable server-side encryption for Kinesis Data Streams using an AWS KMS key. Configure the application to connect via the Direct Connect connection" is incorrect.

Server-side encryption is used for encryption at rest rather than encryption in transit. There is also no way to use the DX connection unless an interface VPC endpoint is provisioned.

References:

https://docs.aws.amazon.com/streams/latest/dev/vpc.html

Bibliography

I. Official

II. Unofficial

III. Critical

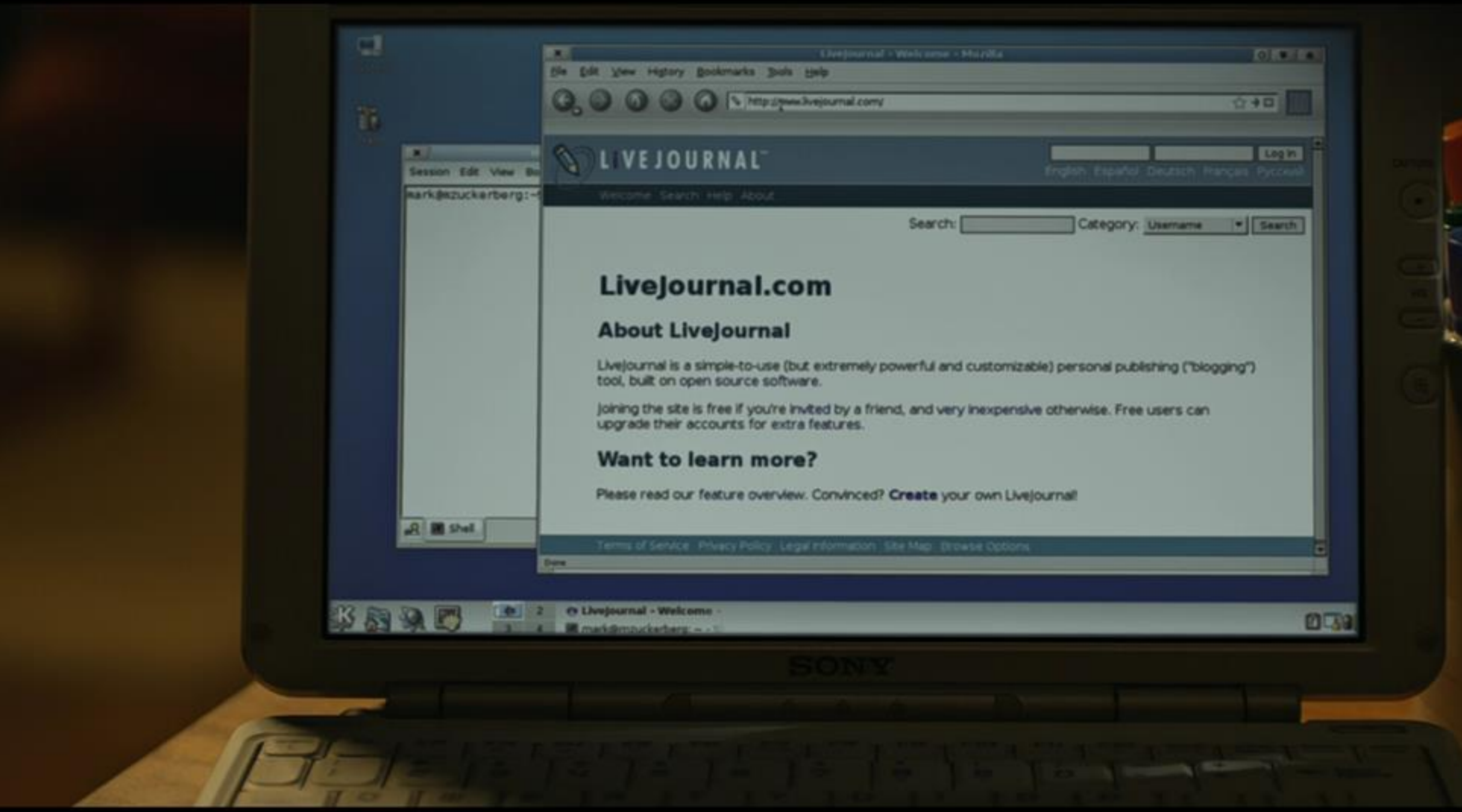
IV. General

[Dzieza 2024]

The Cloud Under the Sea. *The Verge*. April 16th 2024. Available at: <https://www.theverge.com/c/24070570/internet-cables-undersea-deep-repair-ships>

ElastiCache

EcoPulse


LiveJournal - Welcome - Mozilla
File Edit View History Bookmarks Tools Help
http://www.livejournal.com/
LIVEJOURNAL
Log in
English Español Deutsch Français Русский
Welcome Search Help About
Search:
Category: Username
Search
Terms of Service Privacy Policy Legal Information Site Map Browse Options
Done
Session Edit View
mark@mzuckerberg:~
Shell
LiveJournal - Welcome -
mark@mzuckerberg: ~
SONY

LiveJournal.com

About LiveJournal

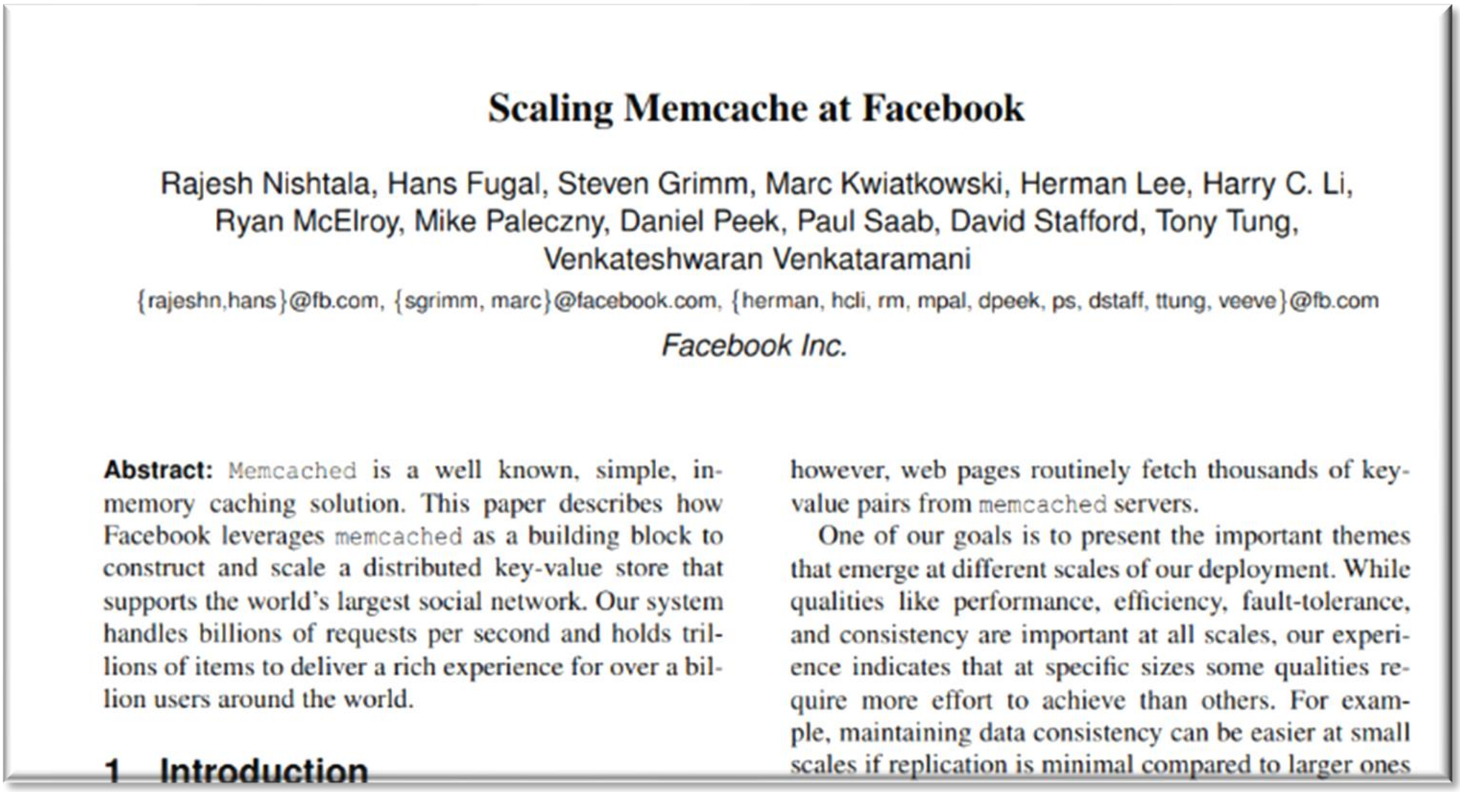
LiveJournal is a simple-to-use (but extremely powerful and customizable) personal publishing ("blogging") tool, built on open source software.

Joining the site is free if you're invited by a friend, and very inexpensive otherwise. Free users can upgrade their accounts for extra features.

Want to learn more?

Please read our feature overview. Convinced? **Create** your own LiveJournal!

Scene from the film *The Social Network*, which shows Mark Zuckerberg using LiveJournal. Why is this important? Because LiveJournal was created by Brad Fitzpatrick, who is responsible for Memcached.



Scaling Memcache at Facebook

Rajesh Nishtala, Hans Fugal, Steven Grimm, Marc Kwiatkowski, Herman Lee, Harry C. Li,
Ryan McElroy, Mike Paleczny, Daniel Peek, Paul Saab, David Stafford, Tony Tung,
Venkateshwaran Venkataramani

{rajeshn,hans}@fb.com, {sgrimm, marc}@facebook.com, {herman, hcli, rm, mpal, dpeek, ps, dstaff, ttung, veeve}@fb.com

Facebook Inc.

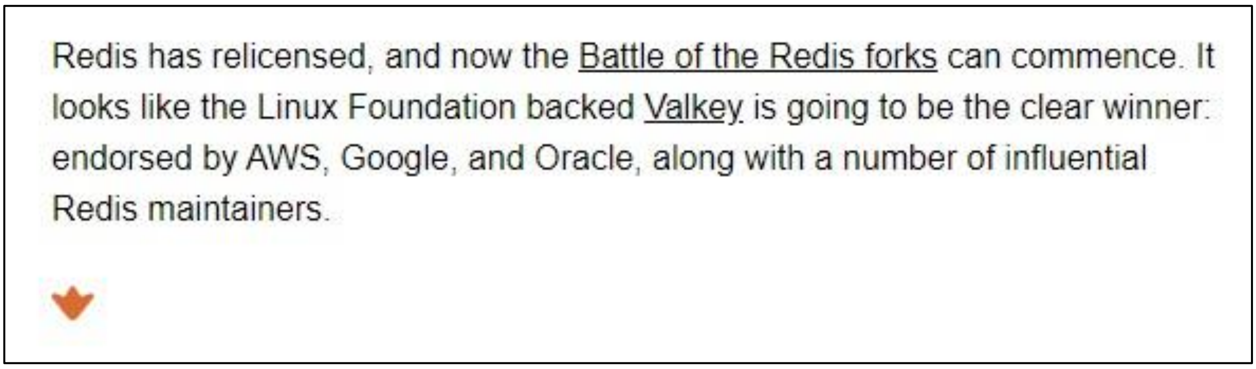
Abstract: `Memcached` is a well known, simple, in-memory caching solution. This paper describes how Facebook leverages `memcached` as a building block to construct and scale a distributed key-value store that supports the world's largest social network. Our system handles billions of requests per second and holds trillions of items to deliver a rich experience for over a billion users around the world.

1 Introduction

however, web pages routinely fetch thousands of key-value pairs from `memcached` servers.

One of our goals is to present the important themes that emerge at different scales of our deployment. While qualities like performance, efficiency, fault-tolerance, and consistency are important at all scales, our experience indicates that at specific sizes some qualities require more effort to achieve than others. For example, maintaining data consistency can be easier at small scales if replication is minimal compared to larger ones

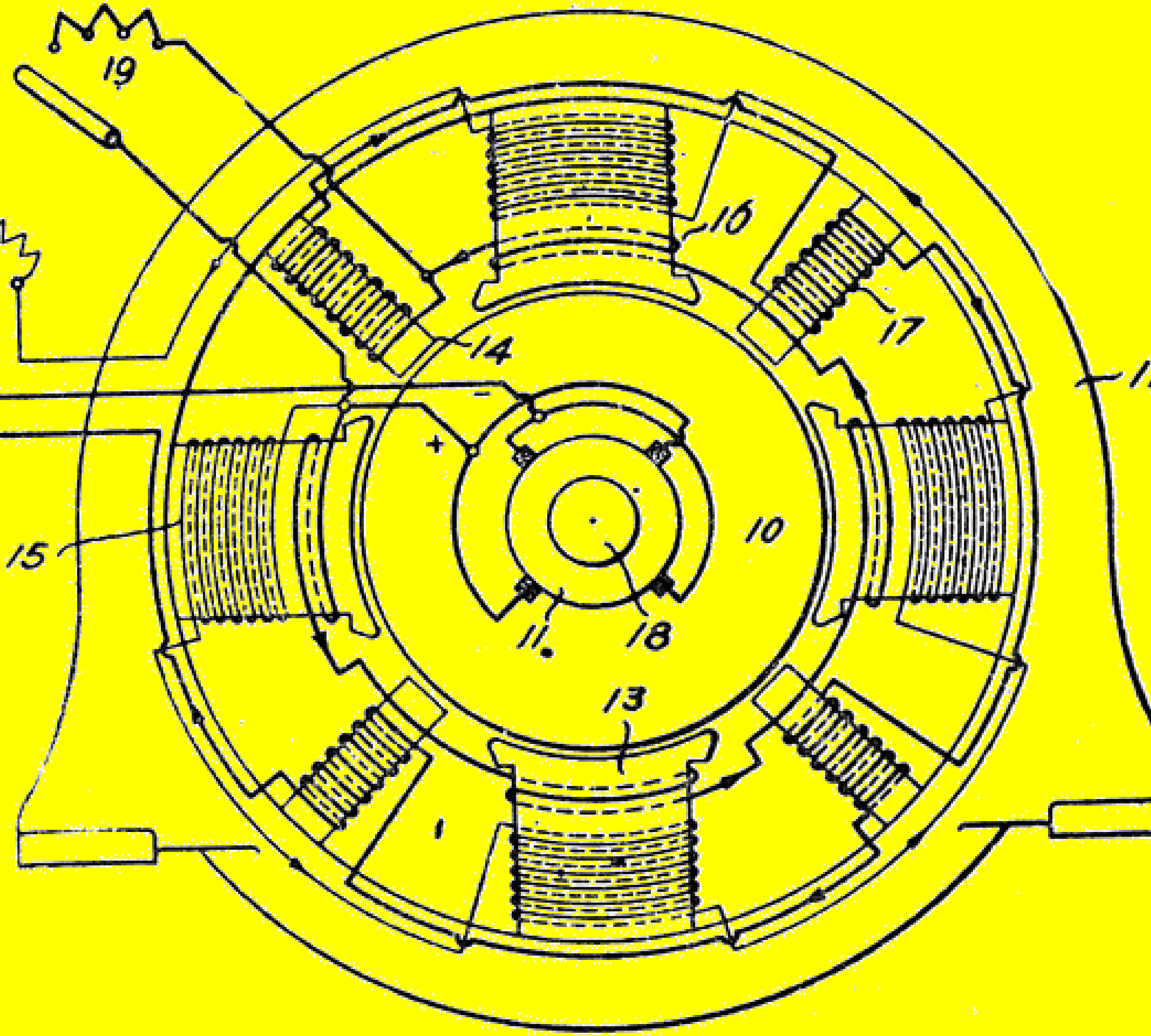
Look at all my notes on Redis and bring them here.

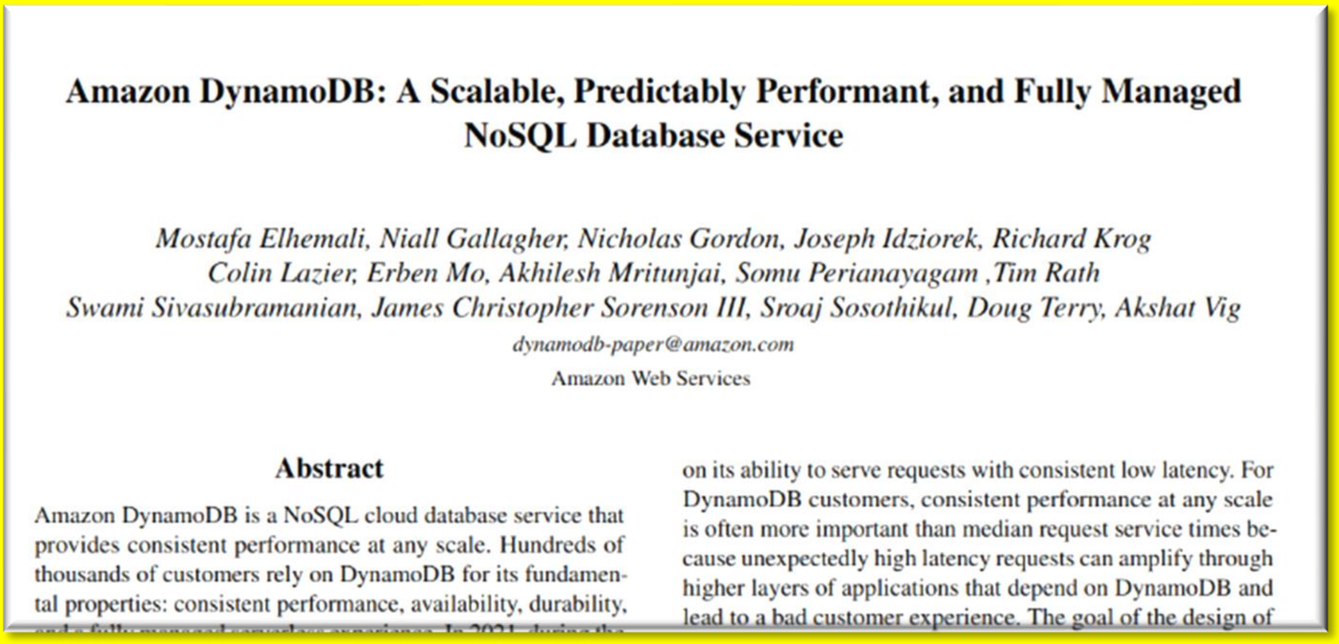


Redis has relicensed, and now the Battle of the Redis forks can commence. It looks like the Linux Foundation backed Valkey is going to be the clear winner: endorsed by AWS, Google, and Oracle, along with a number of influential Redis maintainers.

Email from Corey Quinn on 1st April 2024

DynamoDB





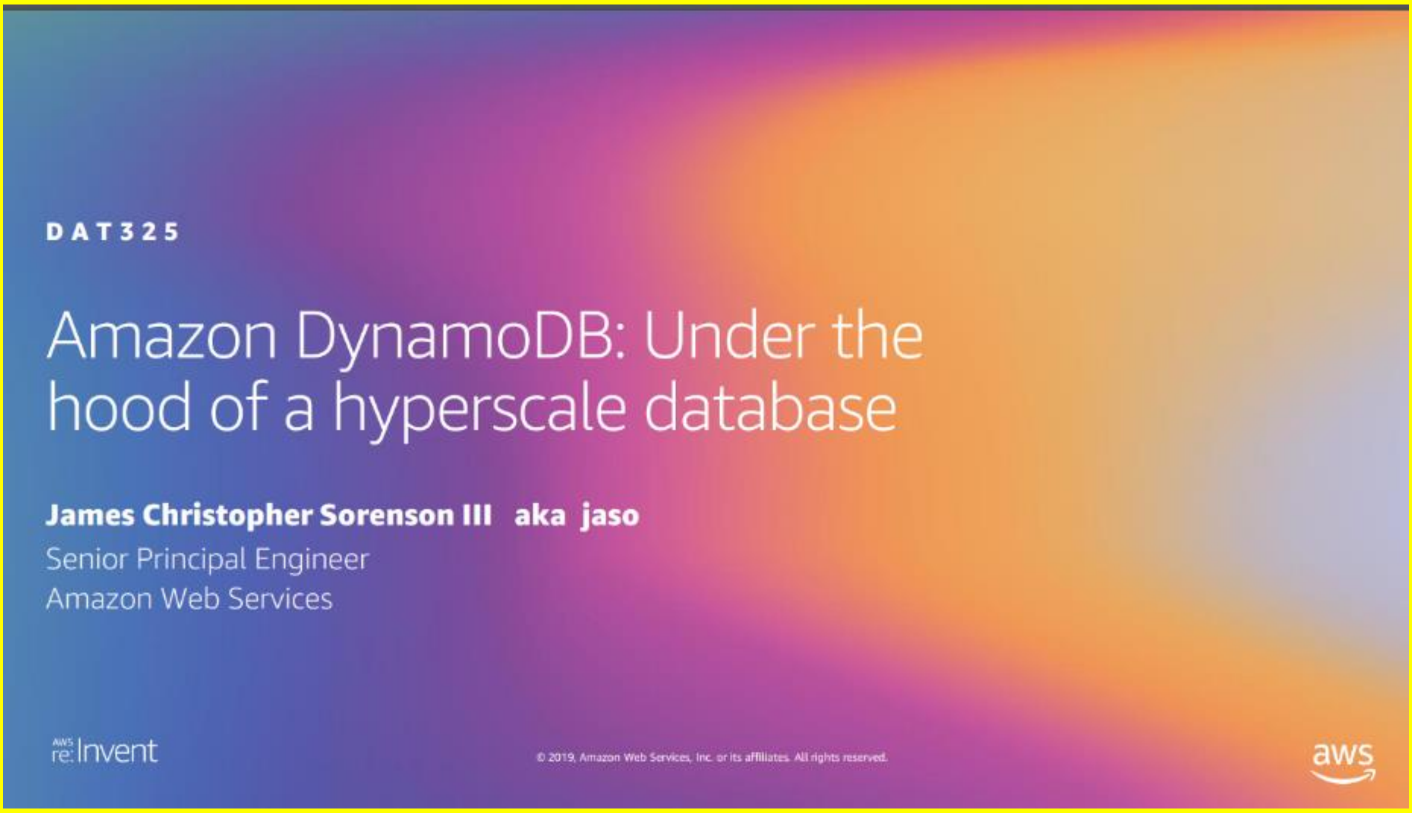
Amazon DynamoDB: A Scalable, Predictably Performant, and Fully Managed NoSQL Database Service

Mostafa Elhemali, Niall Gallagher, Nicholas Gordon, Joseph Idziorek, Richard Krog
Colin Lazier, Erben Mo, Akhilesh Mritunjai, Somu Perianayagam ,Tim Rath
Swami Sivasubramanian, James Christopher Sorenson III, Sroaj Sosothikul, Doug Terry, Akshat Vig
dynamodb-paper@amazon.com
Amazon Web Services

Abstract

Amazon DynamoDB is a NoSQL cloud database service that provides consistent performance at any scale. Hundreds of thousands of customers rely on DynamoDB for its fundamental properties: consistent performance, availability, durability, and a fully managed serverless experience. In 2021, during the

on its ability to serve requests with consistent low latency. For DynamoDB customers, consistent performance at any scale is often more important than median request service times because unexpectedly high latency requests can amplify through higher layers of applications that depend on DynamoDB and lead to a bad customer experience. The goal of the design of

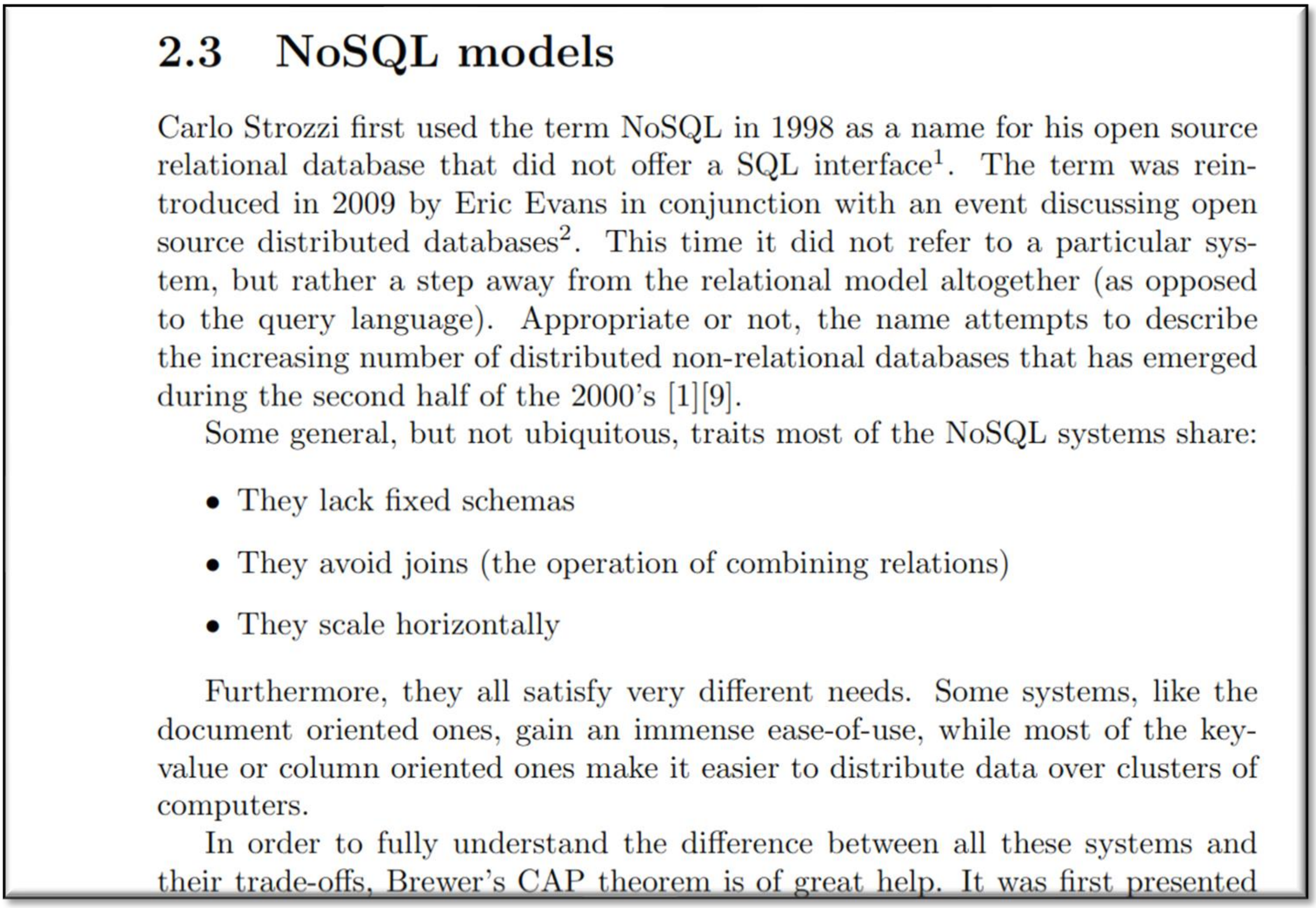


DAT325

Amazon DynamoDB: Under the hood of a hyperscale database

James Christopher Sorenson III aka jaso
Senior Principal Engineer
Amazon Web Services

Jaso Sorenson was one of the architects of DynamoDB. Above he explains how it works at *Reinvent*



2.3 NoSQL models

Carlo Strozzi first used the term NoSQL in 1998 as a name for his open source relational database that did not offer a SQL interface[1]. The term was reintroduced in 2009 by Eric Evans in conjunction with an event discussing open source distributed databases[2]. This time it did not refer to a particular system, but rather a step away from the relational model altogether (as opposed to the query language). Appropriate or not, the name attempts to describe the increasing number of distributed non-relational databases that has emerged during the second half of the 2000's [1][9].

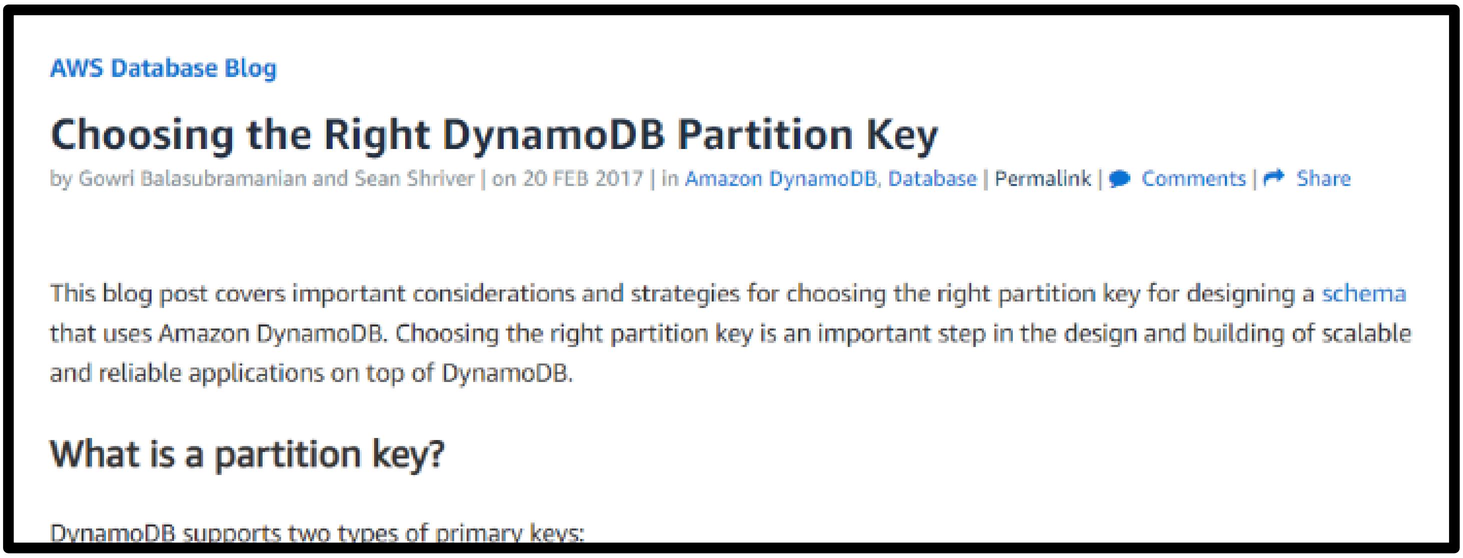
Some general, but not ubiquitous, traits most of the NoSQL systems share:

- They lack fixed schemas
- They avoid joins (the operation of combining relations)
- They scale horizontally

Furthermore, they all satisfy very different needs. Some systems, like the document oriented ones, gain an immense ease-of-use, while most of the key-value or column oriented ones make it easier to distribute data over clusters of computers.

In order to fully understand the difference between all these systems and their trade-offs, Brewer's CAP theorem is of great help. It was first presented

Figure 2 From a piece by Adam Lith and Jakob Mattsson



AWS Database Blog

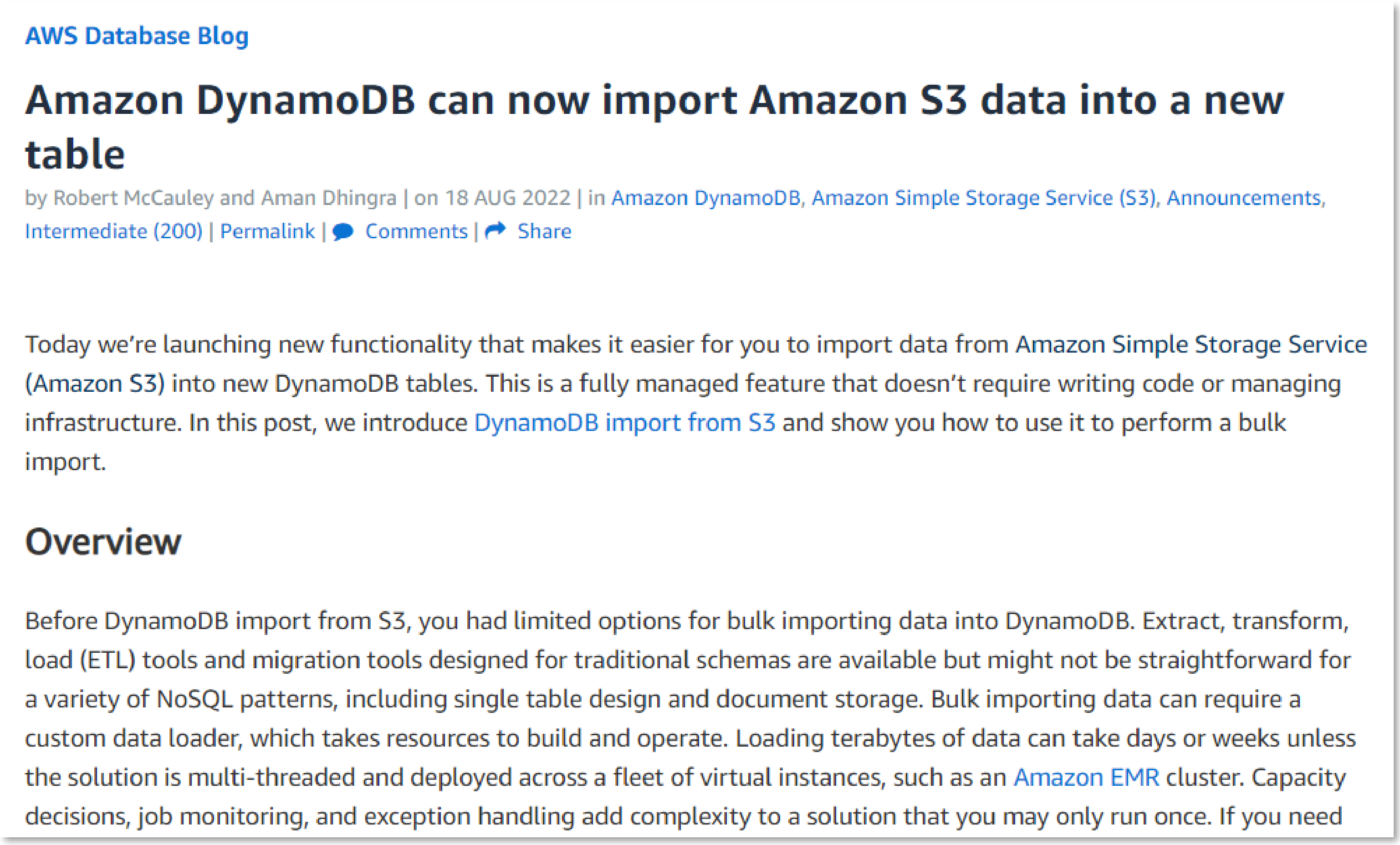
Choosing the Right DynamoDB Partition Key

by Gowri Balasubramanian and Sean Shriver | on 20 FEB 2017 | in Amazon DynamoDB, Database | Permalink | Comments | Share

This blog post covers important considerations and strategies for choosing the right partition key for designing a schema that uses Amazon DynamoDB. Choosing the right partition key is an important step in the design and building of scalable and reliable applications on top of DynamoDB.

What is a partition key?

DynamoDB supports two types of primary keys:



AWS Database Blog

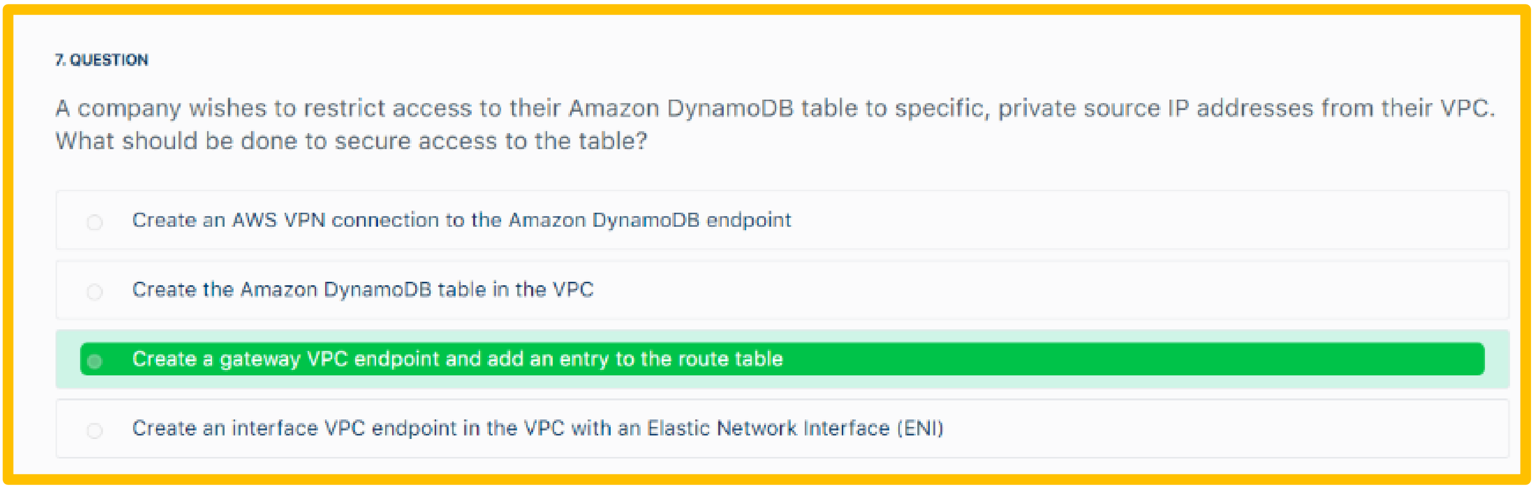
Amazon DynamoDB can now import Amazon S3 data into a new table

by Robert McCauley and Aman Dhingra | on 18 AUG 2022 | in Amazon DynamoDB, Amazon Simple Storage Service (S3), Announcements, Intermediate (200) | Permalink | Comments | Share

Today we're launching new functionality that makes it easier for you to import data from Amazon Simple Storage Service (Amazon S3) into new DynamoDB tables. This is a fully managed feature that doesn't require writing code or managing infrastructure. In this post, we introduce DynamoDB import from S3 and show you how to use it to perform a bulk import.

Overview

Before DynamoDB import from S3, you had limited options for bulk importing data into DynamoDB. Extract, transform, load (ETL) tools and migration tools designed for traditional schemas are available but might not be straightforward for a variety of NoSQL patterns, including single table design and document storage. Bulk importing data can require a custom data loader, which takes resources to build and operate. Loading terabytes of data can take days or weeks unless the solution is multi-threaded and deployed across a fleet of virtual instances, such as an Amazon EMR cluster. Capacity decisions, job monitoring, and exception handling add complexity to a solution that you may only run once. If you need



7. QUESTION

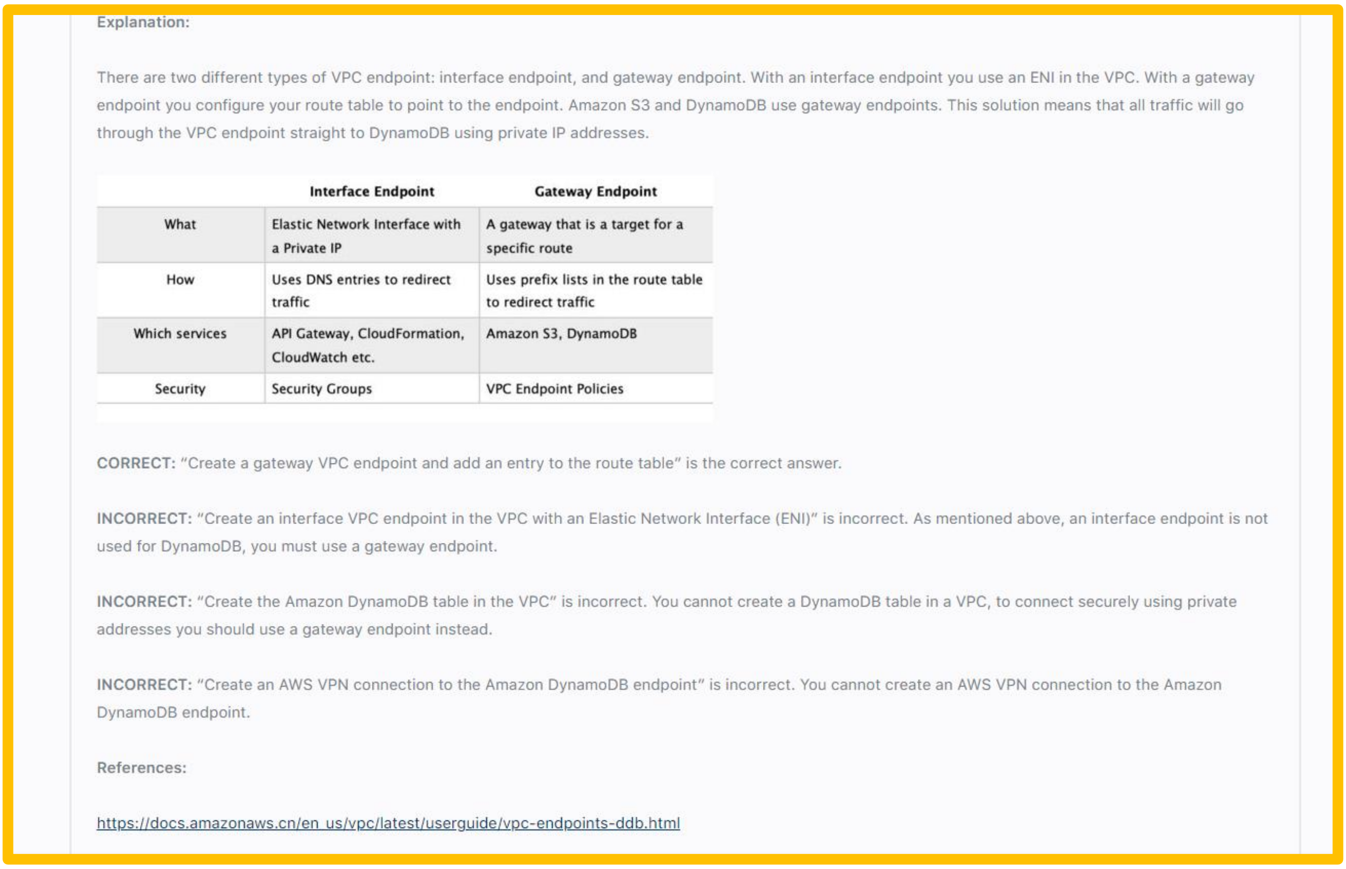
A company wishes to restrict access to their Amazon DynamoDB table to specific, private source IP addresses from their VPC. What should be done to secure access to the table?

○ Create an AWS VPN connection to the Amazon DynamoDB endpoint

○ Create the Amazon DynamoDB table in the VPC

● Create a gateway VPC endpoint and add an entry to the route table

○ Create an interface VPC endpoint in the VPC with an Elastic Network Interface (ENI)



Explanation:

There are two different types of VPC endpoint: interface endpoint, and gateway endpoint. With an interface endpoint you use an ENI in the VPC. With a gateway endpoint you configure your route table to point to the endpoint. Amazon S3 and DynamoDB use gateway endpoints. This solution means that all traffic will go through the VPC endpoint straight to DynamoDB using private IP addresses.

	Interface Endpoint	Gateway Endpoint
What	Elastic Network Interface with a Private IP	A gateway that is a target for a specific route
How	Uses DNS entries to redirect traffic	Uses prefix lists in the route table to redirect traffic
Which services	API Gateway, CloudFormation, CloudWatch etc.	Amazon S3, DynamoDB
Security	Security Groups	VPC Endpoint Policies

CORRECT: "Create a gateway VPC endpoint and add an entry to the route table" is the correct answer.

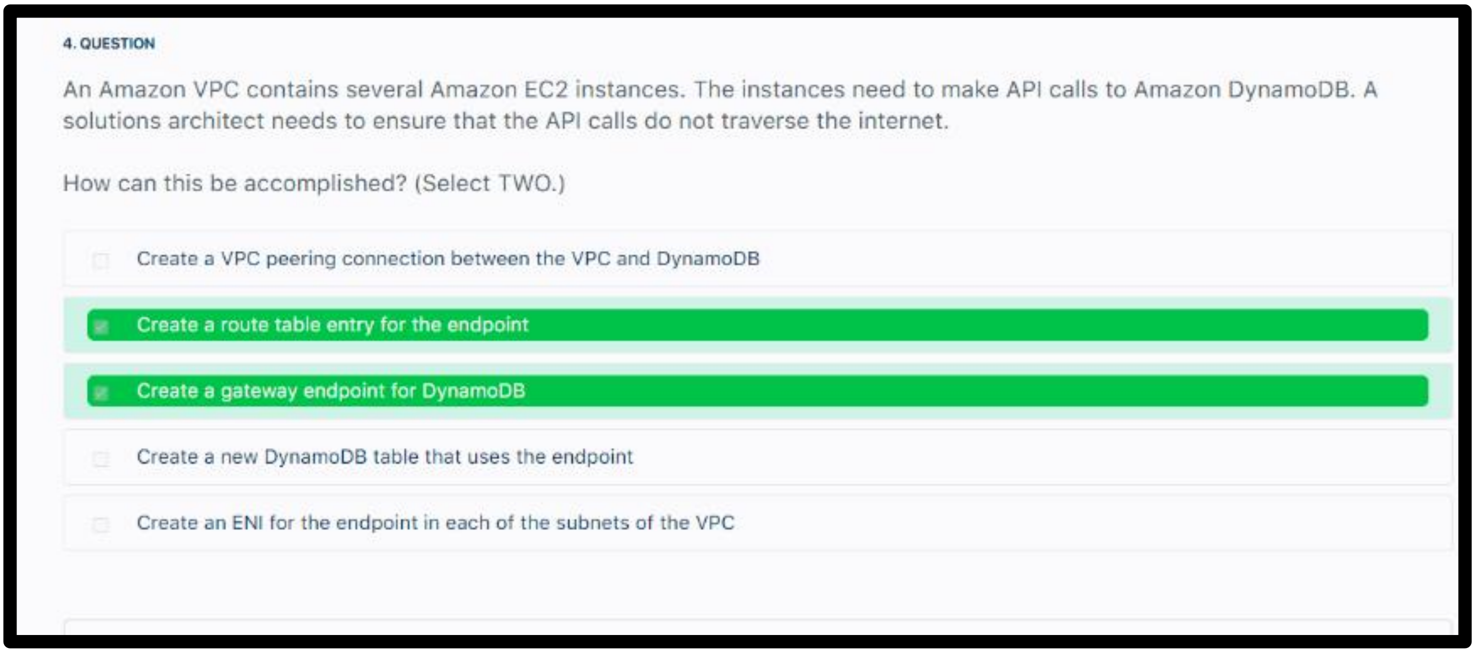
INCORRECT: "Create an interface VPC endpoint in the VPC with an Elastic Network Interface (ENI)" is incorrect. As mentioned above, an interface endpoint is not used for DynamoDB, you must use a gateway endpoint.

INCORRECT: "Create the Amazon DynamoDB table in the VPC" is incorrect. You cannot create a DynamoDB table in a VPC, to connect securely using private addresses you should use a gateway endpoint instead.

INCORRECT: "Create an AWS VPN connection to the Amazon DynamoDB endpoint" is incorrect. You cannot create an AWS VPN connection to the Amazon DynamoDB endpoint.

References:

https://docs.amazonaws.cn/en_us/vpc/latest/userguide/vpc-endpoints-ddb.html



4. QUESTION

An Amazon VPC contains several Amazon EC2 instances. The instances need to make API calls to Amazon DynamoDB. A solutions architect needs to ensure that the API calls do not traverse the internet.

How can this be accomplished? (Select TWO.)

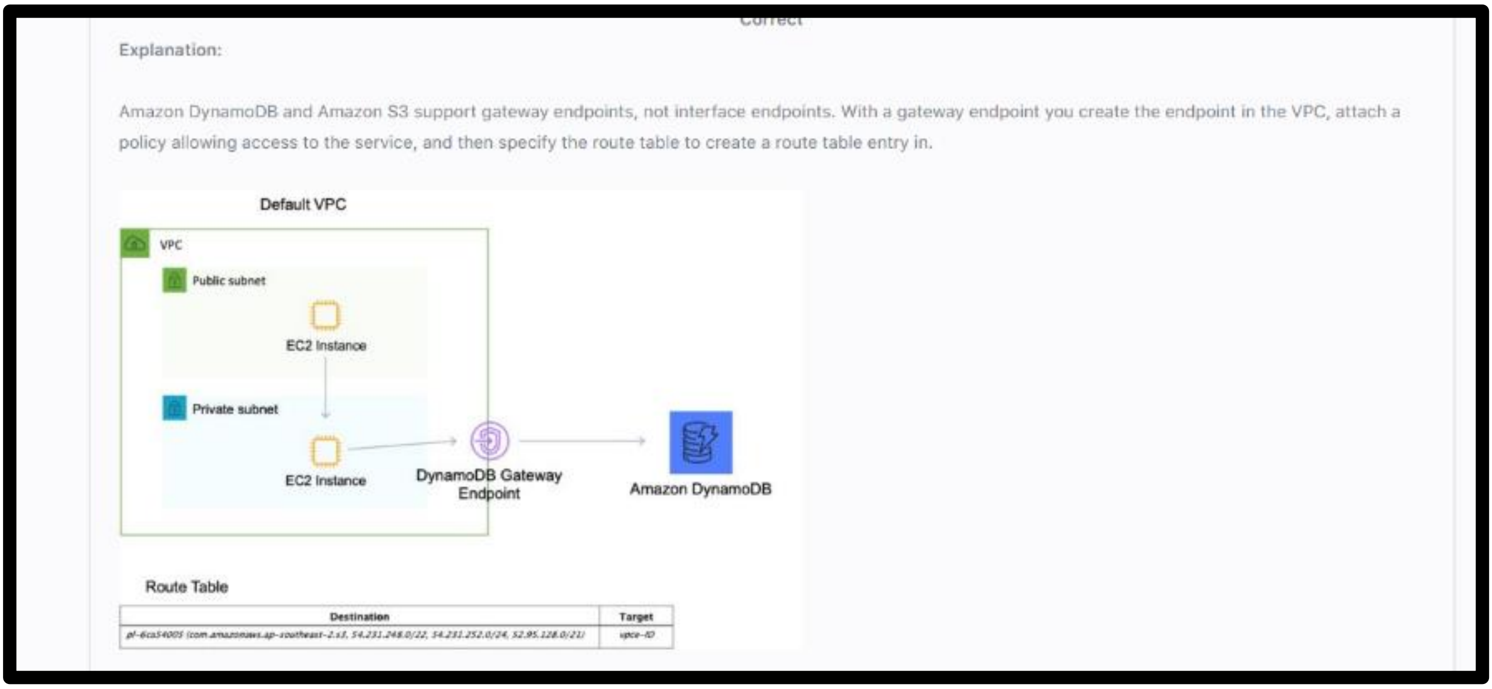
□ Create a VPC peering connection between the VPC and DynamoDB

☑ Create a route table entry for the endpoint

☑ Create a gateway endpoint for DynamoDB

□ Create a new DynamoDB table that uses the endpoint

□ Create an ENI for the endpoint in each of the subnets of the VPC



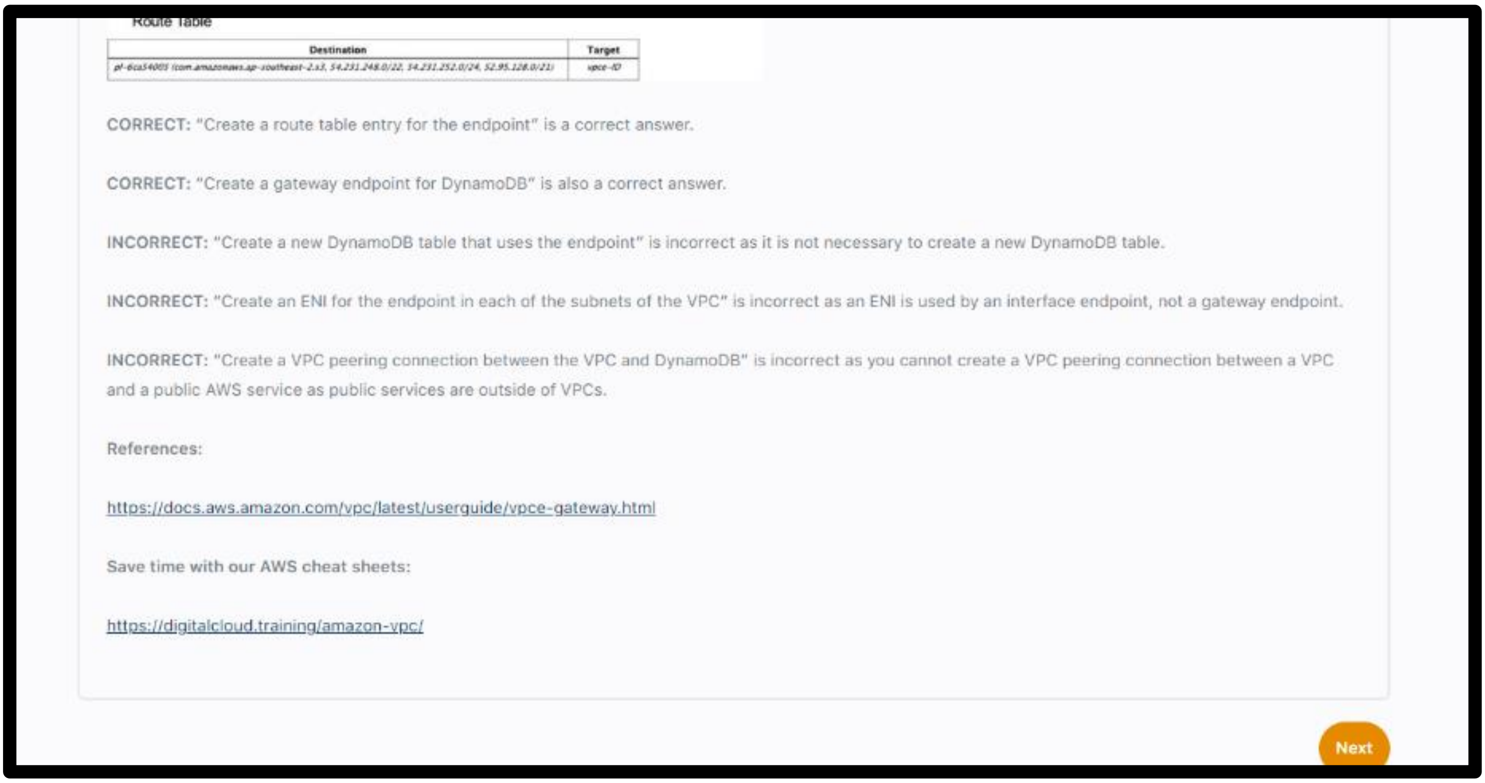
Correct

Explanation:

Amazon DynamoDB and Amazon S3 support gateway endpoints, not interface endpoints. With a gateway endpoint you create the endpoint in the VPC, attach a policy allowing access to the service, and then specify the route table to create a route table entry in.

Route Table

Destination	Target
pl-6ca54005 (com.amazonaws.ap-southeast-2.s3, 54.231.248.0/22, 54.231.252.0/24, 52.95.128.0/21)	vpce-ID



Route Table

Destination	Target
pl-6ca54005 (com.amazonaws.ap-southeast-2.s3, 54.231.248.0/22, 54.231.252.0/24, 52.95.128.0/21)	vpce-ID

CORRECT: "Create a route table entry for the endpoint" is a correct answer.

CORRECT: "Create a gateway endpoint for DynamoDB" is also a correct answer.

INCORRECT: "Create a new DynamoDB table that uses the endpoint" is incorrect as it is not necessary to create a new DynamoDB table.

INCORRECT: "Create an ENI for the endpoint in each of the subnets of the VPC" is incorrect as an ENI is used by an interface endpoint, not a gateway endpoint.

INCORRECT: "Create a VPC peering connection between the VPC and DynamoDB" is incorrect as you cannot create a VPC peering connection between a VPC and a public AWS service as public services are outside of VPCs.

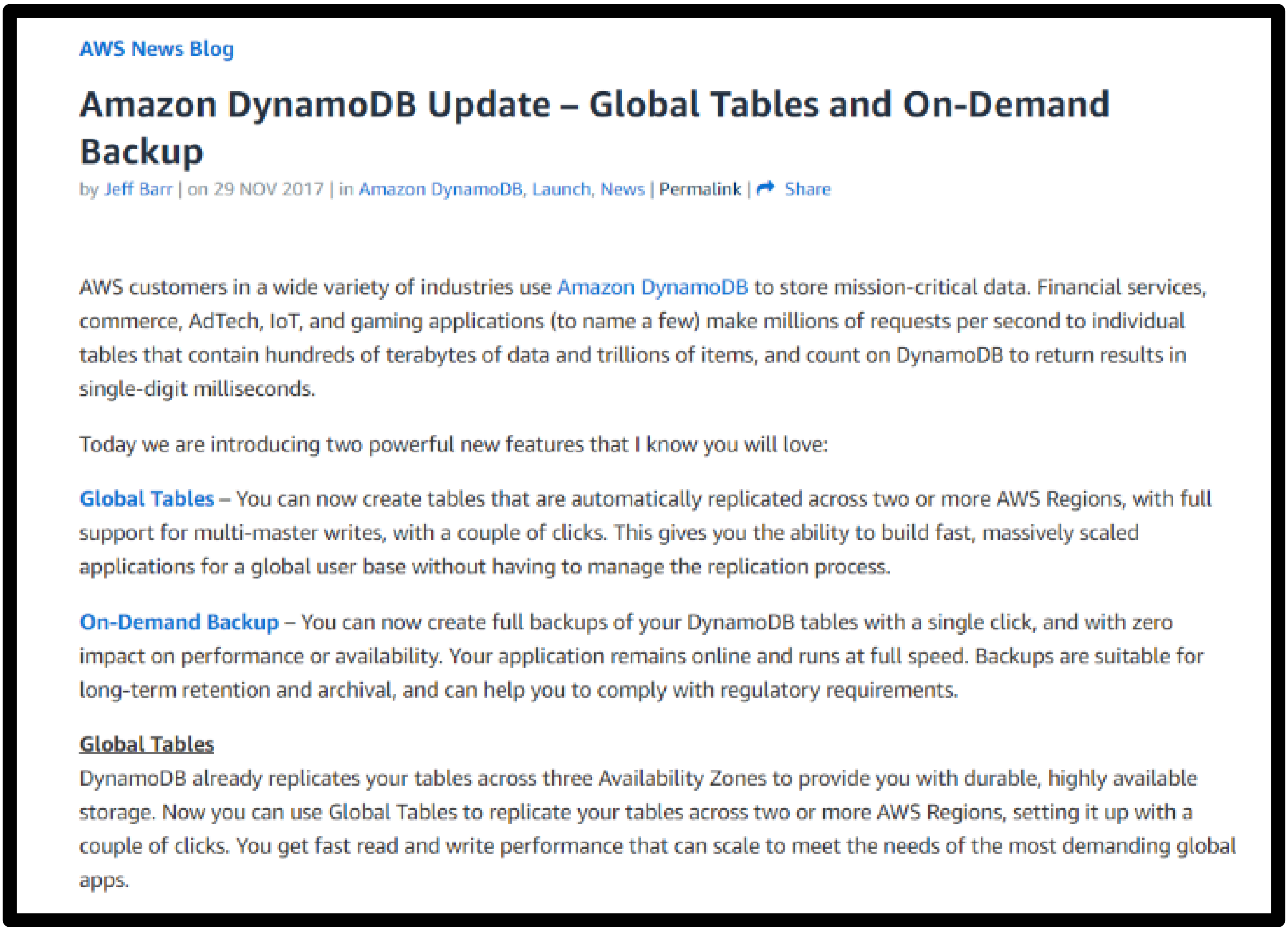
References:

https://docs.aws.amazon.com/vpc/latest/userguide/vpce-gateway.html

Save time with our AWS cheat sheets:

https://digitalcloud.training/amazon-vpc/

Introducing
GLOBAL TABLES



AWS News Blog

Amazon DynamoDB Update – Global Tables and On-Demand Backup

by Jeff Barr | on 29 NOV 2017 | in Amazon DynamoDB, Launch, News | Permalink | Share

AWS customers in a wide variety of industries use Amazon DynamoDB to store mission-critical data. Financial services, commerce, AdTech, IoT, and gaming applications (to name a few) make millions of requests per second to individual tables that contain hundreds of terabytes of data and trillions of items, and count on DynamoDB to return results in single-digit milliseconds.

Today we are introducing two powerful new features that I know you will love:

Global Tables – You can now create tables that are automatically replicated across two or more AWS Regions, with full support for multi-master writes, with a couple of clicks. This gives you the ability to build fast, massively scaled applications for a global user base without having to manage the replication process.

On-Demand Backup – You can now create full backups of your DynamoDB tables with a single click, and with zero impact on performance or availability. Your application remains online and runs at full speed. Backups are suitable for long-term retention and archival, and can help you to comply with regulatory requirements.

Global Tables

DynamoDB already replicates your tables across three Availability Zones to provide you with durable, highly available storage. Now you can use Global Tables to replicate your tables across two or more AWS Regions, setting it up with a couple of clicks. You get fast read and write performance that can scale to meet the needs of the most demanding global apps.

On demand mode

Provisioned mode

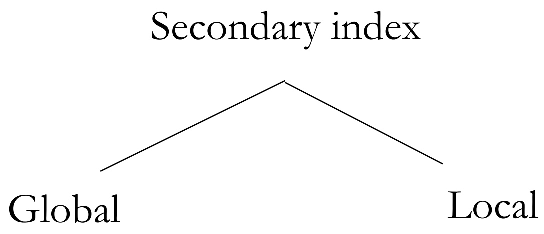
You can switch between provisioned and on-demand mode only once every 24 hours.

What on earth are “read capacity units”?

What on earth is a “secondary index”?

What is the difference between a SCAN and QUERY?

What on earth is a “secondary index”?



What on earth is "Global Tables"?

This is a feature of DynamoDB that allows you to replicate your table across multiple Regions. Clearly, this is somewhat similar to the "Multi-AZ" feature within RDS. But note how this is at the *Region* level, not the AZ level.

We're told:

> A global table is a collection of replica tables, and a global table can have only one replica table per region. Whenever you write an item to a replica table, it's replicated to replica tables in other regions.
>
> Piper and Clinton 2021, p158

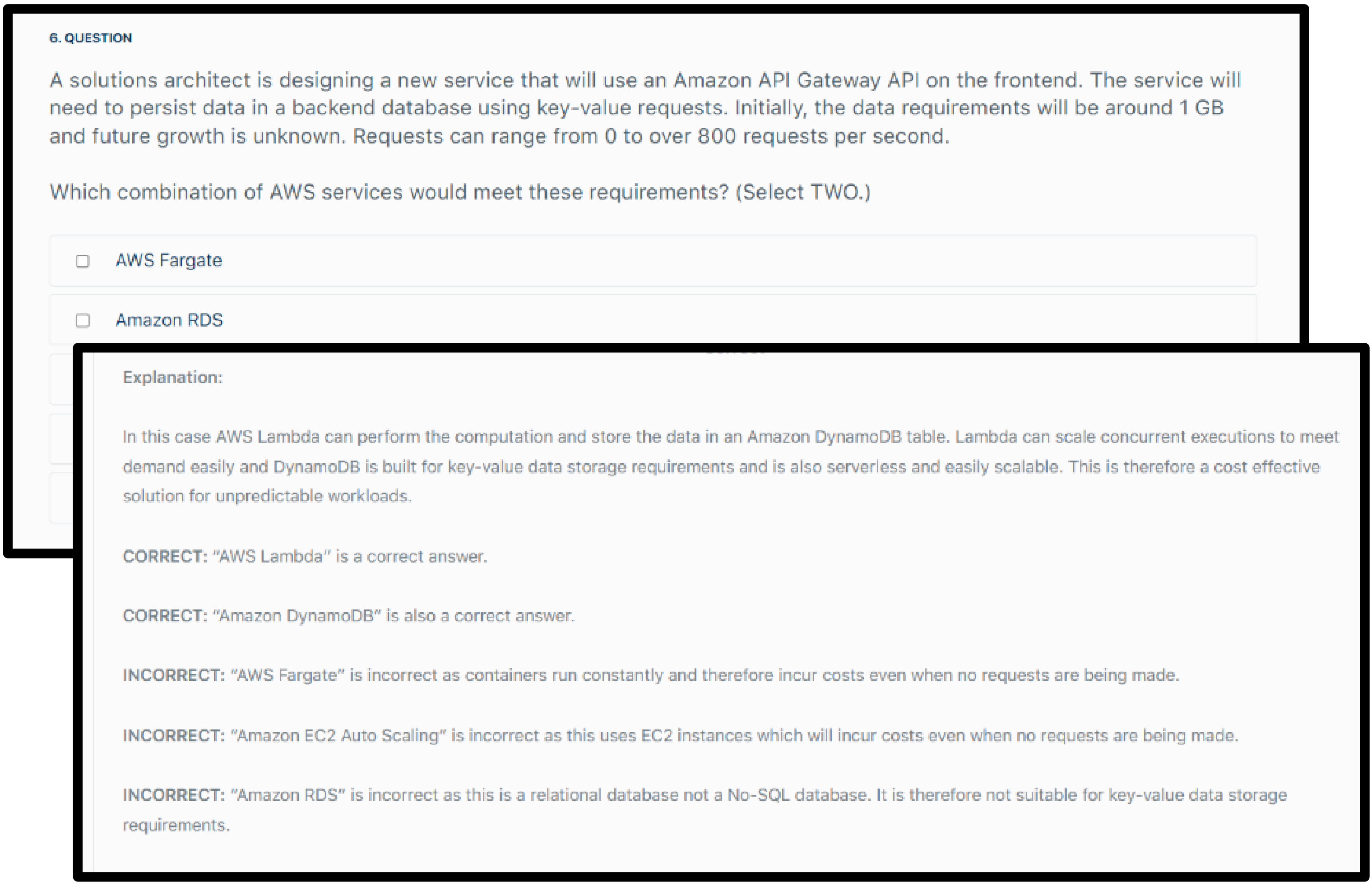
Is “Global Tables” designed to optimize performance, or help with failover?

We talk about “replica tables”, and usually when we talk of “replicas”, this is about fully functioning entities that improve performance. If these tables were merely on standby, ready to be failed over to, we would call them “standbys” or something. So, I think we can be certain that this “global tables” feature helps with performance.

However, I do think it brings a benefit in terms of failover as well. Similar to the Amazon Aurora feature known as “multi-master”, it may be that “global tables” removes the need for any failover to occur. We have a presence in multiple regions, so if a table houses in one AZ goes down, this is not a problem. Notice how Piper and Clinton (2021) introduce this feature by talking about *availability* (not, say, performance). They write:

> To improve **availability**, you can use global tables to replicate a table across multiple regions.
>
> Piper and Clinton, p. 158

https://qconsf.com/presentation/oct2022/amazon-dynamodb-evolution-hyper-scale-cloud-database-service



6. QUESTION

A solutions architect is designing a new service that will use an Amazon API Gateway API on the frontend. The service will need to persist data in a backend database using key-value requests. Initially, the data requirements will be around 1 GB and future growth is unknown. Requests can range from 0 to over 800 requests per second.

Which combination of AWS services would meet these requirements? (Select TWO.)

□ AWS Fargate

□ Amazon RDS

Explanation:

In this case AWS Lambda can perform the computation and store the data in an Amazon DynamoDB table. Lambda can scale concurrent executions to meet demand easily and DynamoDB is built for key-value data storage requirements and is also serverless and easily scalable. This is therefore a cost effective solution for unpredictable workloads.

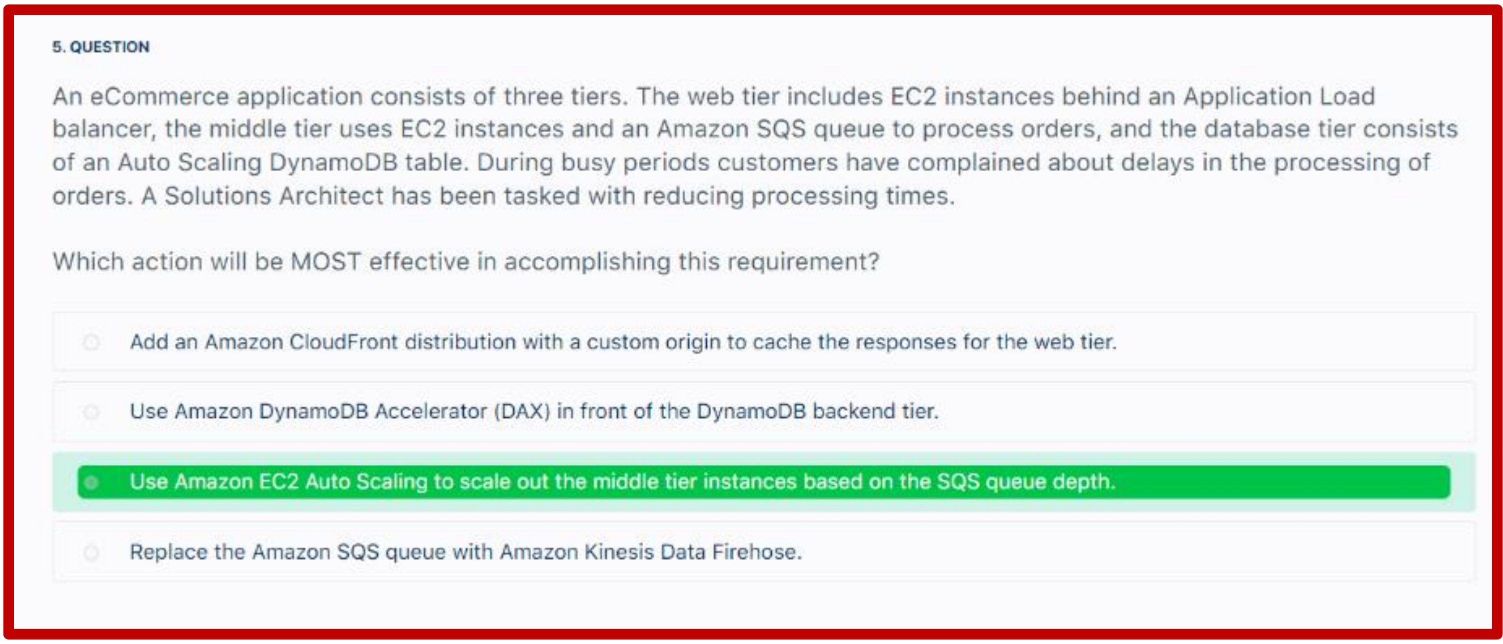
CORRECT: "AWS Lambda" is a correct answer.

CORRECT: "Amazon DynamoDB" is also a correct answer.

INCORRECT: "AWS Fargate" is incorrect as containers run constantly and therefore incur costs even when no requests are being made.

INCORRECT: "Amazon EC2 Auto Scaling" is incorrect as this uses EC2 instances which will incur costs even when no requests are being made.

INCORRECT: "Amazon RDS" is incorrect as this is a relational database not a No-SQL database. It is therefore not suitable for key-value data storage requirements.



5. QUESTION

An eCommerce application consists of three tiers. The web tier includes EC2 instances behind an Application Load balancer, the middle tier uses EC2 instances and an Amazon SQS queue to process orders, and the database tier consists of an Auto Scaling DynamoDB table. During busy periods customers have complained about delays in the processing of orders. A Solutions Architect has been tasked with reducing processing times.

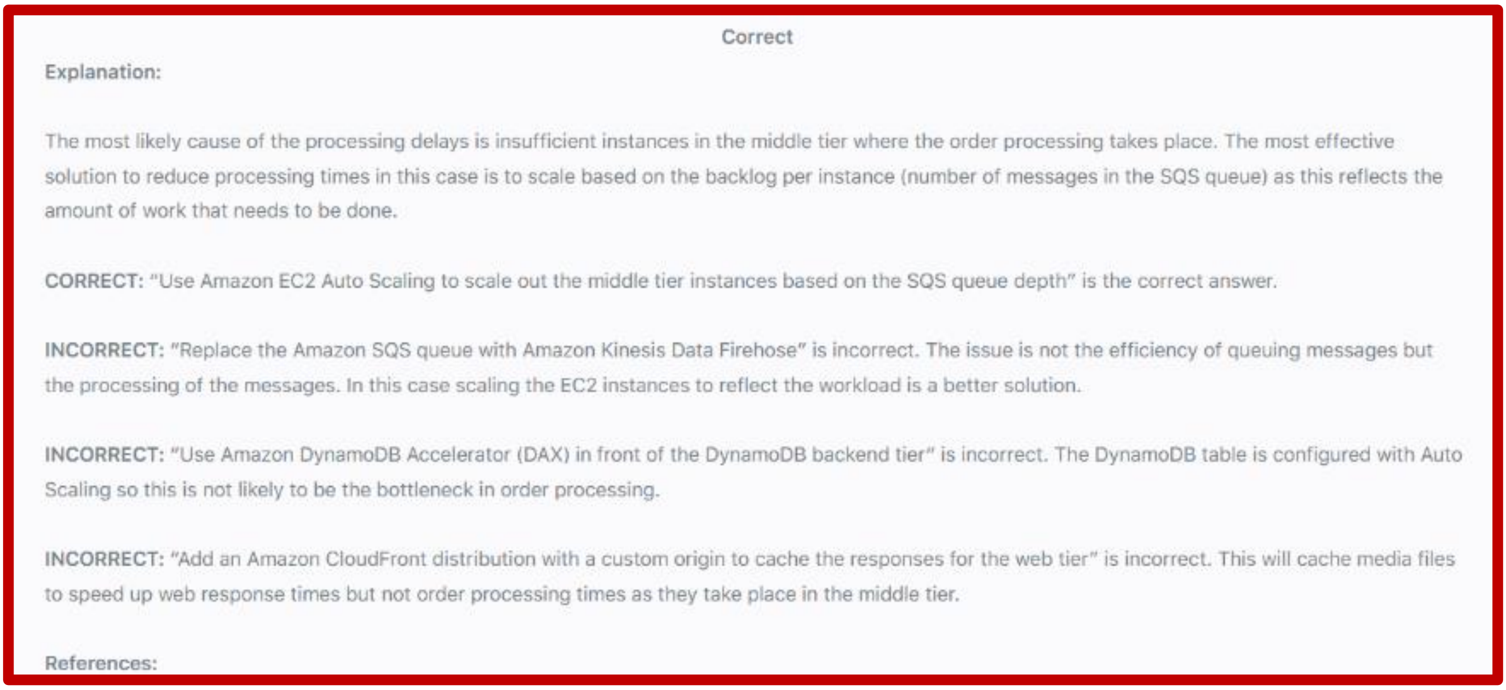
Which action will be MOST effective in accomplishing this requirement?

○ Add an Amazon CloudFront distribution with a custom origin to cache the responses for the web tier.

○ Use Amazon DynamoDB Accelerator (DAX) in front of the DynamoDB backend tier.

● Use Amazon EC2 Auto Scaling to scale out the middle tier instances based on the SQS queue depth.

○ Replace the Amazon SQS queue with Amazon Kinesis Data Firehose.



Correct

Explanation:

The most likely cause of the processing delays is insufficient instances in the middle tier where the order processing takes place. The most effective solution to reduce processing times in this case is to scale based on the backlog per instance (number of messages in the SQS queue) as this reflects the amount of work that needs to be done.

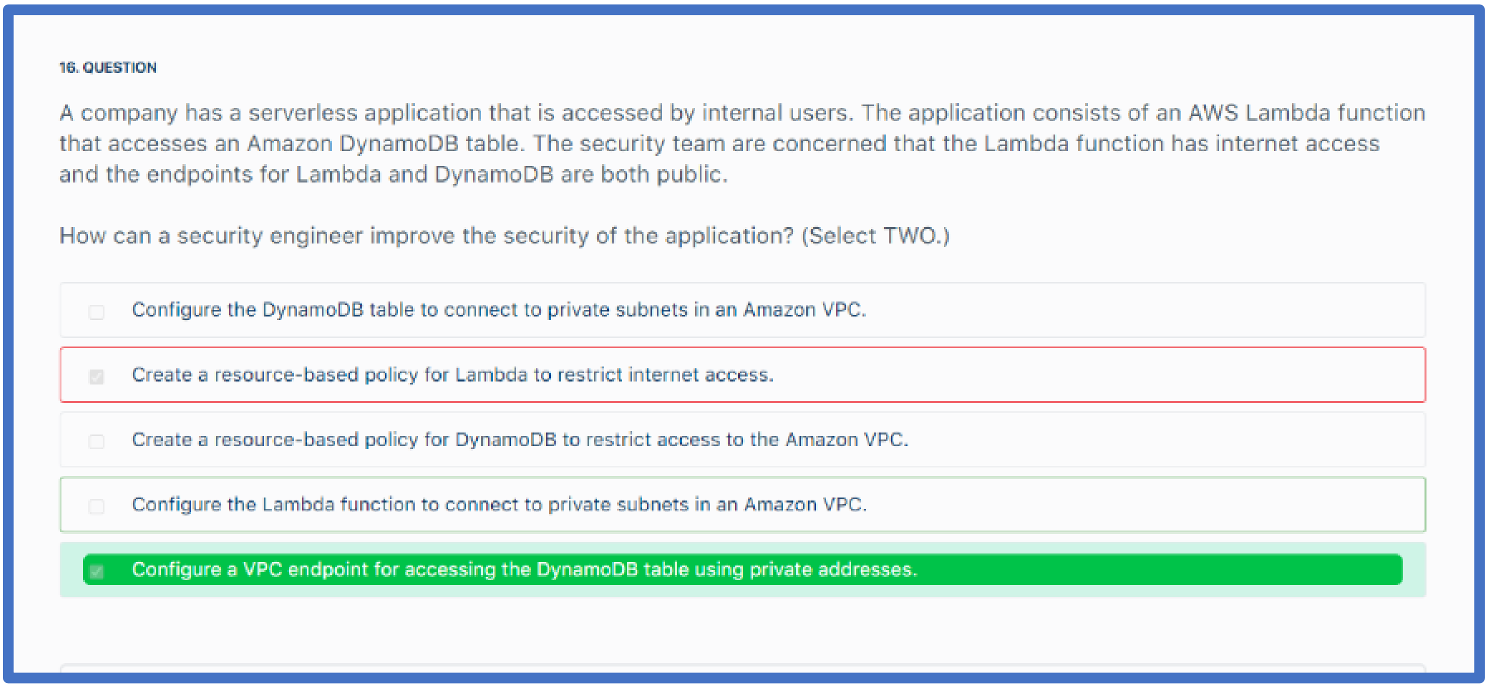
CORRECT: "Use Amazon EC2 Auto Scaling to scale out the middle tier instances based on the SQS queue depth" is the correct answer.

INCORRECT: "Replace the Amazon SQS queue with Amazon Kinesis Data Firehose" is incorrect. The issue is not the efficiency of queuing messages but the processing of the messages. In this case scaling the EC2 instances to reflect the workload is a better solution.

INCORRECT: "Use Amazon DynamoDB Accelerator (DAX) in front of the DynamoDB backend tier" is incorrect. The DynamoDB table is configured with Auto Scaling so this is not likely to be the bottleneck in order processing.

INCORRECT: "Add an Amazon CloudFront distribution with a custom origin to cache the responses for the web tier" is incorrect. This will cache media files to speed up web response times but not order processing times as they take place in the middle tier.

References:

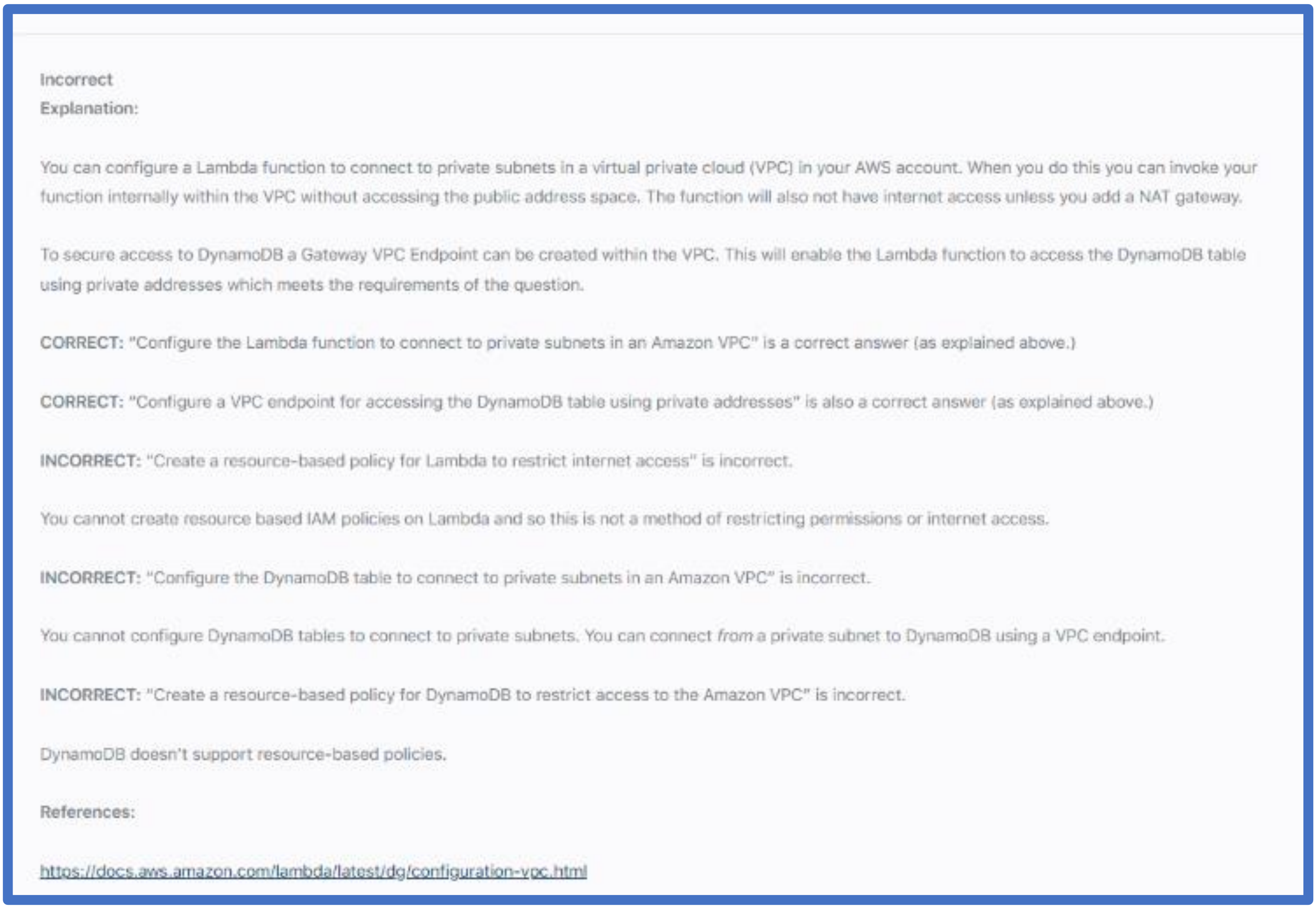


16. QUESTION

A company has a serverless application that is accessed by internal users. The application consists of an AWS Lambda function that accesses an Amazon DynamoDB table. The security team are concerned that the Lambda function has internet access and the endpoints for Lambda and DynamoDB are both public.

How can a security engineer improve the security of the application? (Select TWO.)

- ☐ Configure the DynamoDB table to connect to private subnets in an Amazon VPC.
- ☑ Create a resource-based policy for Lambda to restrict internet access.
- ☐ Create a resource-based policy for DynamoDB to restrict access to the Amazon VPC.
- ☐ Configure the Lambda function to connect to private subnets in an Amazon VPC.
- ☑ Configure a VPC endpoint for accessing the DynamoDB table using private addresses.



Incorrect
Explanation:

You can configure a Lambda function to connect to private subnets in a virtual private cloud (VPC) in your AWS account. When you do this you can invoke your function internally within the VPC without accessing the public address space. The function will also not have internet access unless you add a NAT gateway.

To secure access to DynamoDB a Gateway VPC Endpoint can be created within the VPC. This will enable the Lambda function to access the DynamoDB table using private addresses which meets the requirements of the question.

CORRECT: "Configure the Lambda function to connect to private subnets in an Amazon VPC" is a correct answer (as explained above.)

CORRECT: "Configure a VPC endpoint for accessing the DynamoDB table using private addresses" is also a correct answer (as explained above.)

INCORRECT: "Create a resource-based policy for Lambda to restrict internet access" is incorrect.

You cannot create resource based IAM policies on Lambda and so this is not a method of restricting permissions or internet access.

INCORRECT: "Configure the DynamoDB table to connect to private subnets in an Amazon VPC" is incorrect.

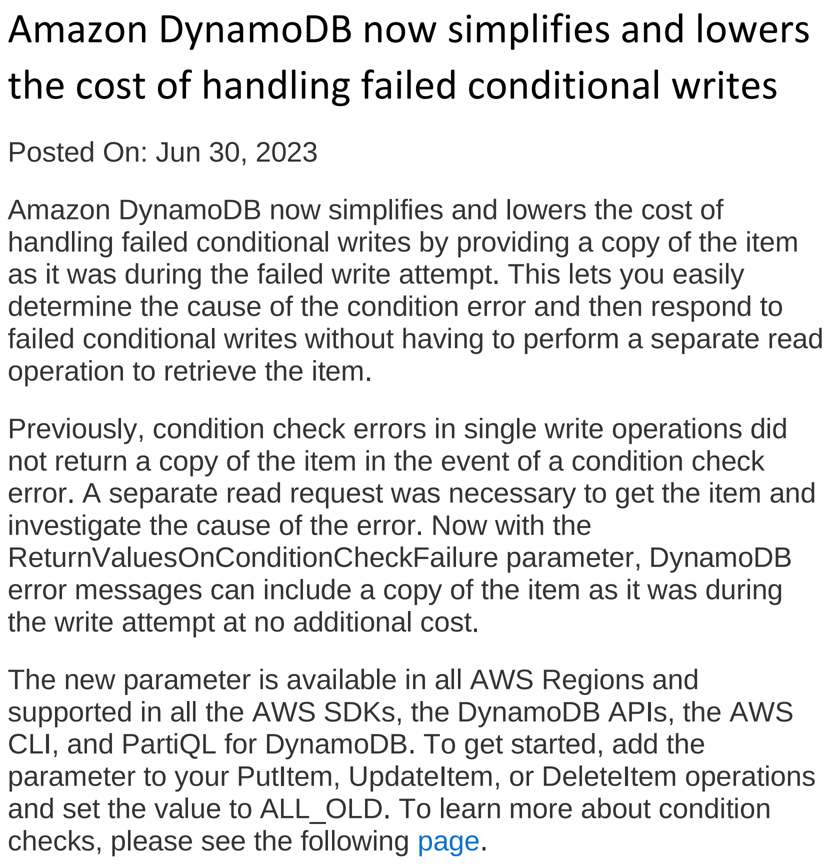
You cannot configure DynamoDB tables to connect to private subnets. You can connect *from* a private subnet to DynamoDB using a VPC endpoint.

INCORRECT: "Create a resource-based policy for DynamoDB to restrict access to the Amazon VPC" is incorrect.

DynamoDB doesn't support resource-based policies.

References:

https://docs.aws.amazon.com/lambda/latest/dg/configuration-vpc.html



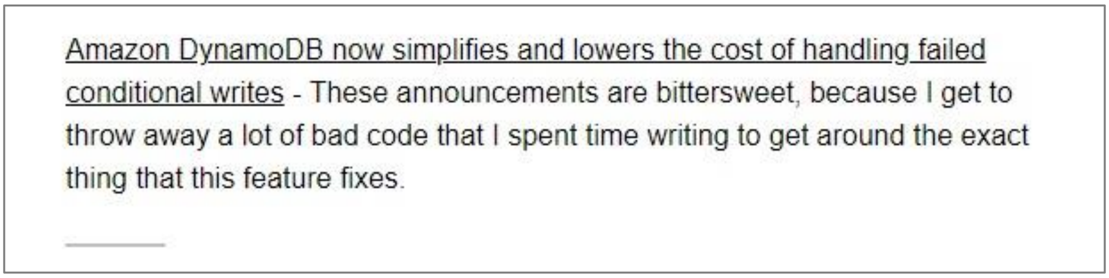
Amazon DynamoDB now simplifies and lowers the cost of handling failed conditional writes

Posted On: Jun 30, 2023

Amazon DynamoDB now simplifies and lowers the cost of handling failed conditional writes by providing a copy of the item as it was during the failed write attempt. This lets you easily determine the cause of the condition error and then respond to failed conditional writes without having to perform a separate read operation to retrieve the item.

Previously, condition check errors in single write operations did not return a copy of the item in the event of a condition check error. A separate read request was necessary to get the item and investigate the cause of the error. Now with the ReturnValuesOnConditionCheckFailure parameter, DynamoDB error messages can include a copy of the item as it was during the write attempt at no additional cost.

The new parameter is available in all AWS Regions and supported in all the AWS SDKs, the DynamoDB APIs, the AWS CLI, and PartiQL for DynamoDB. To get started, add the parameter to your PutItem, UpdateItem, or DeleteItem operations and set the value to ALL_OLD. To learn more about condition checks, please see the following page.



Amazon DynamoDB now simplifies and lowers the cost of handling failed conditional writes - These announcements are bittersweet, because I get to throw away a lot of bad code that I spent time writing to get around the exact thing that this feature fixes.

Email from Corey Quinn on 10th July 2023

Let's understand this announcement

What is a "conditional write"? A conditional write is a write which only occurs if certain conditions are met. The User Guide for DynamoDB tells us:

> To manipulate data in an Amazon DynamoDB table, you use the `PutItem`, `UpdateItem`, and `DeleteItem` operations. (You can also use `BatchWriteItem` to perform

> multiple `PutItem` or `DeleteItem` operations in a single call.)
>
> For these data manipulation operations, you can specify a *condition expression* to determine which items should be modified. If the condition expression evaluates to true, the operation succeeds; otherwise, the operation fails.

So, there are a number of API calls which are used to manipulate data in a DynamoDB table. These are PutItem, UpdateItem, and DeleteItem. You can make it so that these operations only succeed if certain conditions are met.

This articulation of certain conditions is known as a condition *expression*. So, what are some examples of conditions which we might install?

Before we look at an example, we need to understand something about the PutItem operation. This operation tries to reduce replicas. It attempts to preserve uniqueness. It tries to keep things clean (like a golfing clubhouse – putting). Thus:

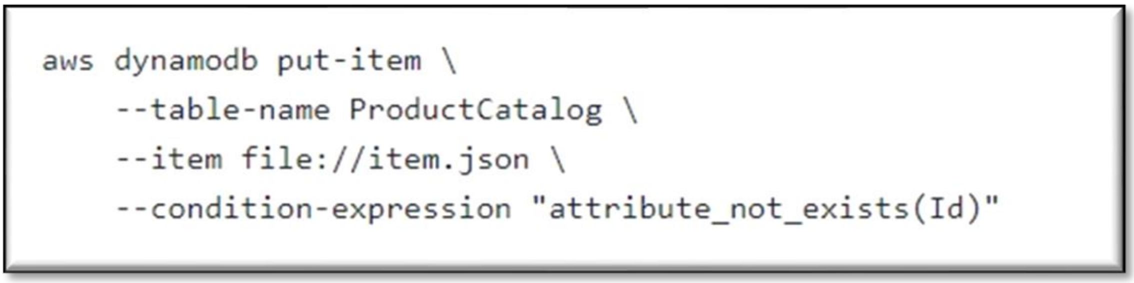
> The `PutItem` operation overwrites an item with the same primary key (if it exists).

This is radical behaviour. It means that if two items differ in terms of all of their attributes, save for the primary key, then they will be treated as identical. Clearly, you might want to avoid this. We can achieve this by specifying a condition which must be met if the operation is to succeed:

> If you want to avoid this, use a condition expression.
>
> This allows the write to proceed only if the item in question does not already have the same primary key.

Let's bring that point out. The write will proceed only if the primary key *differs* ('does not already have the same primary key'). If the primary key is the same, the write will not occur. The item will remain as it was.

Without the condition installed, if the item to be written had the same primary key as another item, the latter would be overwritten. *With* the condition installed, the overwrite would not occur.



```
aws dynamodb put-item \
    --table-name ProductCatalog \
    --item file://item.json \
    --condition-expression "attribute_not_exists(Id)"
```

Now let's reconsider the announcement. Presumably, a failed conditional write is a write which does not succeed. It does not succeed because one of the conditions was not met.

Thus, it would be helpful to know which condition was not met. And this is what we are getting with this new feature – knowledge of what condition is not met. Previously, you could only *infer* why the write failed by doing a separate read request:

> Previously, condition check errors in single write operations did not return a copy of the item in the event of a condition check error.
>
> A separate read request was necessary to get the item and investigate the cause of the error.

Introducing the AWS .NET Distributed Cache Provider for DynamoDB

Posted On: Jul 7, 2023

We are happy to announce the preview release of the AWS .NET Distributed Cache Provider for DynamoDB. This library enables Amazon DynamoDB to be used as the storage for ASP.NET Core's distributed cache framework.

A cache can improve the performance of an application; an external cache allows the data to be shared across application servers and helps to avoid cache misses when the application is restarted or redeployed. Customers can now use DynamoDB tables to cache their session state using ASP.NET Core's distributed cache framework.

Get started by installing the AWS.DistributedCacheProvider package from NuGet.org. It is open sourced, and we welcome community contributions! To learn more, go to our blog post, visit our GitHub page and our developer documentation.

Let's understand this announcement

First of all, what is .NET? This is pronounced "dot net". This is a software framework developed by Microsoft. The project is mainly developed by Microsoft employees. The term *.NET Framework* is no longer used, and now it is simply known as .NET.

The .NET platform fully supports C# and F# as well as Visual Basic.

There are many versions of .NET. For example, .NET 7 was released in August 2022. We saw .NET 5 released in November 2020.

What is ASP.net? This is a framework for developing web applications. It was first released in January 2002. ASP.net was initially released as part of version 1.0 of the .NET Framework.

Wikipedia tells us that:

> **ASP.NET** is an open-source,[2] server-side web-application framework designed for web development to produce dynamic web pages.
>
> It was developed by Microsoft to allow programmers to build dynamic web sites, applications and services. The name stands for Active Server Pages Network Enabled Technologies.

So, now that we understand what .NET and ASP.net is, we should be in a better position to understand this announcement. This announcement is for those people who build web applications using ASP.net, and need some sort of database to store the data of the application. The database which they use is DynamoDB:

> The AWS .NET Distributed Cache Provider provides an implementation of the ASP.NET Core interface IDistributedCache backed by Amazon DynamoDB.

This above extract is from the GitHub repository. Let's break it down. A cache is simply a store of data, kept closer to the user, and usually containing a subset of the data in the database. The subset consists of those items which are more likely to be requested.

What is a *distributed* cache? Wikipedia tells us that:

> In computing, a **distributed cache** is an extension of the traditional concept of cache used in a single locale. A distributed cache may span multiple servers so that it can grow in size and in transactional capacity. It is mainly used to store application data residing in database and web session data.

So, it seems that if you use .NET on AWS, then there is a way to achieve a distributed cache. Indeed, Chitikesi explains that:

> Caching works by storing data that is accessed frequently but changes infrequently – e.g., static data from an API – in a high-read, performance data store known as the cache. ASP.NET Core provides several ways to incorporate caching into your applications.
>
> [Chitikesi 2022]

ASP.NET Core provides multiple ways to build a cache. Presumably, not all these ways involve a distributed cache. Nevertheless, if you do want to use a distributed cache, then this is possible:

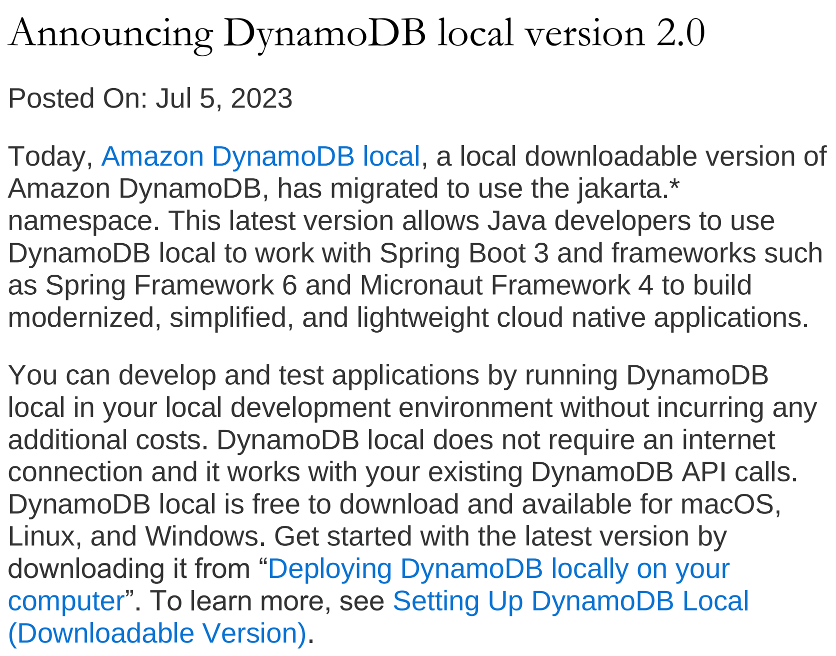
> The recommended approach to cache data in a web server farm for ASP.NET Core web applications is to use distributed caching.
>
> A distributed cache is maintained as an external service separate from your application servers. It enables independent scaling of your application and cache environments, improves fault tolerance, and ensures availability of cached data across deployments and server restarts in load-balanced environments.
>
> [Chitikesi 2022]

Perhaps trying to be reminiscent of Apple branding, or the movie iRobot, the interface is known as IDistibutedCache.

> Distributed caches in ASP.NET Core implement the IDistributedCache interface.
>
> Two implementations are provided with .NET Core: the SQL Server Distributed Cache and the Redis Distributed Cache.

> In the following sections, we will step through implementing a SQL Server Distributed Cache backed by an Amazon RDS for SQL Server database and a Redis Distributed Cache backed by Amazon ElastiCache for Redis.

Here, Chitikesi tells us that we might use Redis to implement a distributed cache. Alternatively, we might use SQL Server to implement a distributed cache. SQL Server is a Microsoft product.

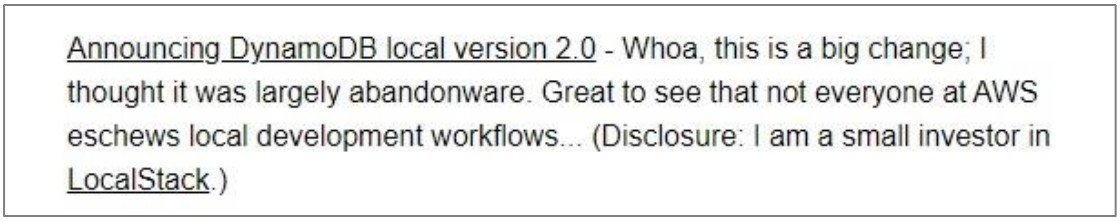


Announcing DynamoDB local version 2.0

Posted On: Jul 5, 2023

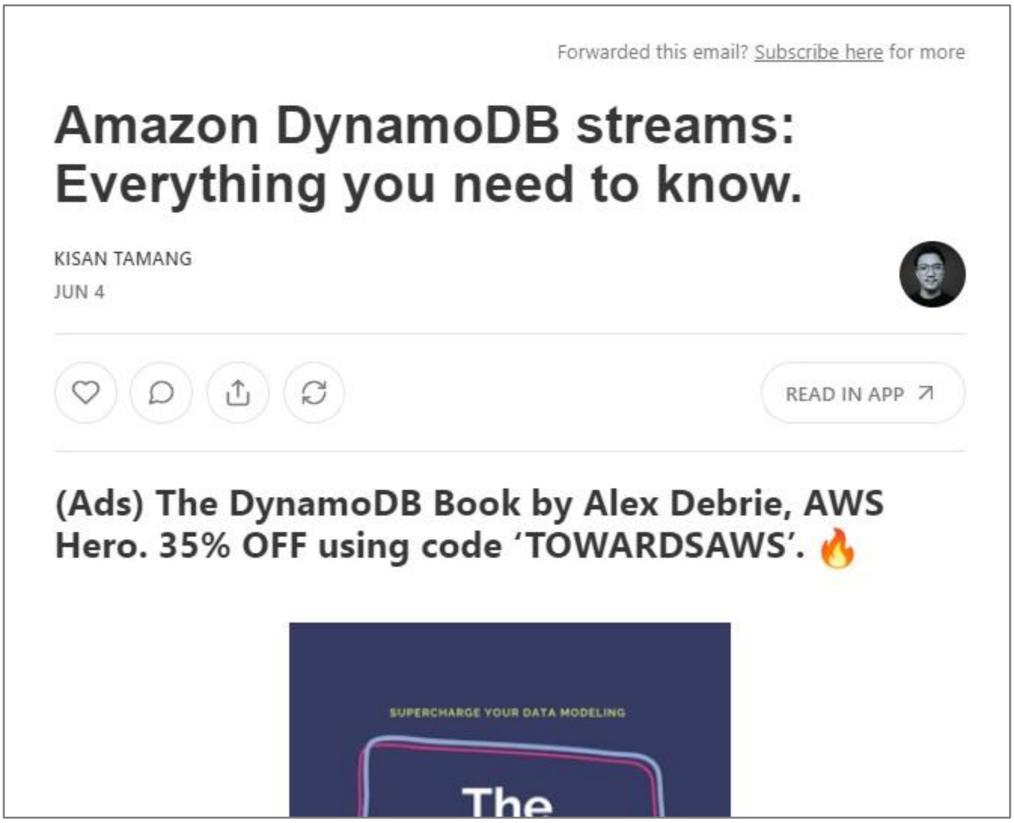
Today, Amazon DynamoDB local, a local downloadable version of Amazon DynamoDB, has migrated to use the jakarta.* namespace. This latest version allows Java developers to use DynamoDB local to work with Spring Boot 3 and frameworks such as Spring Framework 6 and Micronaut Framework 4 to build modernized, simplified, and lightweight cloud native applications.

You can develop and test applications by running DynamoDB local in your local development environment without incurring any additional costs. DynamoDB local does not require an internet connection and it works with your existing DynamoDB API calls. DynamoDB local is free to download and available for macOS, Linux, and Windows. Get started with the latest version by downloading it from “Deploying DynamoDB locally on your computer”. To learn more, see Setting Up DynamoDB Local (Downloadable Version).



Announcing DynamoDB local version 2.0 - Whoa, this is a big change; I thought it was largely abandonware. Great to see that not everyone at AWS eschews local development workflows... (Disclosure: I am a small investor in LocalStack.)

Email from Corey Quinn on July 10th 2023



Forwarded this email? Subscribe here for more

Amazon DynamoDB streams: Everything you need to know.

KISAN TAMANG
JUN 4

Amazon DynamoDB introduces configurable maximum throughput for On-demand tables

Posted On: May 3, 2024

Amazon DynamoDB on-demand is a serverless, pay-per-request billing option that can serve thousands of requests per second without capacity planning. Previously, the on-demand request rate was only limited by the default throughput quota (40K read request units and 40K write request units), which uniformly applied to all tables within the account, and could not be customized or tailored for diverse workloads and differing requirements. Since on-demand mode scales instantly to accommodate varying traffic patterns, a piece of hastily written or unoptimized code could rapidly scale up and consume resources, making it difficult to keep costs and usage bounded.

Starting today, you can optionally configure maximum read or write (or both) throughput for individual on-demand tables and associated secondary indexes, making it easy to balance costs and performance. Throughput requests in excess of the maximum table throughput will automatically get throttled, but you can easily modify the table-specific maximum throughput at any time based on your application requirements. Customers can use this feature for predictable cost management, protection against accidental surge in consumed resources and excessive use, and safe guarding downstream services with fixed capacities from potential overloading and performance bottlenecks.

On-Demand throughput is available in all AWS Regions.
See Amazon DynamoDB Pricing page for on-demand pricing.
See the Developer Guide to learn more.

Bibliography

I. Official

[AWS 2023]

Amazon DynamoDB now simplifies and lowers the cost of handling failed conditional writes. [Announcement]. June 30th 2023. Available at: <https://aws.amazon.com/about-aws/whats-new/2023/06/amazon-dynamodb-cost-failed-conditional-writes/?utm_source=substack&utm_medium=email>

[AWS 2023b]

https://aws.amazon.com/about-aws/whats-new/2023/07/aws-net-distributed-cache-provider-dynamodb/?utm_source=substack&utm_medium=email

[AWS 2024]

Amazon DynamoDB now supports an AWS FIS action to pause global table replication [Announcement]. Apr 30th 2024. Available at: <https://aws.amazon.com/about-aws/whats-new/2024/04/amazon-dynamodb-fis-action-pause-global-table-replication/>.

[AWS 2024b]

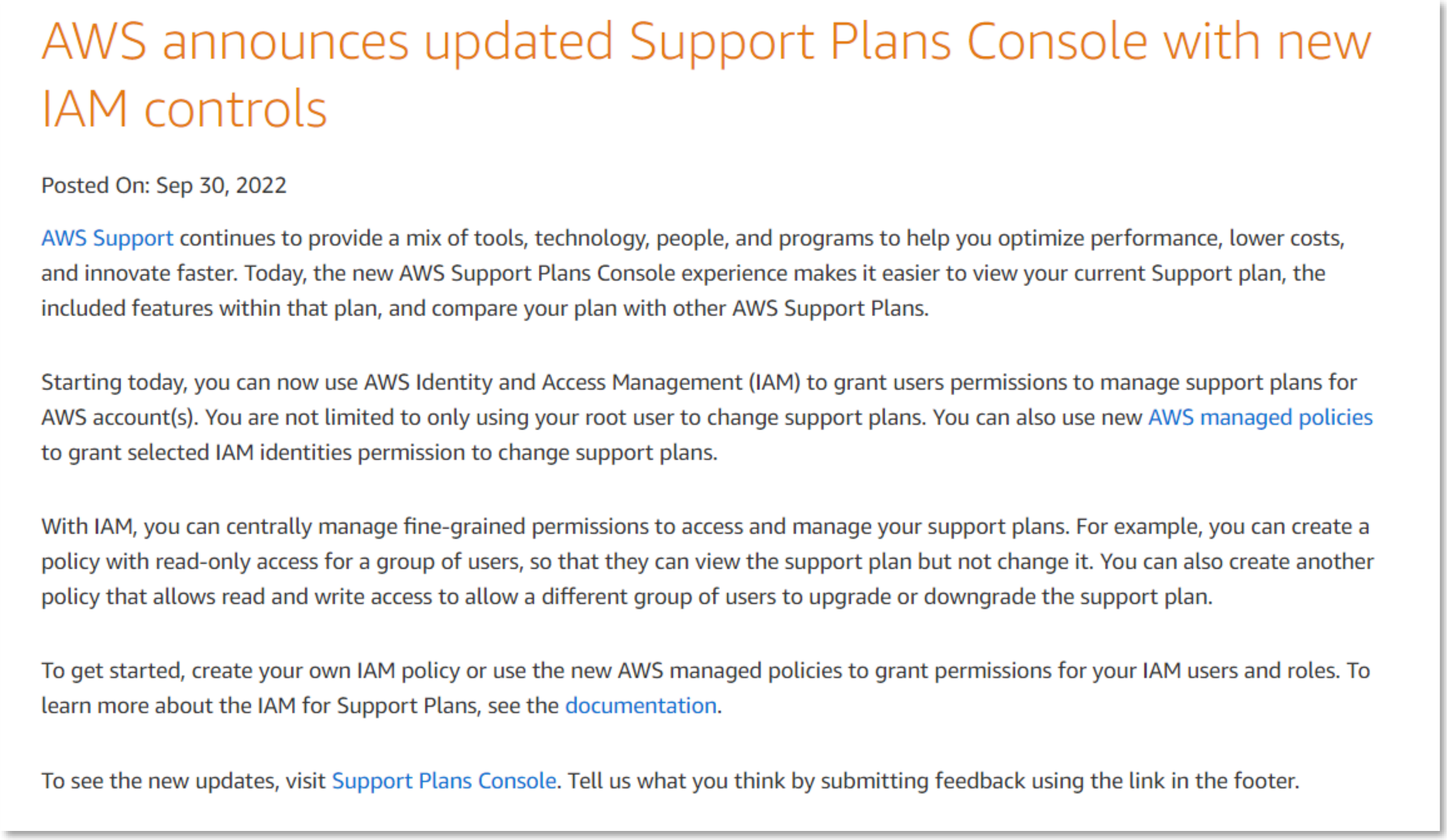
Amazon DynamoDB introduces configurable maximum throughput for On-demand tables [Announcement]. Available at: < https://aws.amazon.com/about-aws/whats-new/2024/05/dynamodb-configurable-maximum-throughput-on-demand-tables/>

II. Unofficial

[Tamang 2024]

Tamang, Kisan (2024). Amazon DynamoDB streams: Everything you need to know. *The Cloud Handbook* [Newsletter]. June 4th 2024.

Trusted Advisor



AWS announces updated Support Plans Console with new IAM controls

Posted On: Sep 30, 2022

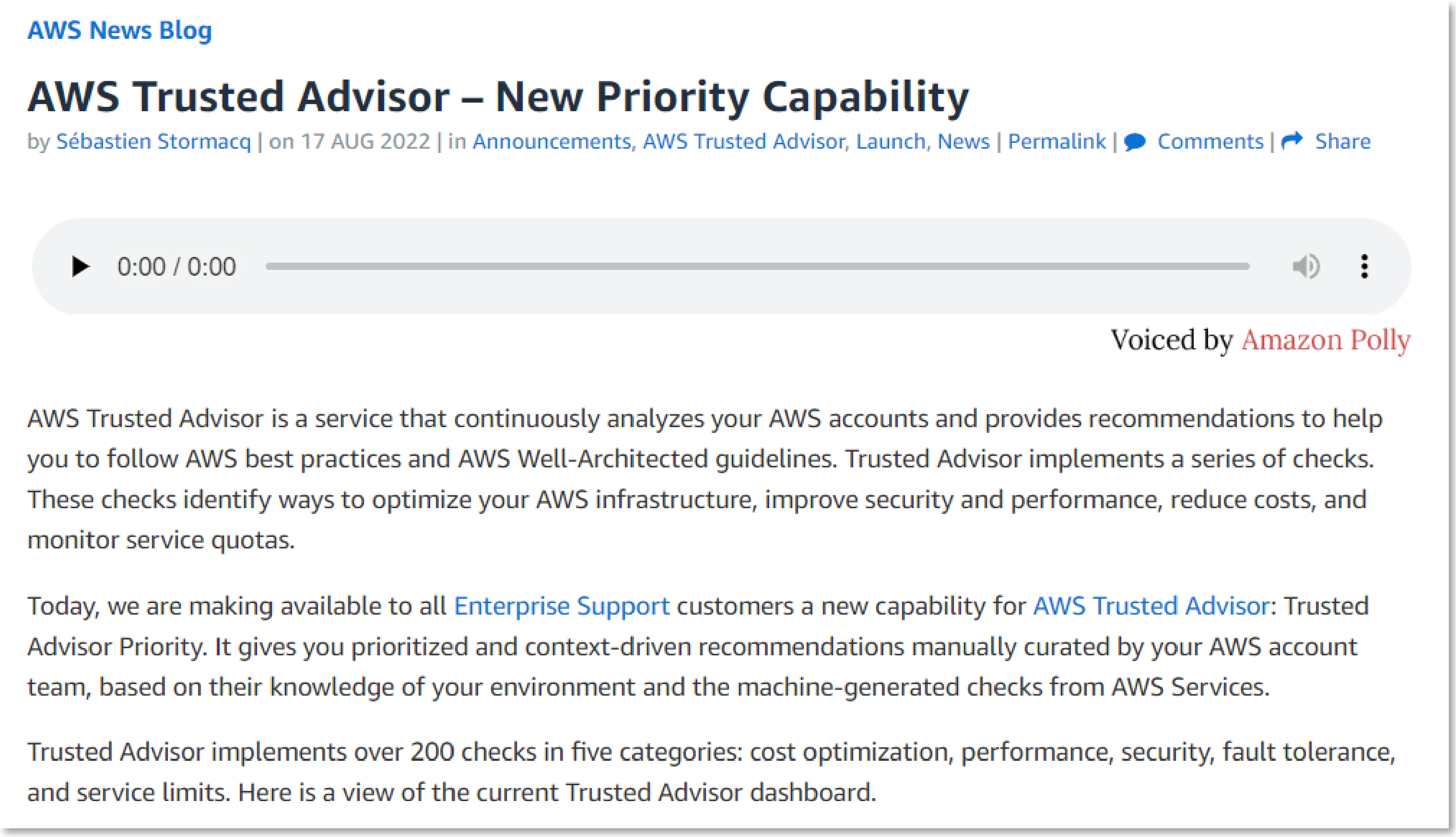
AWS Support continues to provide a mix of tools, technology, people, and programs to help you optimize performance, lower costs, and innovate faster. Today, the new AWS Support Plans Console experience makes it easier to view your current Support plan, the included features within that plan, and compare your plan with other AWS Support Plans.

Starting today, you can now use AWS Identity and Access Management (IAM) to grant users permissions to manage support plans for AWS account(s). You are not limited to only using your root user to change support plans. You can also use new AWS managed policies to grant selected IAM identities permission to change support plans.

With IAM, you can centrally manage fine-grained permissions to access and manage your support plans. For example, you can create a policy with read-only access for a group of users, so that they can view the support plan but not change it. You can also create another policy that allows read and write access to allow a different group of users to upgrade or downgrade the support plan.

To get started, create your own IAM policy or use the new AWS managed policies to grant permissions for your IAM users and roles. To learn more about the IAM for Support Plans, see the documentation.

To see the new updates, visit Support Plans Console. Tell us what you think by submitting feedback using the link in the footer.



AWS News Blog

AWS Trusted Advisor – New Priority Capability

by Sébastien Stormacq | on 17 AUG 2022 | in Announcements, AWS Trusted Advisor, Launch, News | Permalink | Comments | Share

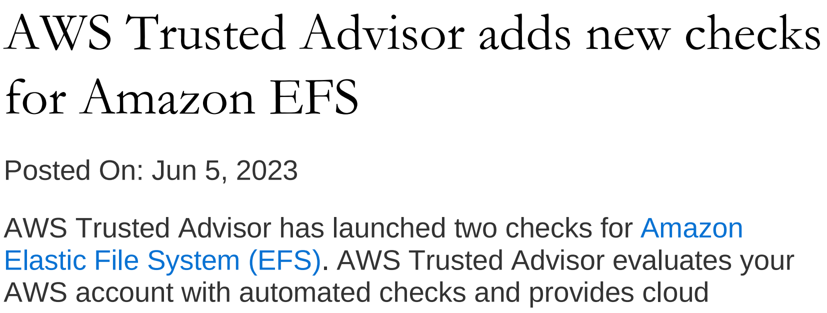
Voiced by Amazon Polly

AWS Trusted Advisor is a service that continuously analyzes your AWS accounts and provides recommendations to help you to follow AWS best practices and AWS Well-Architected guidelines. Trusted Advisor implements a series of checks. These checks identify ways to optimize your AWS infrastructure, improve security and performance, reduce costs, and monitor service quotas.

Today, we are making available to all Enterprise Support customers a new capability for AWS Trusted Advisor: Trusted Advisor Priority. It gives you prioritized and context-driven recommendations manually curated by your AWS account team, based on their knowledge of your environment and the machine-generated checks from AWS Services.

Trusted Advisor implements over 200 checks in five categories: cost optimization, performance, security, fault tolerance, and service limits. Here is a view of the current Trusted Advisor dashboard.

A link to the article shown was provided in Corey Quinn's email on 22nd August 2022



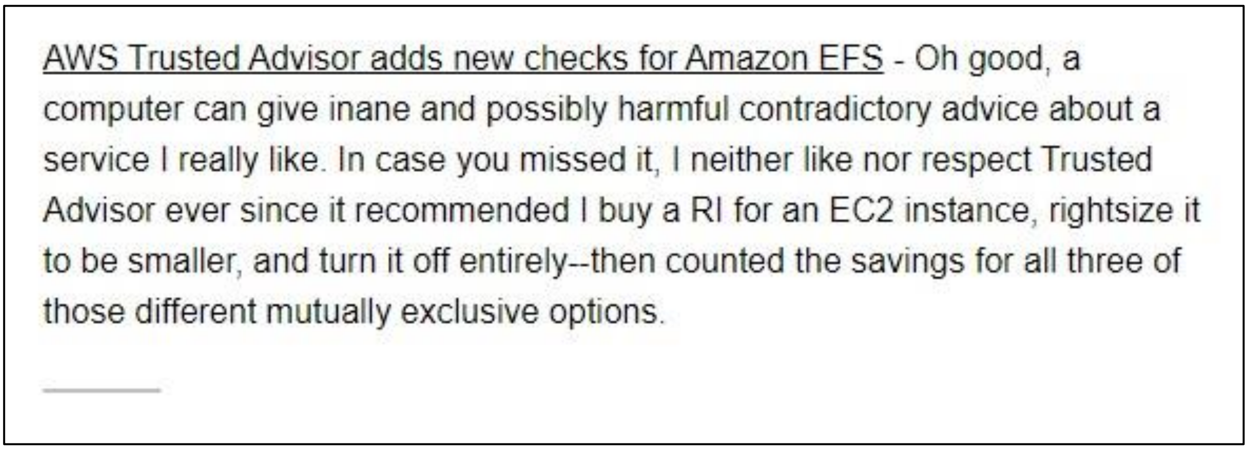
AWS Trusted Advisor adds new checks for Amazon EFS

Posted On: Jun 5, 2023

AWS Trusted Advisor has launched two checks for Amazon Elastic File System (EFS). AWS Trusted Advisor evaluates your AWS account with automated checks and provides cloud optimization recommendations to reduce costs, improve performance, increase security and fault tolerance, and monitor service quotas.

The fault tolerance check for Amazon EFS No Mount Target Redundancy checks if mount targets exist in multiple Availability Zones. The performance check for Amazon EFS Throughput Mode Optimization checks to determine if your Amazon EFS file system is not configured to use Elastic, or Provisioned Throughput mode. The checks are available in all commercial Regions.

AWS Premium Support customers can access the fault tolerance checks from the AWS Trusted Advisor Console, or via the AWS Support API. For more information please visit the AWS Trusted Advisor webpage and the documentation site for a complete list of check references.



AWS Trusted Advisor adds new checks for Amazon EFS - Oh good, a computer can give inane and possibly harmful contradictory advice about a service I really like. In case you missed it, I neither like nor respect Trusted Advisor ever since it recommended I buy a RI for an EC2 instance, rightsize it to be smaller, and turn it off entirely--then counted the savings for all three of those different mutually exclusive options.

Email from Quinn on 12th June 2023

https://aws.amazon.com/blogs/aws/aws-trusted-advisor-new-priority-capability/?ck_subscriber_id=1560524742

TPN

111. ***Phenomenon1*** – the tendency of X to Y.
112. ***Phen2*** – the tendency of X to Y.
113. ***Phen3*** – the tendency of X to Y.
114. ***Phen4*** – the tendency of X to Y.
115. ***Phen5*** – the tendency of X to Y.
116. ***Phen6*** – the tendency of X to Y.
117. ***Phen7*** – the tendency of X to Y.
118. ***Phen8*** – the tendency of X to Y.
119. ***Phen9*** – the tendency of X to Y.
120. ***Phen10*** – the tendency of X to Y.

Glossary

Term1

Description of what term means here.

Term2

Description of what term means here.

Term3

Description of what term means here.

Bibliography

I. Official

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

II. Unofficial

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

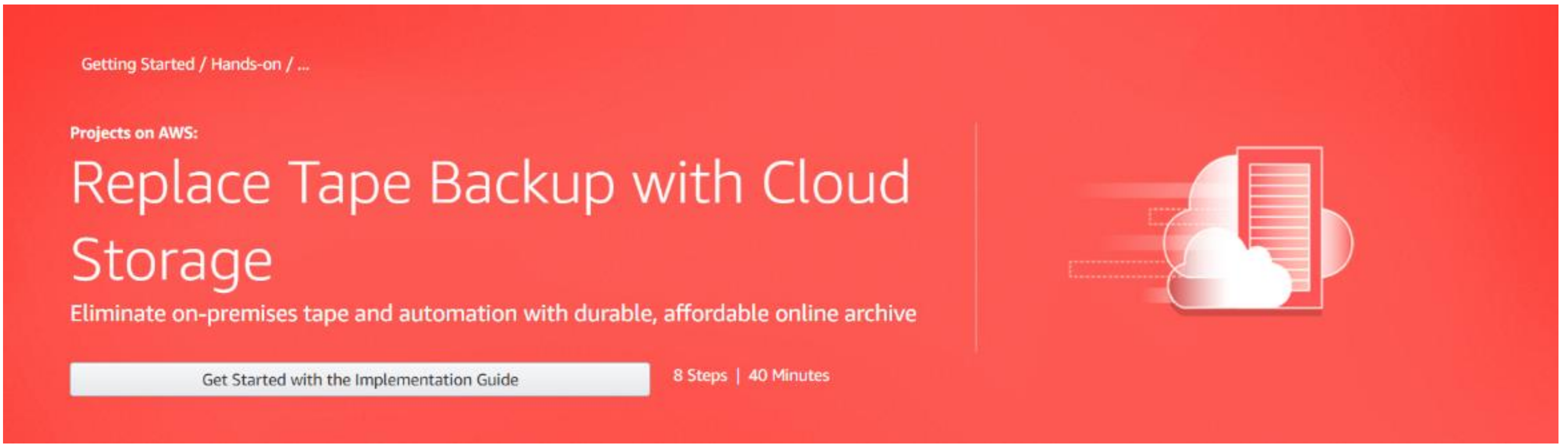
IV. General

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

Storage Gateway

Photograph provided on the AWS website of the AWS Storage Gateway hardware appliance



Getting Started / Hands-on / ...

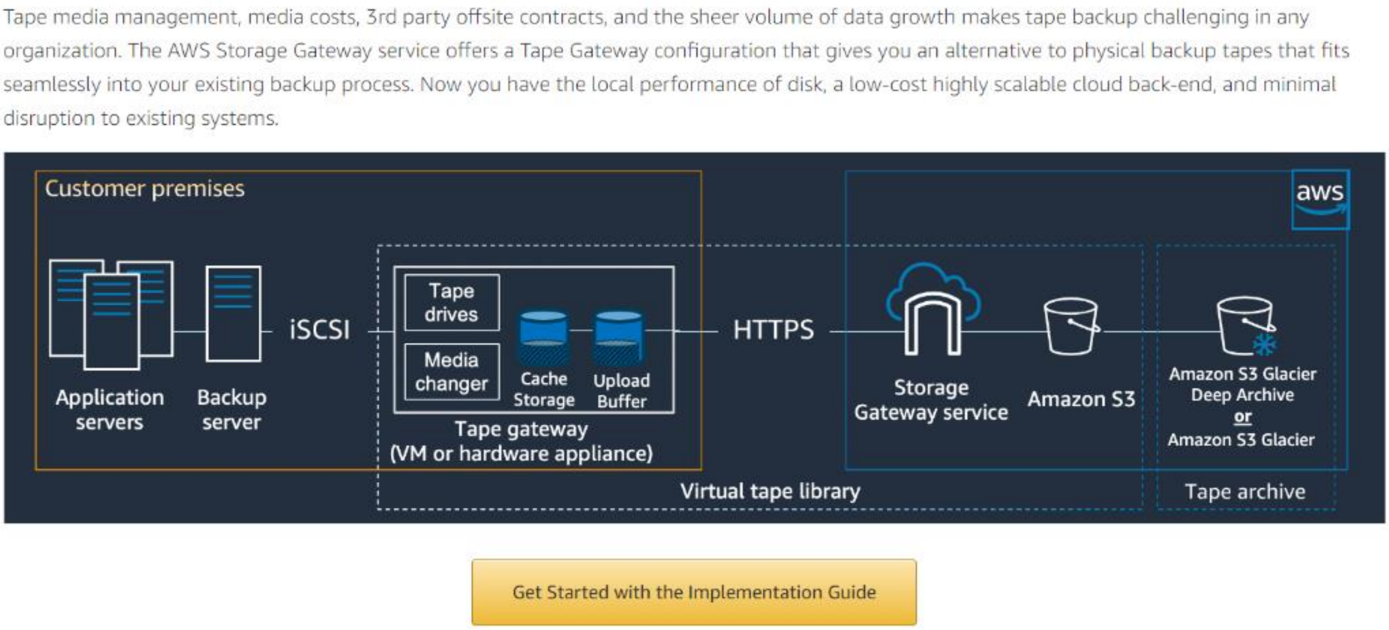
Projects on AWS:

Replace Tape Backup with Cloud Storage

Eliminate on-premises tape and automation with durable, affordable online archive

Get Started with the Implementation Guide

8 Steps | 40 Minutes



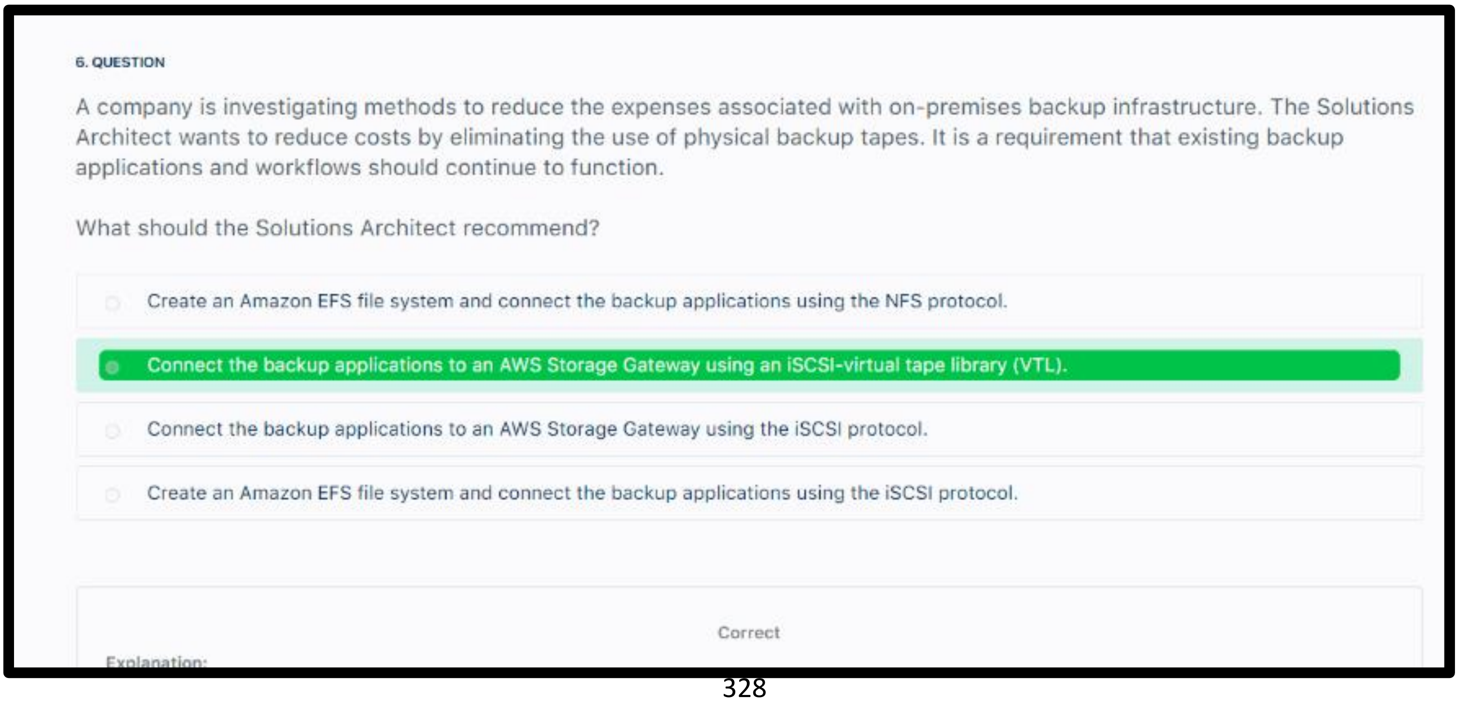
Tape media management, media costs, 3rd party offsite contracts, and the sheer volume of data growth makes tape backup challenging in any organization. The AWS Storage Gateway service offers a Tape Gateway configuration that gives you an alternative to physical backup tapes that fits seamlessly into your existing backup process. Now you have the local performance of disk, a low-cost highly scalable cloud back-end, and minimal disruption to existing systems.

NOTE

Comparison with FSx

AWS announced their "Storage Gateway" service in 2012. Note that in 2018 they're going to announce something called FSx. The "FS" stands for filesystem and the "x" simply denotes that many different filesystems are going to be provided for. For example, some values of the "x" include the Lustre file system and Microsoft Windows.

Why was FSx required? Well, Storage Gateway is for hybrid architectures, which are partly on-premises and partly in the cloud. The



6. QUESTION

A company is investigating methods to reduce the expenses associated with on-premises backup infrastructure. The Solutions Architect wants to reduce costs by eliminating the use of physical backup tapes. It is a requirement that existing backup applications and workflows should continue to function.

What should the Solutions Architect recommend?

○ Create an Amazon EFS file system and connect the backup applications using the NFS protocol.

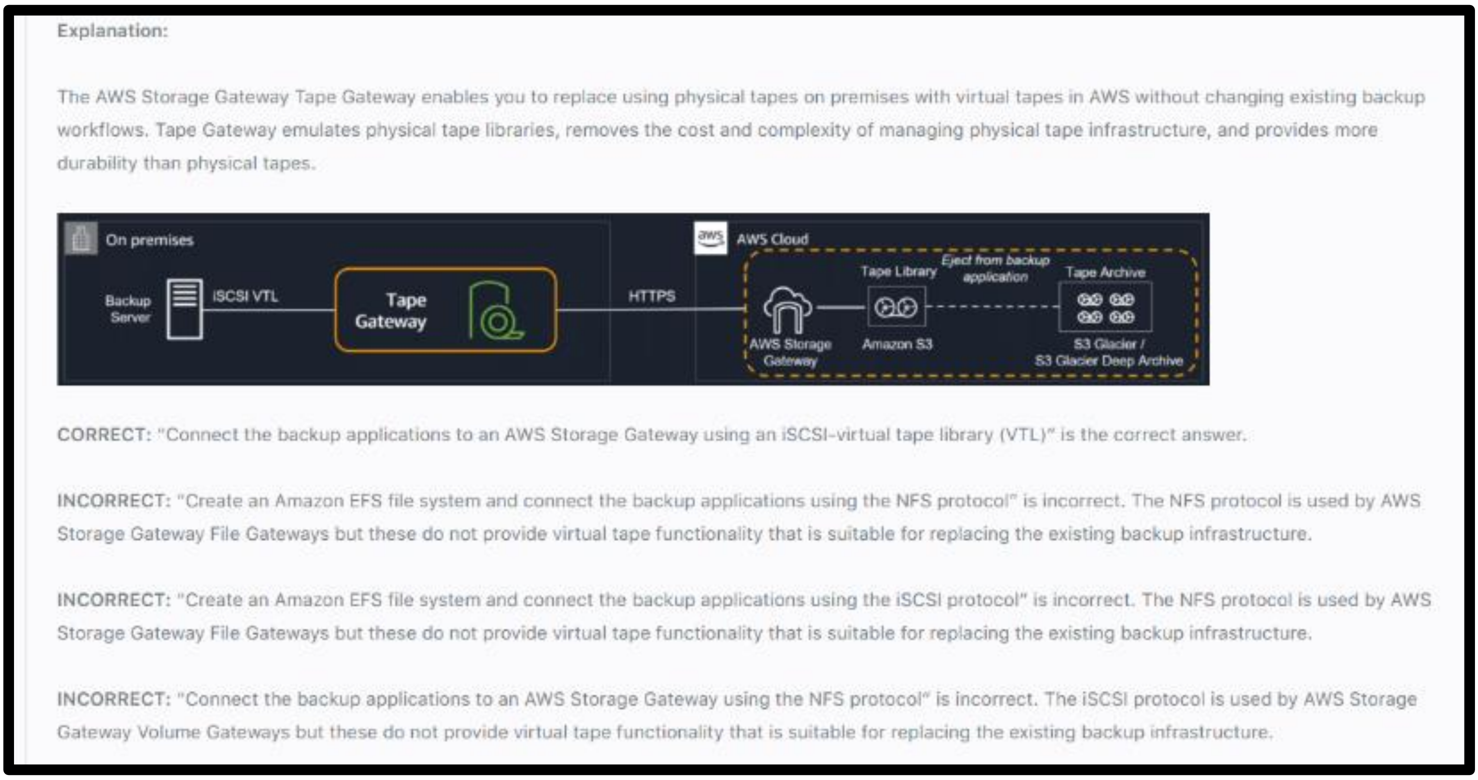
● Connect the backup applications to an AWS Storage Gateway using an iSCSI-virtual tape library (VTL).

○ Connect the backup applications to an AWS Storage Gateway using the iSCSI protocol.

○ Create an Amazon EFS file system and connect the backup applications using the iSCSI protocol.

Correct

Explanation:



Explanation:

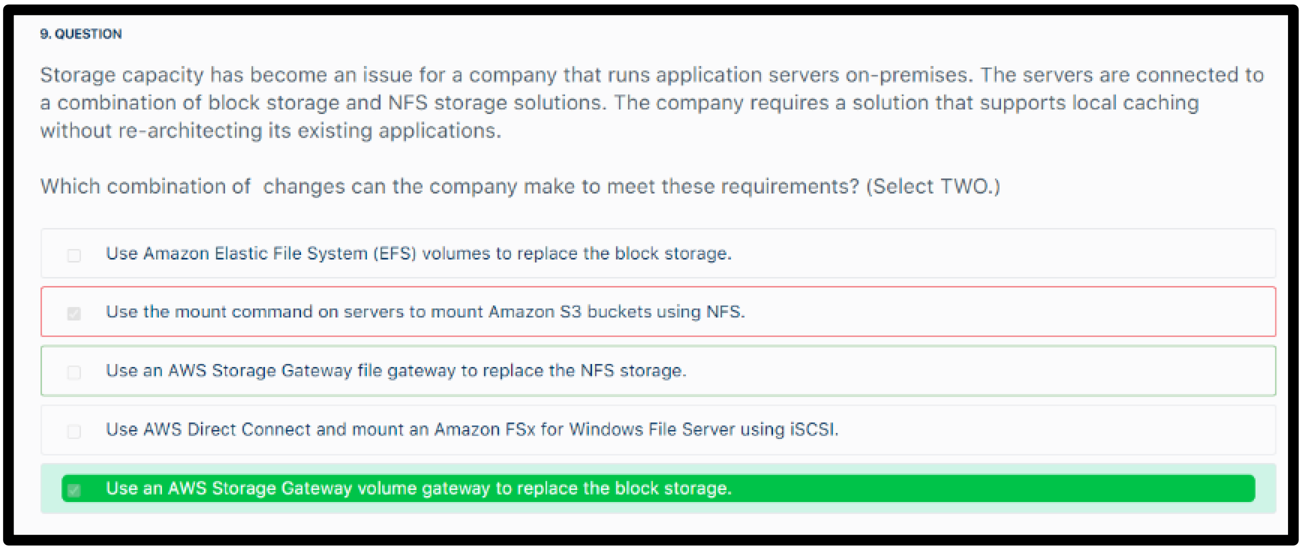
The AWS Storage Gateway Tape Gateway enables you to replace using physical tapes on premises with virtual tapes in AWS without changing existing backup workflows. Tape Gateway emulates physical tape libraries, removes the cost and complexity of managing physical tape infrastructure, and provides more durability than physical tapes.

CORRECT: "Connect the backup applications to an AWS Storage Gateway using an iSCSI-virtual tape library (VTL)" is the correct answer.

INCORRECT: "Create an Amazon EFS file system and connect the backup applications using the NFS protocol" is incorrect. The NFS protocol is used by AWS Storage Gateway File Gateways but these do not provide virtual tape functionality that is suitable for replacing the existing backup infrastructure.

INCORRECT: "Create an Amazon EFS file system and connect the backup applications using the iSCSI protocol" is incorrect. The NFS protocol is used by AWS Storage Gateway File Gateways but these do not provide virtual tape functionality that is suitable for replacing the existing backup infrastructure.

INCORRECT: "Connect the backup applications to an AWS Storage Gateway using the NFS protocol" is incorrect. The iSCSI protocol is used by AWS Storage Gateway Volume Gateways but these do not provide virtual tape functionality that is suitable for replacing the existing backup infrastructure.



9. QUESTION

Storage capacity has become an issue for a company that runs application servers on-premises. The servers are connected to a combination of block storage and NFS storage solutions. The company requires a solution that supports local caching without re-architecting its existing applications.

Which combination of changes can the company make to meet these requirements? (Select TWO.)

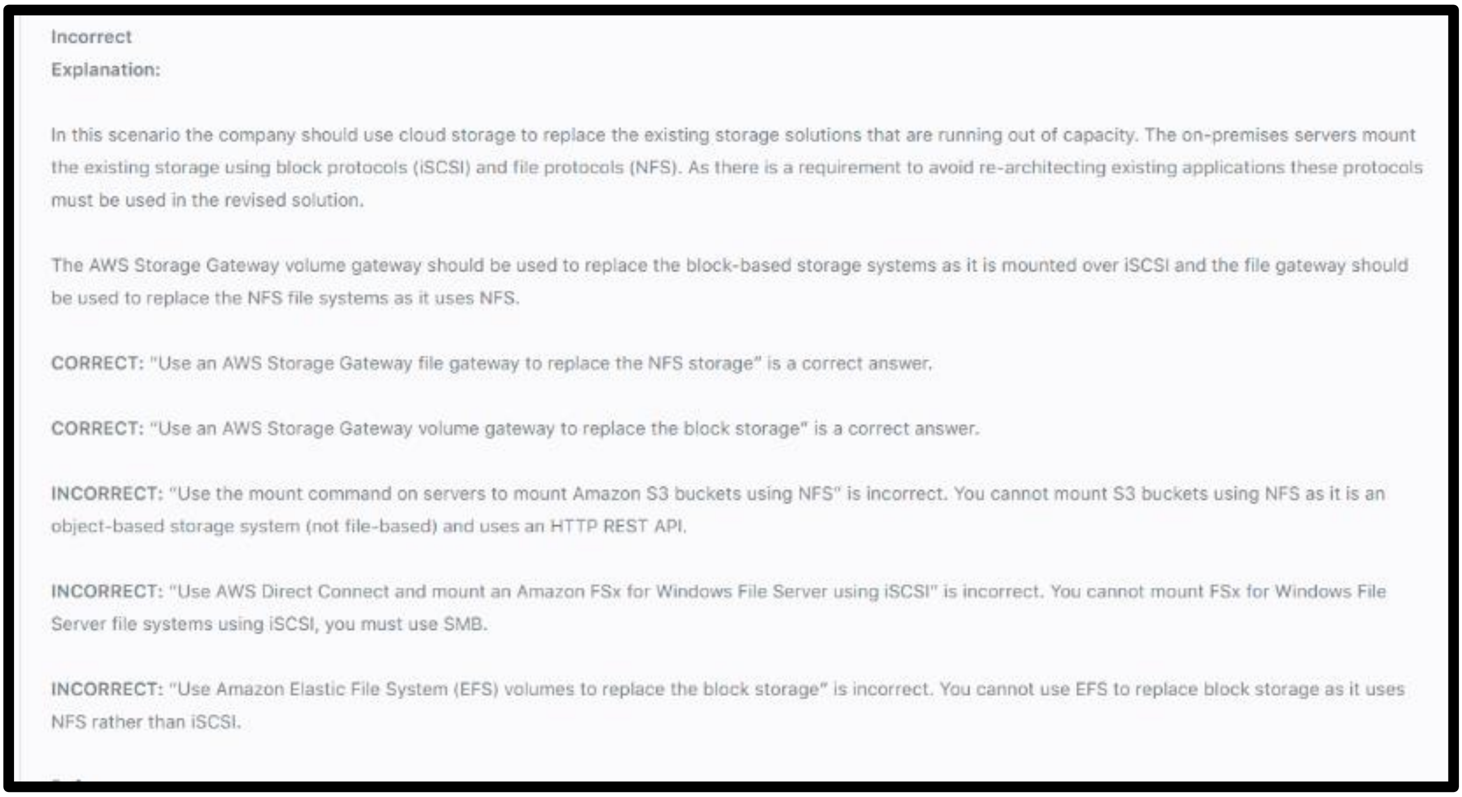
☐ Use Amazon Elastic File System (EFS) volumes to replace the block storage.

☑ Use the mount command on servers to mount Amazon S3 buckets using NFS.

☐ Use an AWS Storage Gateway file gateway to replace the NFS storage.

☐ Use AWS Direct Connect and mount an Amazon FSx for Windows File Server using iSCSI.

☑ Use an AWS Storage Gateway volume gateway to replace the block storage.



Incorrect

Explanation:

In this scenario the company should use cloud storage to replace the existing storage solutions that are running out of capacity. The on-premises servers mount the existing storage using block protocols (iSCSI) and file protocols (NFS). As there is a requirement to avoid re-architecting existing applications these protocols must be used in the revised solution.

The AWS Storage Gateway volume gateway should be used to replace the block-based storage systems as it is mounted over iSCSI and the file gateway should be used to replace the NFS file systems as it uses NFS.

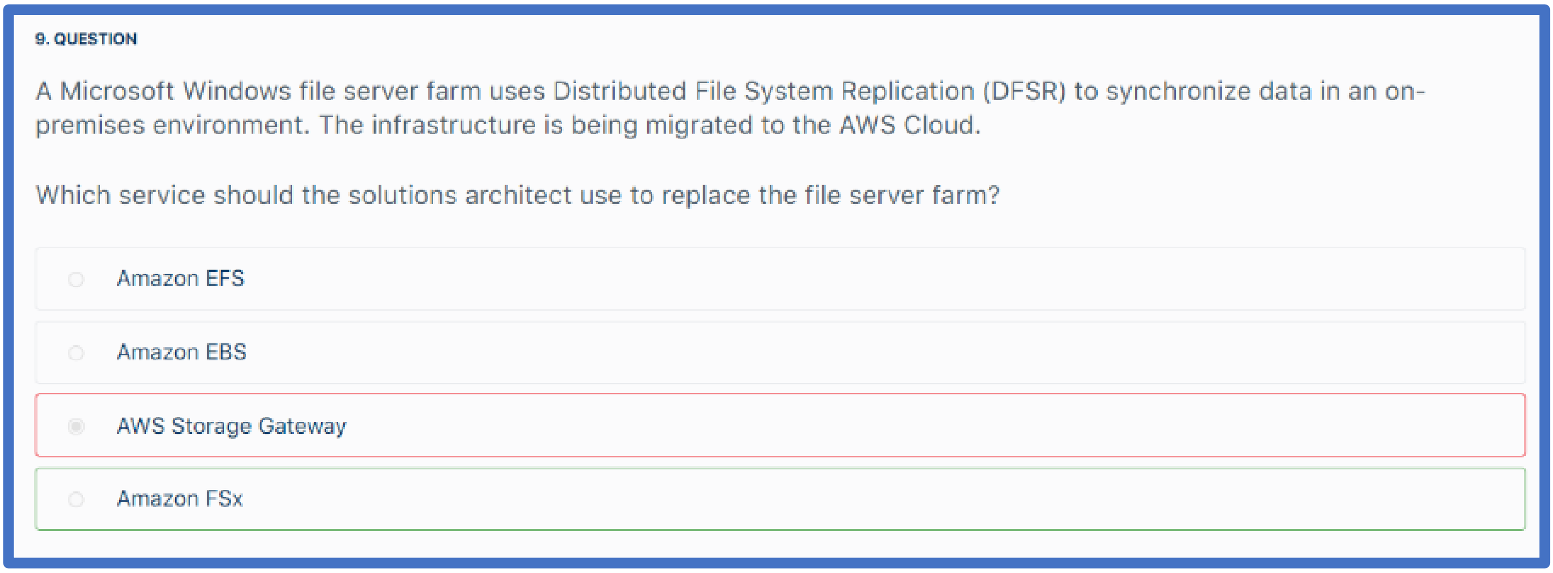
CORRECT: "Use an AWS Storage Gateway file gateway to replace the NFS storage" is a correct answer.

CORRECT: "Use an AWS Storage Gateway volume gateway to replace the block storage" is a correct answer.

INCORRECT: "Use the mount command on servers to mount Amazon S3 buckets using NFS" is incorrect. You cannot mount S3 buckets using NFS as it is an object-based storage system (not file-based) and uses an HTTP REST API.

INCORRECT: "Use AWS Direct Connect and mount an Amazon FSx for Windows File Server using iSCSI" is incorrect. You cannot mount FSx for Windows File Server file systems using iSCSI, you must use SMB.

INCORRECT: "Use Amazon Elastic File System (EFS) volumes to replace the block storage" is incorrect. You cannot use EFS to replace block storage as it uses NFS rather than iSCSI.



9. QUESTION

A Microsoft Windows file server farm uses Distributed File System Replication (DFSR) to synchronize data in an on-premises environment. The infrastructure is being migrated to the AWS Cloud.

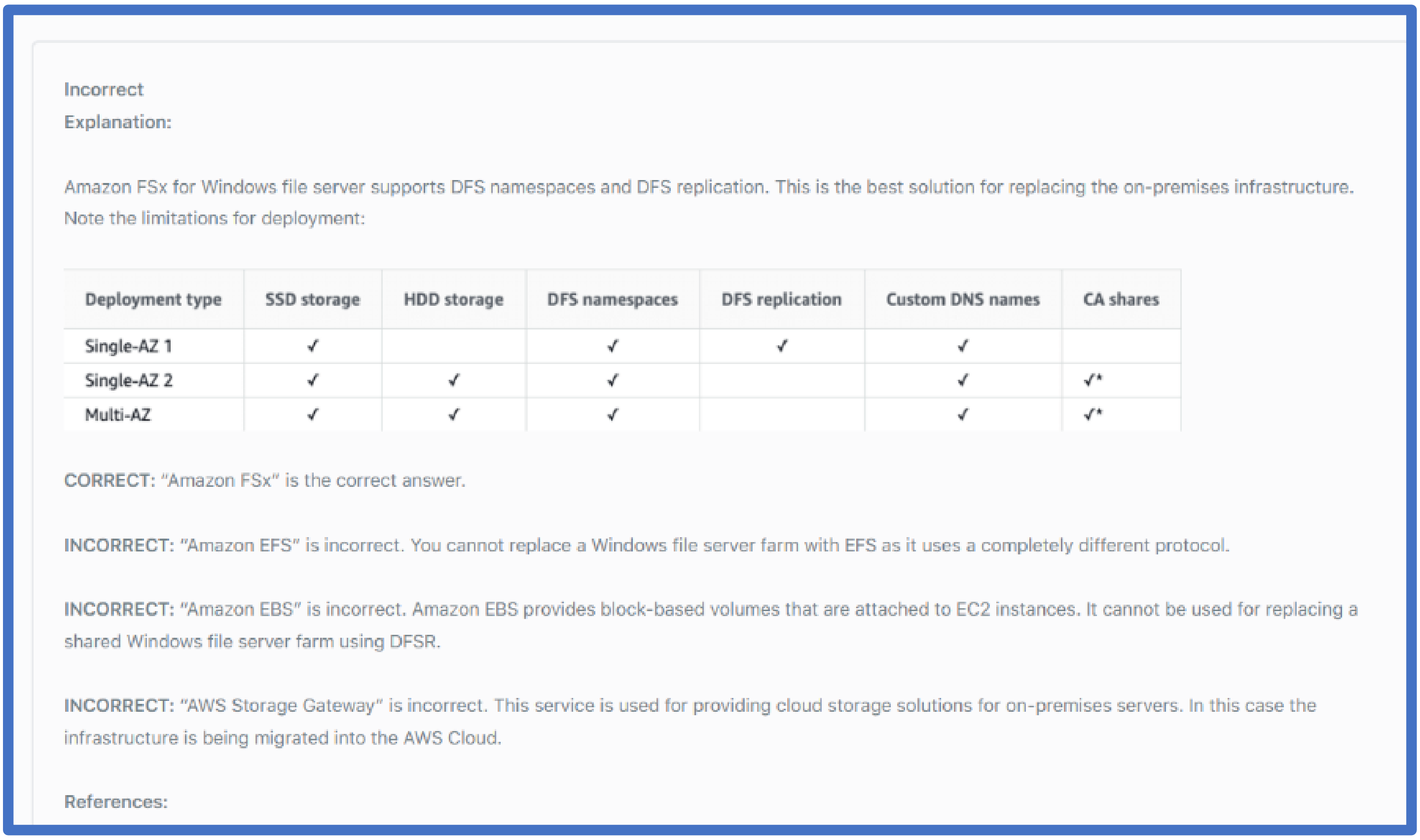
Which service should the solutions architect use to replace the file server farm?

○ Amazon EFS

○ Amazon EBS

● AWS Storage Gateway

○ Amazon FSx



Incorrect

Explanation:

Amazon FSx for Windows file server supports DFS namespaces and DFS replication. This is the best solution for replacing the on-premises infrastructure. Note the limitations for deployment:

Deployment type	SSD storage	HDD storage	DFS namespaces	DFS replication	Custom DNS names	CA shares
Single-AZ 1	✓		✓	✓	✓	
Single-AZ 2	✓	✓	✓		✓	✓*
Multi-AZ	✓	✓	✓		✓	✓*

CORRECT: "Amazon FSx" is the correct answer.

INCORRECT: "Amazon EFS" is incorrect. You cannot replace a Windows file server farm with EFS as it uses a completely different protocol.

INCORRECT: "Amazon EBS" is incorrect. Amazon EBS provides block-based volumes that are attached to EC2 instances. It cannot be used for replacing a shared Windows file server farm using DFSR.

INCORRECT: "AWS Storage Gateway" is incorrect. This service is used for providing cloud storage solutions for on-premises servers. In this case the infrastructure is being migrated into the AWS Cloud.

References:

iSCSI

Ten facts about iSCSI

iSCSI stands for "Internet Small Computer Systems Interface".

1. It is a continuation of SCSI

SCSI is just the plain old "small computer systems interface". This was a protocol designed for communication with storage. The main competitor to "Small Computer Systems Interface" is Fibre Channel. SCSI uses TCP/IP as the basis for transport. The idea is that "Small Computer Systems Interface" commands are encapsulated in a TCP/IP packet.

2. The original specification is RFC 3720

3. iSCSI is a request-response protocol

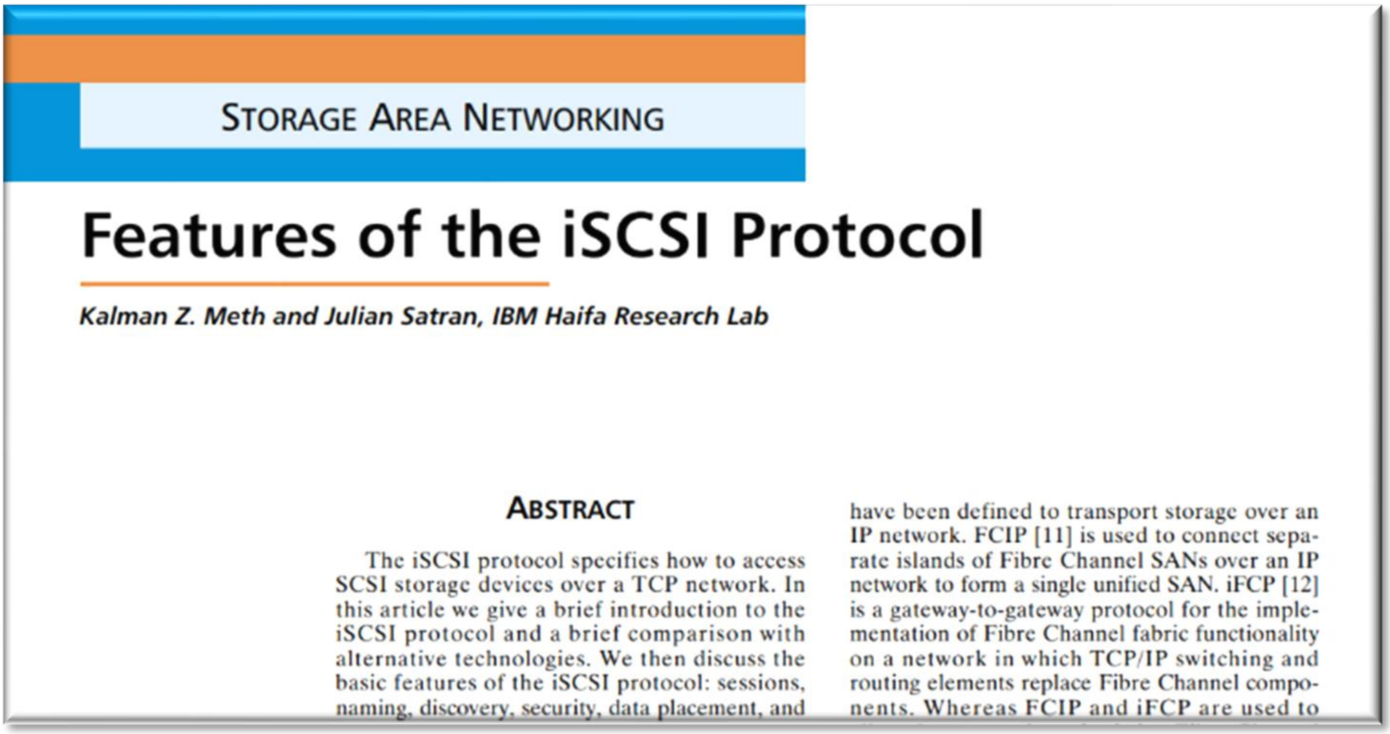
This means that there can be no response until there is a request. "Initiators" are where requests are created; "targets" are where requests are serviced. These two components—iSCSI initiator and iSCSI target—communicate with one another.

The original SCSI does not use requests. Instead, it uses commands. "SCSI commands" are transmitted as "iSCSI requests". "SCSI responses" and "SCSI status messages" are transmitted at iSCSI responses.

4. Clients and servers have different names in iSCSI

A server in an iSCSI storage network is in fact called the "iSCSI target node". Such an iSCSI Target can provide one or more logical units (LUs). The following forms of target are available for the "Internet Small Computer Systems Interface":

1. As hardware, as iSCSI Storage Arrays
2. As software, as iSCSI Target software for installation on a standard server



STORAGE AREA NETWORKING

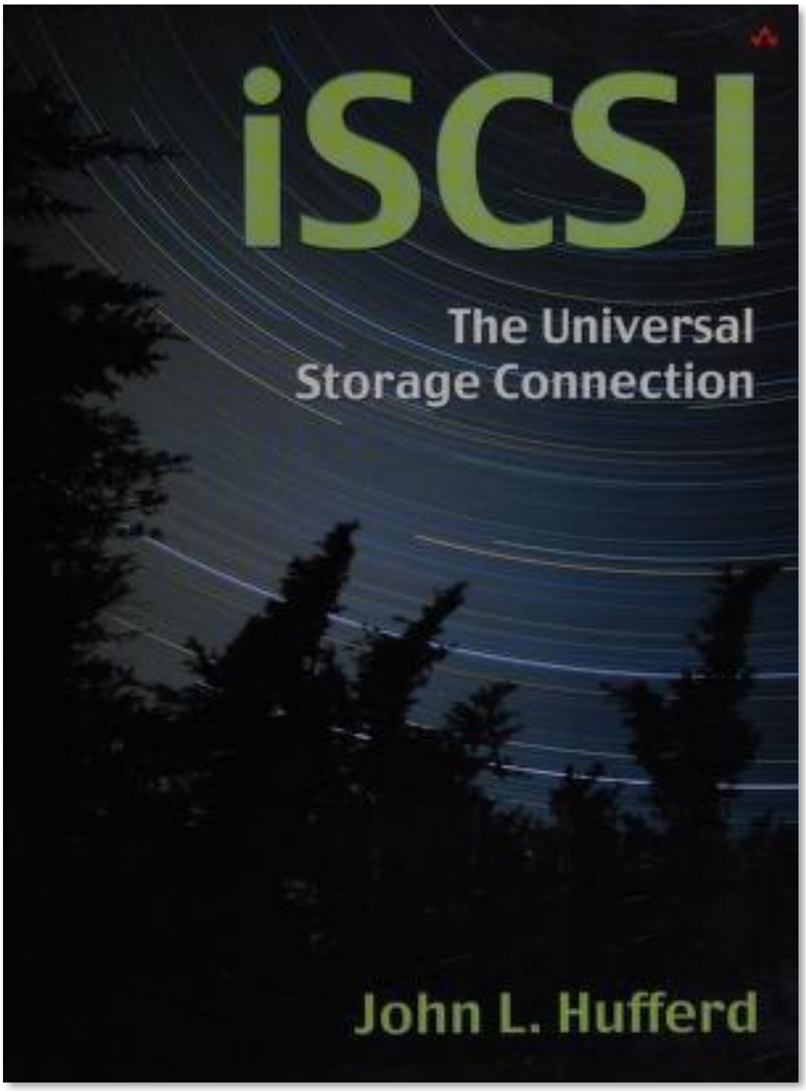
Features of the iSCSI Protocol

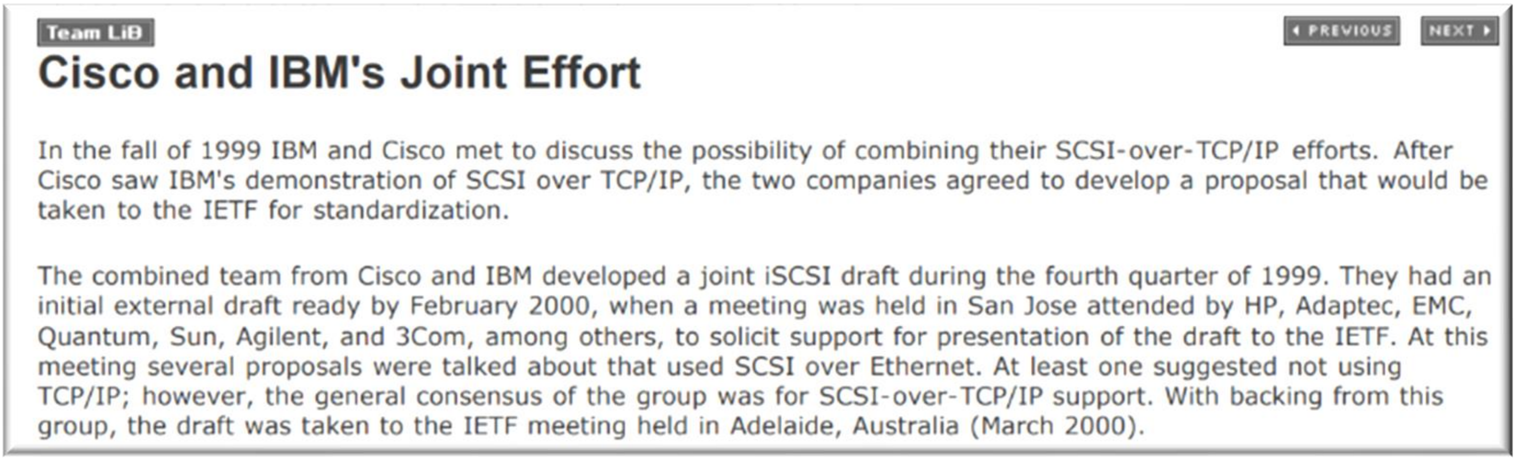
Kalman Z. Meth and Julian Satran, IBM Haifa Research Lab

ABSTRACT

The iSCSI protocol specifies how to access SCSI storage devices over a TCP network. In this article we give a brief introduction to the iSCSI protocol and a brief comparison with alternative technologies. We then discuss the basic features of the iSCSI protocol: sessions, naming, discovery, security, data placement, and

have been defined to transport storage over an IP network. FCIP [11] is used to connect separate islands of Fibre Channel SANs over an IP network to form a single unified SAN. iFCP [12] is a gateway-to-gateway protocol for the implementation of Fibre Channel fabric functionality on a network in which TCP/IP switching and routing elements replace Fibre Channel components. Whereas FCIP and iFCP are used to

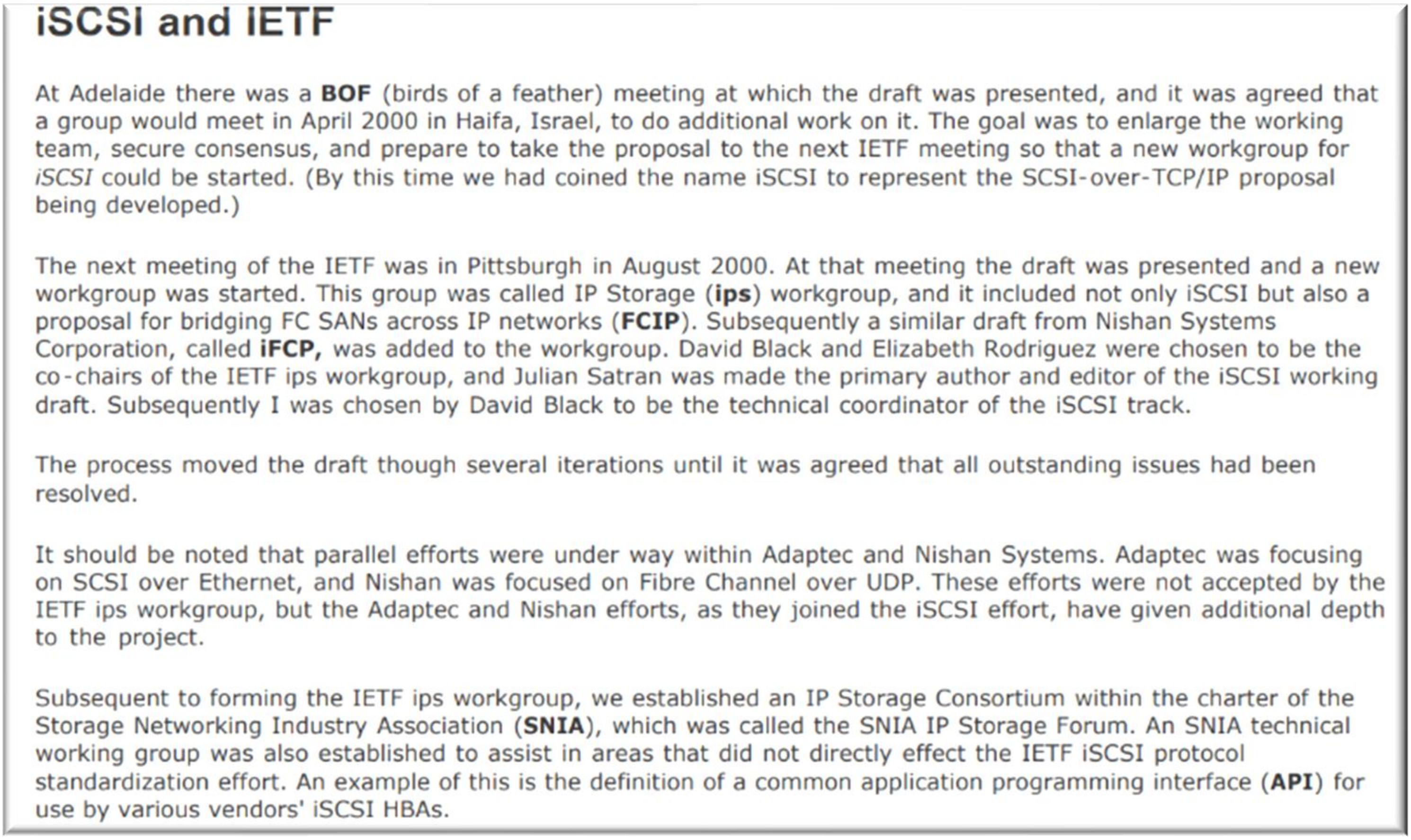




Cisco and IBM's Joint Effort

In the fall of 1999 IBM and Cisco met to discuss the possibility of combining their SCSI-over-TCP/IP efforts. After Cisco saw IBM's demonstration of SCSI over TCP/IP, the two companies agreed to develop a proposal that would be taken to the IETF for standardization.

The combined team from Cisco and IBM developed a joint iSCSI draft during the fourth quarter of 1999. They had an initial external draft ready by February 2000, when a meeting was held in San Jose attended by HP, Adaptec, EMC, Quantum, Sun, Agilent, and 3Com, among others, to solicit support for presentation of the draft to the IETF. At this meeting several proposals were talked about that used SCSI over Ethernet. At least one suggested not using TCP/IP; however, the general consensus of the group was for SCSI-over-TCP/IP support. With backing from this group, the draft was taken to the IETF meeting held in Adelaide, Australia (March 2000).



iSCSI and IETF

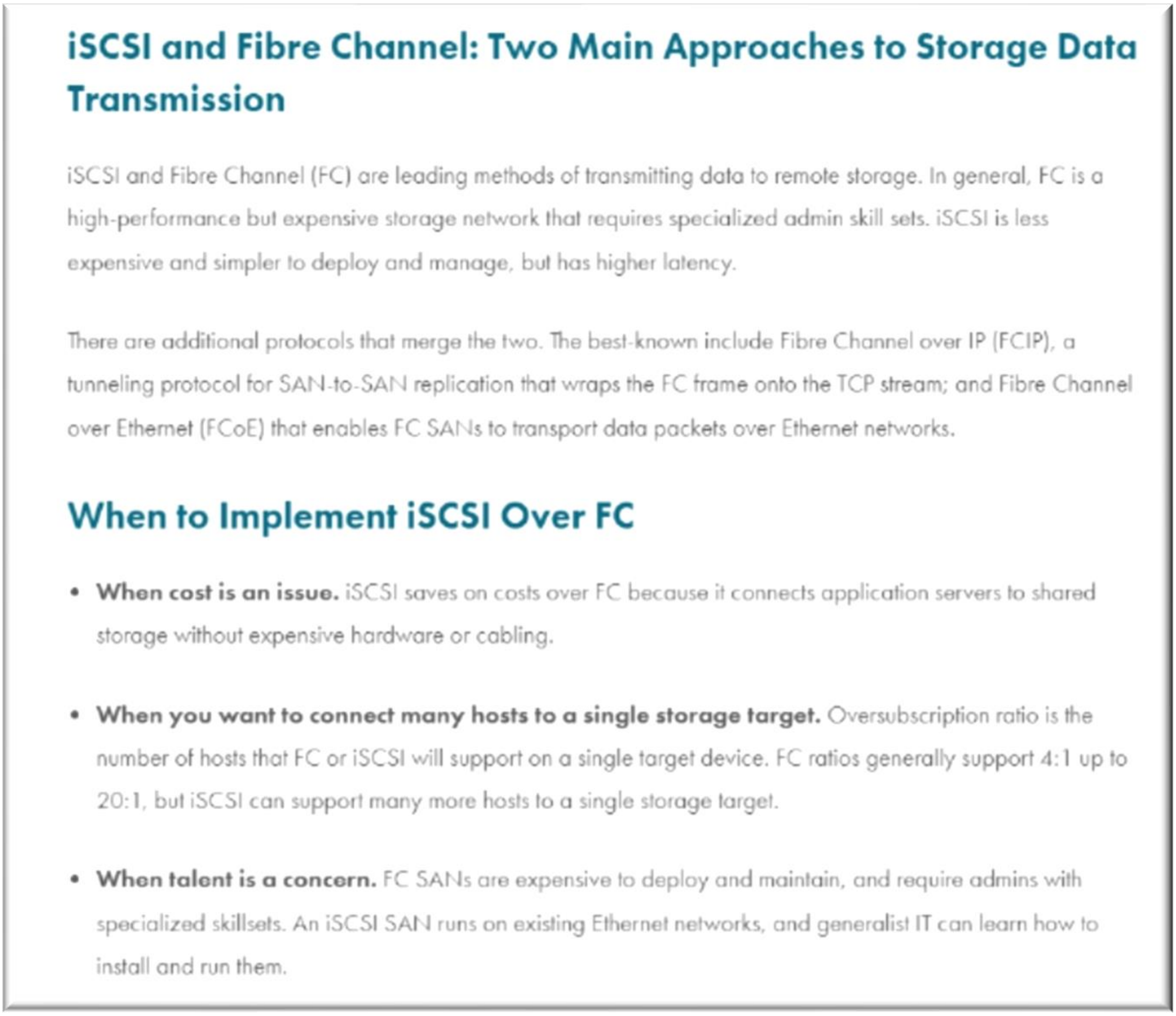
At Adelaide there was a **BOF** (birds of a feather) meeting at which the draft was presented, and it was agreed that a group would meet in April 2000 in Haifa, Israel, to do additional work on it. The goal was to enlarge the working team, secure consensus, and prepare to take the proposal to the next IETF meeting so that a new workgroup for *iSCSI* could be started. (By this time we had coined the name iSCSI to represent the SCSI-over-TCP/IP proposal being developed.)

The next meeting of the IETF was in Pittsburgh in August 2000. At that meeting the draft was presented and a new workgroup was started. This group was called IP Storage (**ips**) workgroup, and it included not only iSCSI but also a proposal for bridging FC SANs across IP networks (**FCIP**). Subsequently a similar draft from Nishan Systems Corporation, called **iFCP,** was added to the workgroup. David Black and Elizabeth Rodriguez were chosen to be the co-chairs of the IETF ips workgroup, and Julian Satran was made the primary author and editor of the iSCSI working draft. Subsequently I was chosen by David Black to be the technical coordinator of the iSCSI track.

The process moved the draft though several iterations until it was agreed that all outstanding issues had been resolved.

It should be noted that parallel efforts were under way within Adaptec and Nishan Systems. Adaptec was focusing on SCSI over Ethernet, and Nishan was focused on Fibre Channel over UDP. These efforts were not accepted by the IETF ips workgroup, but the Adaptec and Nishan efforts, as they joined the iSCSI effort, have given additional depth to the project.

Subsequent to forming the IETF ips workgroup, we established an IP Storage Consortium within the charter of the Storage Networking Industry Association (**SNIA**), which was called the SNIA IP Storage Forum. An SNIA technical working group was also established to assist in areas that did not directly effect the IETF iSCSI protocol standardization effort. An example of this is the definition of a common application programming interface (**API**) for use by various vendors' iSCSI HBAs.



iSCSI and Fibre Channel: Two Main Approaches to Storage Data Transmission

iSCSI and Fibre Channel (FC) are leading methods of transmitting data to remote storage. In general, FC is a high-performance but expensive storage network that requires specialized admin skill sets. iSCSI is less expensive and simpler to deploy and manage, but has higher latency.

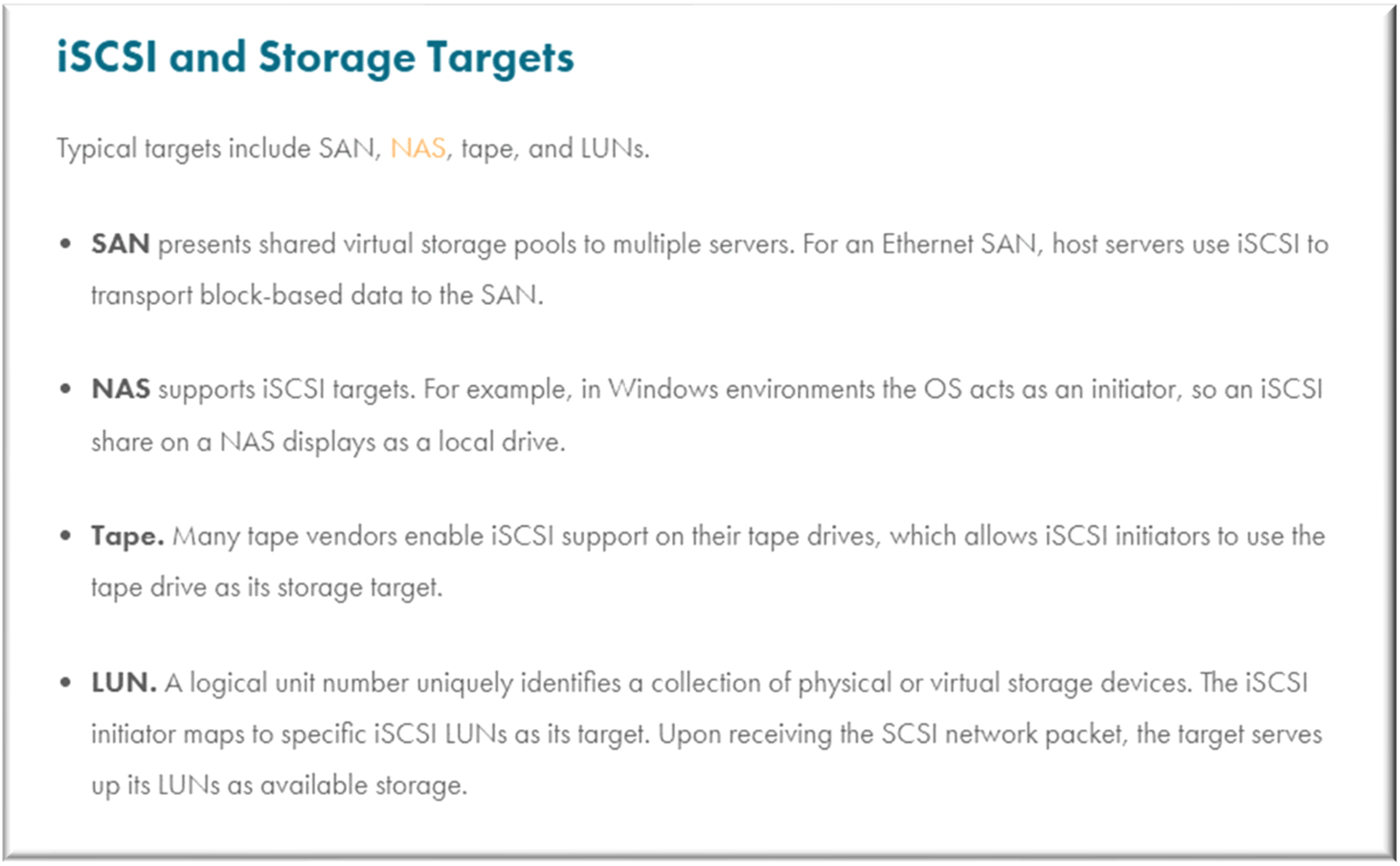
There are additional protocols that merge the two. The best-known include Fibre Channel over IP (FCIP), a tunneling protocol for SAN-to-SAN replication that wraps the FC frame onto the TCP stream; and Fibre Channel over Ethernet (FCoE) that enables FC SANs to transport data packets over Ethernet networks.

When to Implement iSCSI Over FC

- **When cost is an issue.** iSCSI saves on costs over FC because it connects application servers to shared storage without expensive hardware or cabling.

- **When you want to connect many hosts to a single storage target.** Oversubscription ratio is the number of hosts that FC or iSCSI will support on a single target device. FC ratios generally support 4:1 up to 20:1, but iSCSI can support many more hosts to a single storage target.

- **When talent is a concern.** FC SANs are expensive to deploy and maintain, and require admins with specialized skillsets. An iSCSI SAN runs on existing Ethernet networks, and generalist IT can learn how to install and run them.

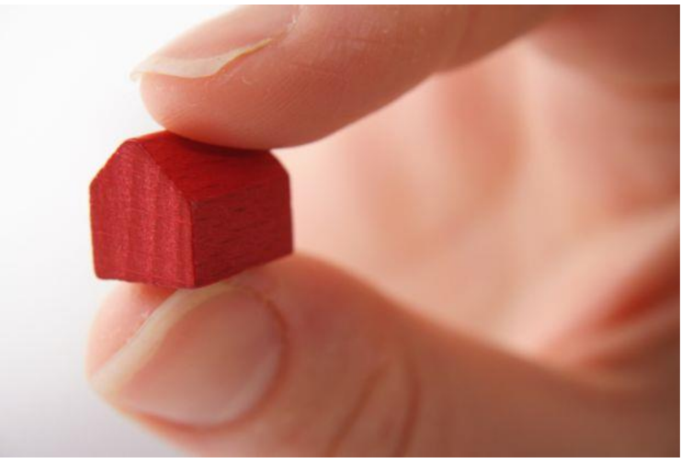


iSCSI and Storage Targets

Typical targets include SAN, NAS, tape, and LUNs.

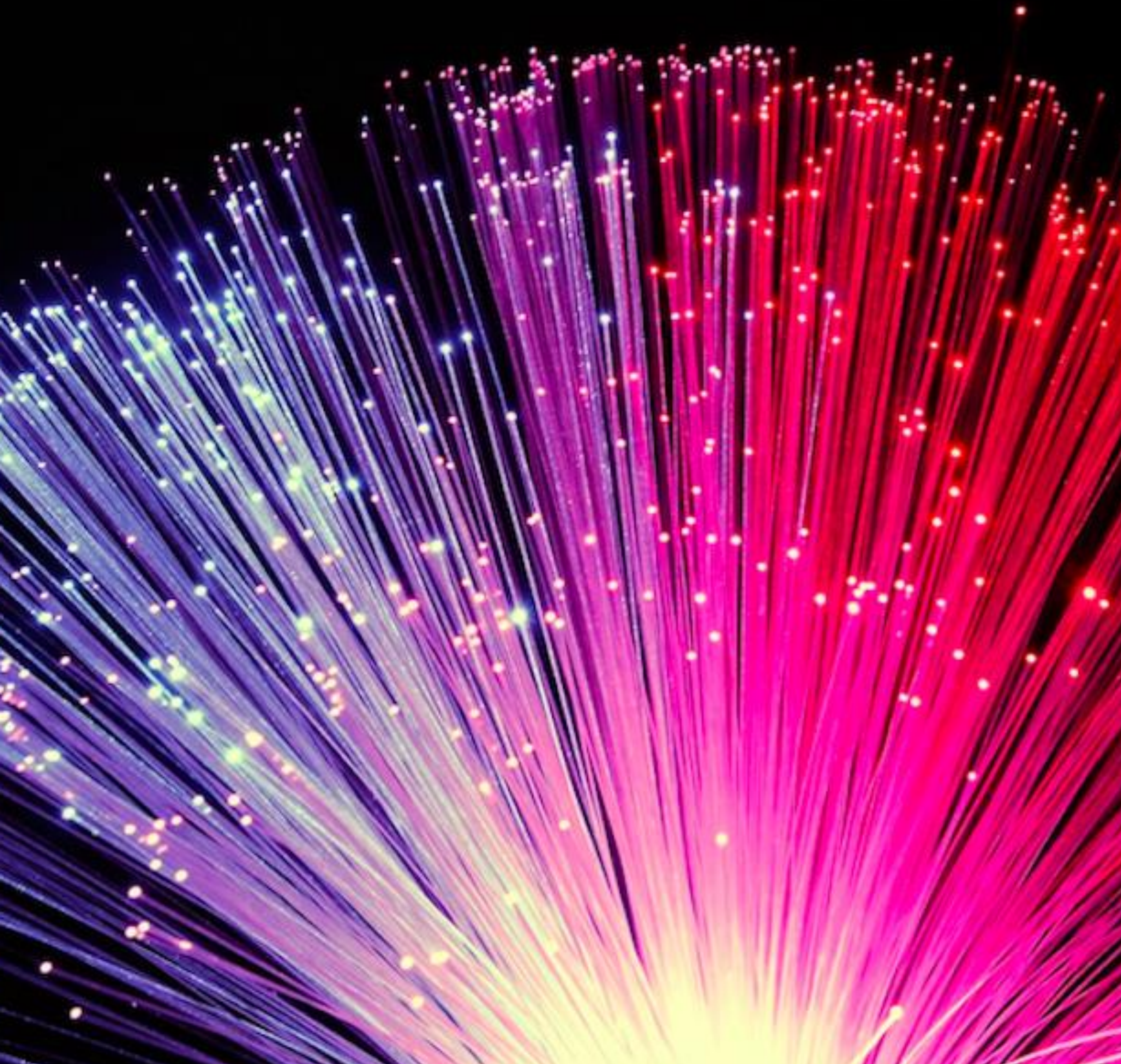
- **SAN** presents shared virtual storage pools to multiple servers. For an Ethernet SAN, host servers use iSCSI to transport block-based data to the SAN.
- **NAS** supports iSCSI targets. For example, in Windows environments the OS acts as an initiator, so an iSCSI share on a NAS displays as a local drive.
- **Tape.** Many tape vendors enable iSCSI support on their tape drives, which allows iSCSI initiators to use the tape drive as its storage target.
- **LUN.** A logical unit number uniquely identifies a collection of physical or virtual storage devices. The iSCSI initiator maps to specific iSCSI LUNs as its target. Upon receiving the SCSI network packet, the target serves up its LUNs as available storage.

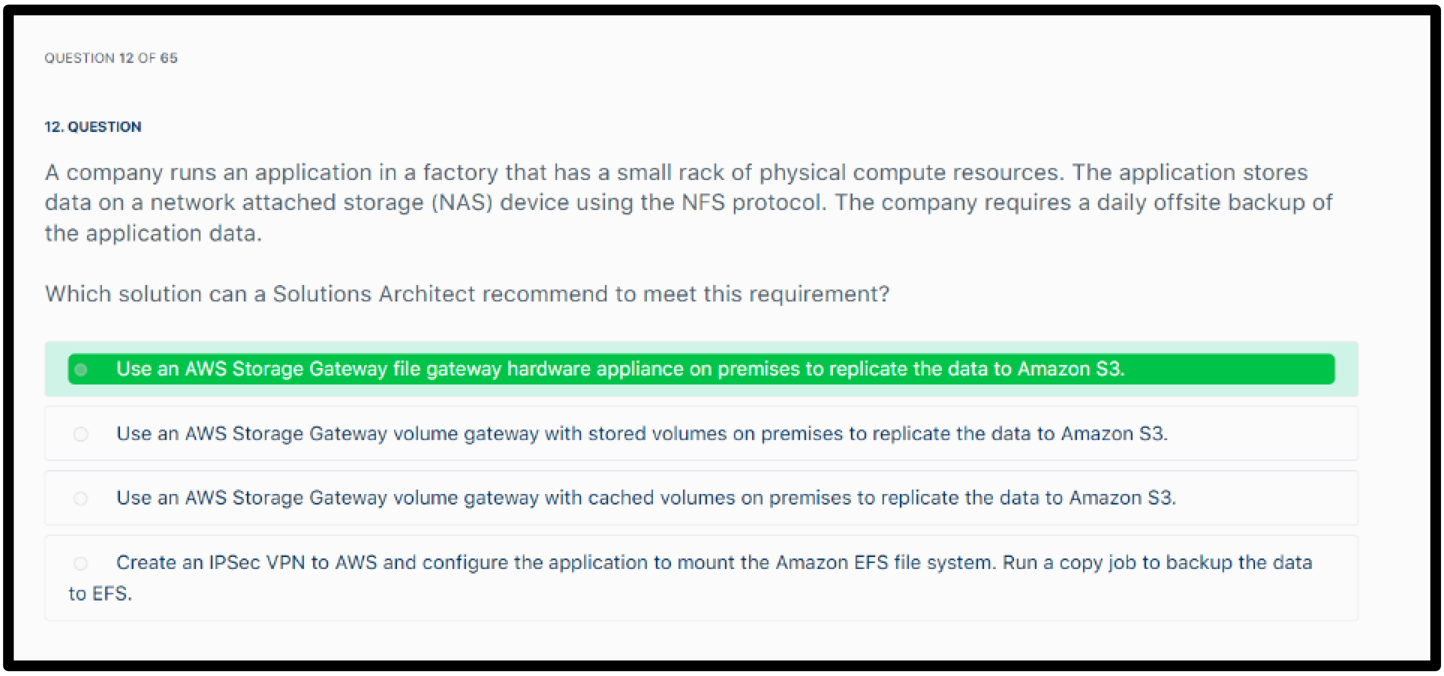
The Haifa research laboratory in Israel



Small Computer Systems Interface

FIBRE CHANNEL





QUESTION 12 OF 65

12. QUESTION

A company runs an application in a factory that has a small rack of physical compute resources. The application stores data on a network attached storage (NAS) device using the NFS protocol. The company requires a daily offsite backup of the application data.

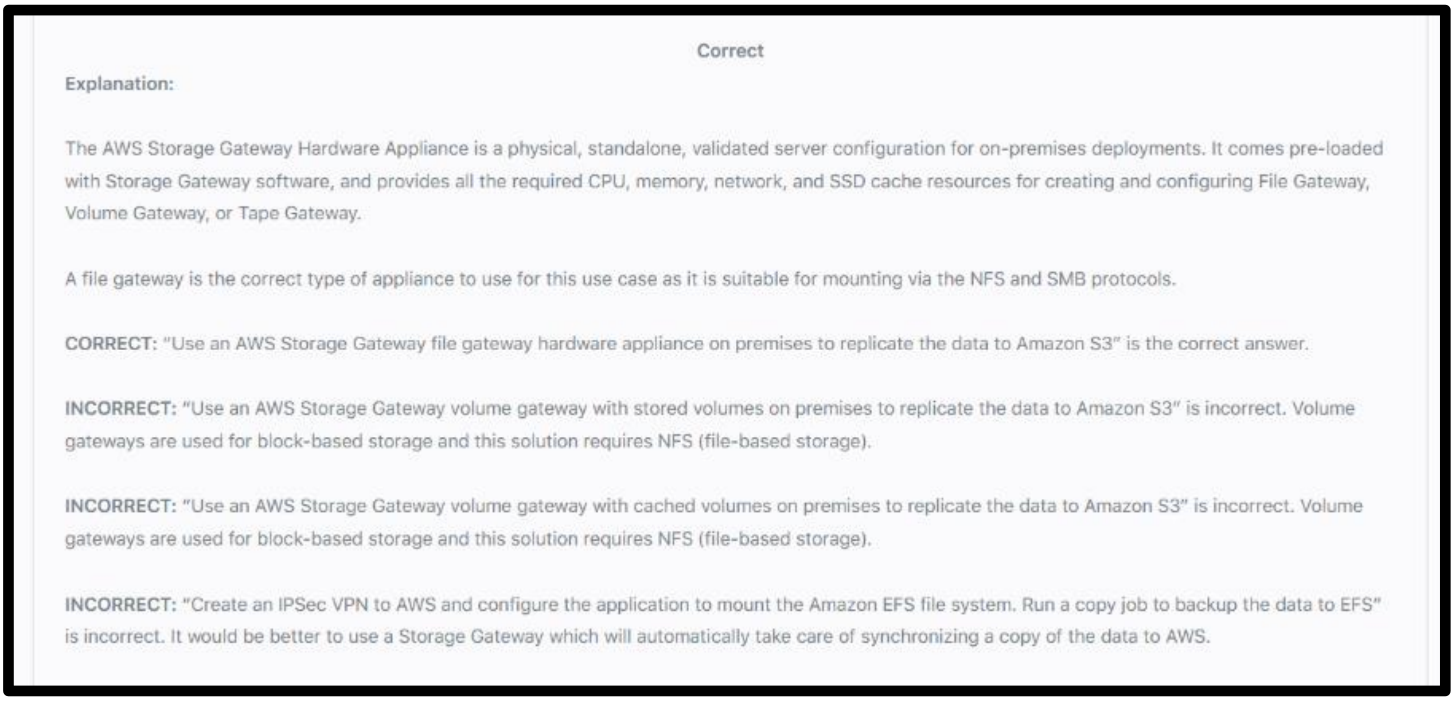
Which solution can a Solutions Architect recommend to meet this requirement?

● Use an AWS Storage Gateway file gateway hardware appliance on premises to replicate the data to Amazon S3.

○ Use an AWS Storage Gateway volume gateway with stored volumes on premises to replicate the data to Amazon S3.

○ Use an AWS Storage Gateway volume gateway with cached volumes on premises to replicate the data to Amazon S3.

○ Create an IPSec VPN to AWS and configure the application to mount the Amazon EFS file system. Run a copy job to backup the data to EFS.



Correct

Explanation:

The AWS Storage Gateway Hardware Appliance is a physical, standalone, validated server configuration for on-premises deployments. It comes pre-loaded with Storage Gateway software, and provides all the required CPU, memory, network, and SSD cache resources for creating and configuring File Gateway, Volume Gateway, or Tape Gateway.

A file gateway is the correct type of appliance to use for this use case as it is suitable for mounting via the NFS and SMB protocols.

CORRECT: "Use an AWS Storage Gateway file gateway hardware appliance on premises to replicate the data to Amazon S3" is the correct answer.

INCORRECT: "Use an AWS Storage Gateway volume gateway with stored volumes on premises to replicate the data to Amazon S3" is incorrect. Volume gateways are used for block-based storage and this solution requires NFS (file-based storage).

INCORRECT: "Use an AWS Storage Gateway volume gateway with cached volumes on premises to replicate the data to Amazon S3" is incorrect. Volume gateways are used for block-based storage and this solution requires NFS (file-based storage).

INCORRECT: "Create an IPSec VPN to AWS and configure the application to mount the Amazon EFS file system. Run a copy job to backup the data to EFS" is incorrect. It would be better to use a Storage Gateway which will automatically take care of synchronizing a copy of the data to AWS.

Bibliography

I. Official

II. Unofficial

III. Critical

IV. General

https://www.techtarget.com/searchstorage/definition/iSCSI

https://www.enterprisestorageforum.com/hardware/what-is-iscsi-and-how-does-it-work/

Simple Workflow

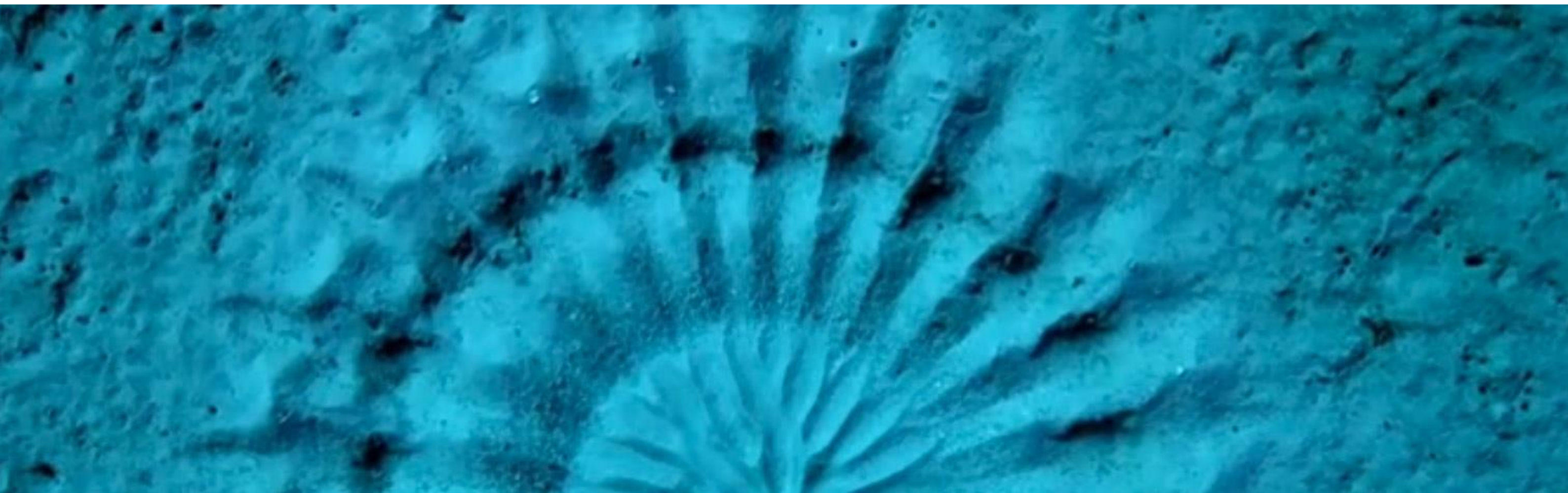
What is a workflow?

It we look this up in older dictionaries, we will not find an entry for "workflow". This is a newer term.

"Flow" is considered a good thing, and "work" a bad thing. Therefore, the proved very useful to marketers, since it suggests that their product can remove work.

The Internet is now littered with articles purporting to explain what a "workflow" is. You will be told that a workflow is a sequence of activities.

This isn't very helpful. I want you to consider getting children ready for school in the morning; the rise of Nazism in Germany in the 1930s; the mating rituals performed by White-spotted pufferfish; and the entire life of Jeffrey Bezos. All of these things could be described as "sequences of activities". So, can we do better in defining a **WORKFLOW**?



Intricate sand art produced by the White-spotted pufferfish

Balan Subramanian explaining how Simple Workflow (SWF) works, in 2012

Your Workers and your Deciders can be written in the programming language of your choice, and they can run in the

> cloud (e.g. on an Amazon EC2 instance), in your data center, or even on your desktop. You need only poll for work, handle it, and return the results to Simple Workflow. In other words, your code can run anywhere, as long as it can "see" the Amazon Simple Workflow HTTPS endpoint. This gives you the flexibility to incorporate existing on-premise systems into new, cloud-based workflows. Simple Workflow lets you do "long polling" to reduce network traffic and unnecessary processing within your code. With this model, requests from your code will be held open for up to 60 seconds if necessary.
>
> Barr 2012

> In order to make it even easier for you to get started with Amazon Simple Workflow, the AWS SDK for Java now includes the new AWS Flow Framework. This new framework includes a number of programming constructs that abstract out a number of task coordination details. For example, it uses a programming model based on Futures to handle dependencies between tasks. Initiating a Worker task is as easy as making a method call, and the framework takes care of the Workers and the Decision Tasks behind the scenes.
>
> Barr 2012

Now, some critical analysis from [Forrester 2012]:

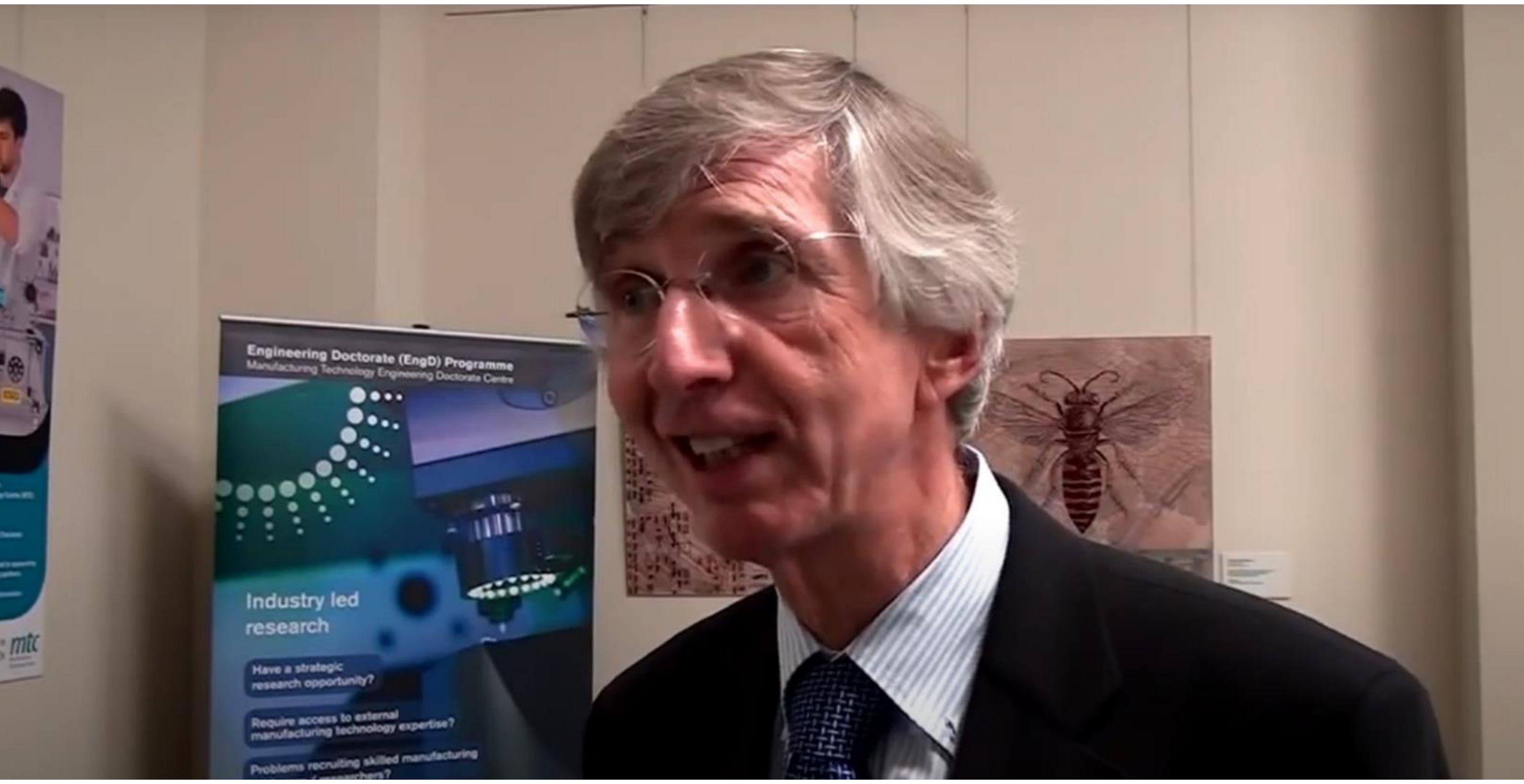
> So Amazon is obviously looking for large scale and transaction throughput. Indeed, this is a significant addition to the business process management (BPM) landscape. If nothing else, the move should rattle the cages of the relatively high-cost model incumbent BPM players. I haven't done a detailed cost analysis, but my perception is that this is potentially an order of magnitude (or two) cheaper than some of the existing cloud offerings out there. Vendors such as Cordys, Appian, and even the likes of IBM and salesforce.com will need to look at the implications of this carefully.
>
> But while it initially sounds like great news for customers, the devil is in the details. As I was trying to wrap my head around it, the descriptions seemed to ask more questions than they answered — until I realized that it is primarily a programming extension framework to AWS. That is, SWF requires relatively strong programming knowledge and deep IT understanding to configure and use. Of course, it

> probably only makes sense if you are leveraging the rest of the AWS platform.
>
> Having discussed this a little internally, we think this offering is light years away from a BPM product and doesn't address many existing BPM barriers — which are not about technology to begin with. "BPM" and "workflow" are not the same thing.
>
> But we can see how emerging vendors of low-cost or open source BPM or other tools could leverage this platform to create distributed applications that meet business process needs. But my sense is that SWF is way beyond the average "power user" business analyst.
>
> Perhaps all that is intentional; the real market they are trying to create is to make Amazon the cloud platform for the orchestration of other in-house BPM tools and applications. That would make sense. So perhaps the opportunity here is for BPM players to develop interfaces to each other through this sort of interface. I am sure there are plenty of vendors exploring that idea right now.

[Forrester 2012]

AWS Simple Workflow(SWF) from Amazon is a unique workflow solution comparing to traditional workflow products such as JBPM and OSWorkflow. SWF is extremely scalable and engineer friendly(in that flow is defined with Java code) while it comes with limitations and lots of gotchas.

[Jiaqi 2012]


Engineering Doctorate (EngD) Programme
Manufacturing Technology Engineering Doctorate Centre
Industry led research
Have a strategic research opportunity?
Require access to external manufacturing technology expertise?
Problems recruiting skilled manufacturing
mtc

This man is called **Kees van Hee**. We pronounce "Kees" similar to the English noun "case" (as in *suitcase*). Kees van Hee was a professor of computer science and based in the Netherlands.

Dddd

Cadence vs SWF

Cadence was conceived and is still led by the original tech leads of the SWF.

SWF had no new features added for the last 5 years. Cadence is open sourced and is under active development.

Cadence was initially based on SWF public API. It uses Thrift and TChannel for communication and SWF uses AWS version of REST. Currently the API is not compatible with SWF as Cadence added a large number of new features and deprecated a few problematic ones. We are planning migrating to gRPC later this year.

Cadence can potentially run on any database that supports single shard multi-row transactions as a backend. Currently it supports Cassandra and MySQL.

SWF has pretty tight throttling limits. Cadence scales very well with use cases in production that require 100s of millions of open workflows and thousands of events per second.

SWF has pretty tight limits on individual payloads and number of events. For example maximum activity input size is 32k. Cadence currently has 256k limit. SWF history size limit is 10k events while Cadence limit 200k. All other limits are also higher.

Cadence has no limit on the activity and workflow execution duration.

Cadence through archival supports unlimited retention after a workflow closure.

SWF has Java and Ruby client libraries. Cadence has Java and Go client libraries.

SWF Java library is fully asynchronous and relies on both code generation (through annotation processor) and AspectJ. It is hard to set up, doesn't play well with IDEs and has very steep learning curve. Cadence Java library (as well as Go one) allow writing workflows as synchronous programs which greatly simplifies the programming model. It also just a library without any need for code generation or AspectJ or similar intrusive technologies.

Cadence client side libraries have much better unit testing support. For example the Java library utilizes an in-memory implementation of the Cadence service.

Cadence features that SWF doesn't have:

Workflow stickiness. SWF replays the whole workflow history on every decision. Which means that a workflow resource usage is proportional to O(n*n) of number of events in the history. Cadence caches workflows on a worker and delivers only new events to them. The whole history is replayed only when a worker goes down or the workflow gets out of cache. So Cadence workflow resource usage is O(n) of number of events in the history. For large workflows it makes a huge difference. It also leads to higher per workflow scale. For example it is not recommended to have workflows that execute over a hundred activities in SWF. Cadence routinely executes workflows that have over thousand activities or child workflows.

Query workflow execution. It allows synchronously get any information out of a workflow. An example of a built-in query is a stack trace of a running workflow.

Cross region (in AWS terminology) replication. SWF in each region is fully independent and if the regional SWF is down all workflows in the region are stuck. Cadence supports asynchronous replication across regions. So even in the event of a complete loss of a region the workflows continue execution without interruption.

Server side retry is an ability to retry an activity or a workflow according to an exponential retry policy without growing the history size.

Reset is an ability to restart a workflow from any point of its execution by creating a new run and copying a part of the history. For example the reset is used to automatically roll back workflows to the point before a bad deployment that was rolled back.

Cron is an ability to schedule a periodic workflow execution by passing cron string to the start method.

Local activity is a short activity that is executed in the context of a decision. It uses 6x less DB operations that a normal activity execution.

Long poll on history allows to efficiently watch for new history events and is also used for efficiently waiting for a workflow completion.

Cadence uses the elastic search for visibility. Soon it is going to support complex searches across multiple customer defined columns which is far superior to the tag based search SWF supports.

If decider constantly fails during a decision SWF records a few events on every failure eventually growing the history beyond the limit and terminating a workflow. Cadence supports transient decision feature that doesn't grow history on such failures. It allows continuing workflows without a problem after the fix to the workflow code is deployed.

Cadence provides command line interface

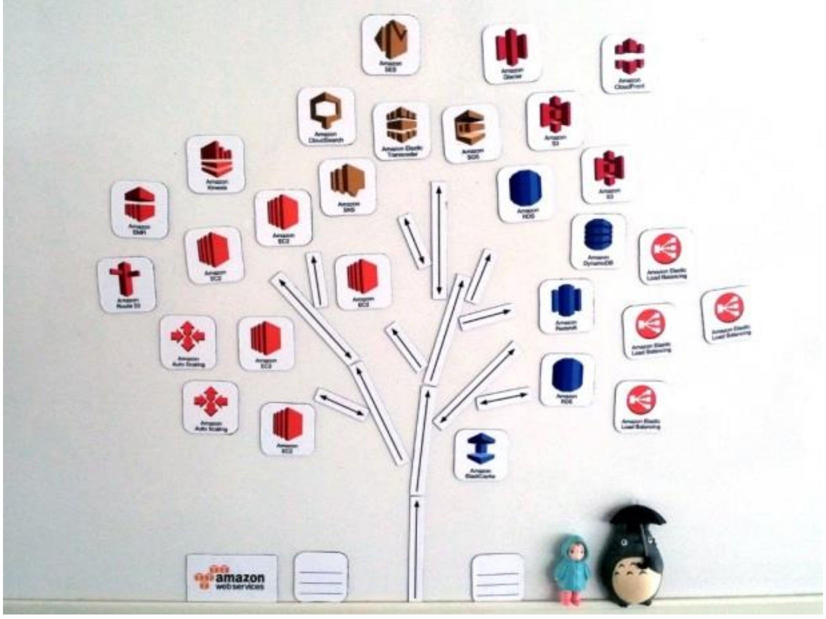
Cadence Web is open sourced and is much nicer than the SWF console.

Cadence supports local development through unit testing as well as using local docker container that contains the full implementation of the Cadence service and the UI.

Cadence doesn’t yet have activity and workflow type registration. The advantage is that changes to activity or workflow scheduling options do not require version bumps that affect clients.

Two Years with Amazon Simple Workflow (SWF)

Posted on **June 18, 2016**



June 12 mark two years of us using Amazon Simple Workflow Service (SWF) in production, and I thought I'd share the experience.
First, let's get this out of the way:

What is SWF not?

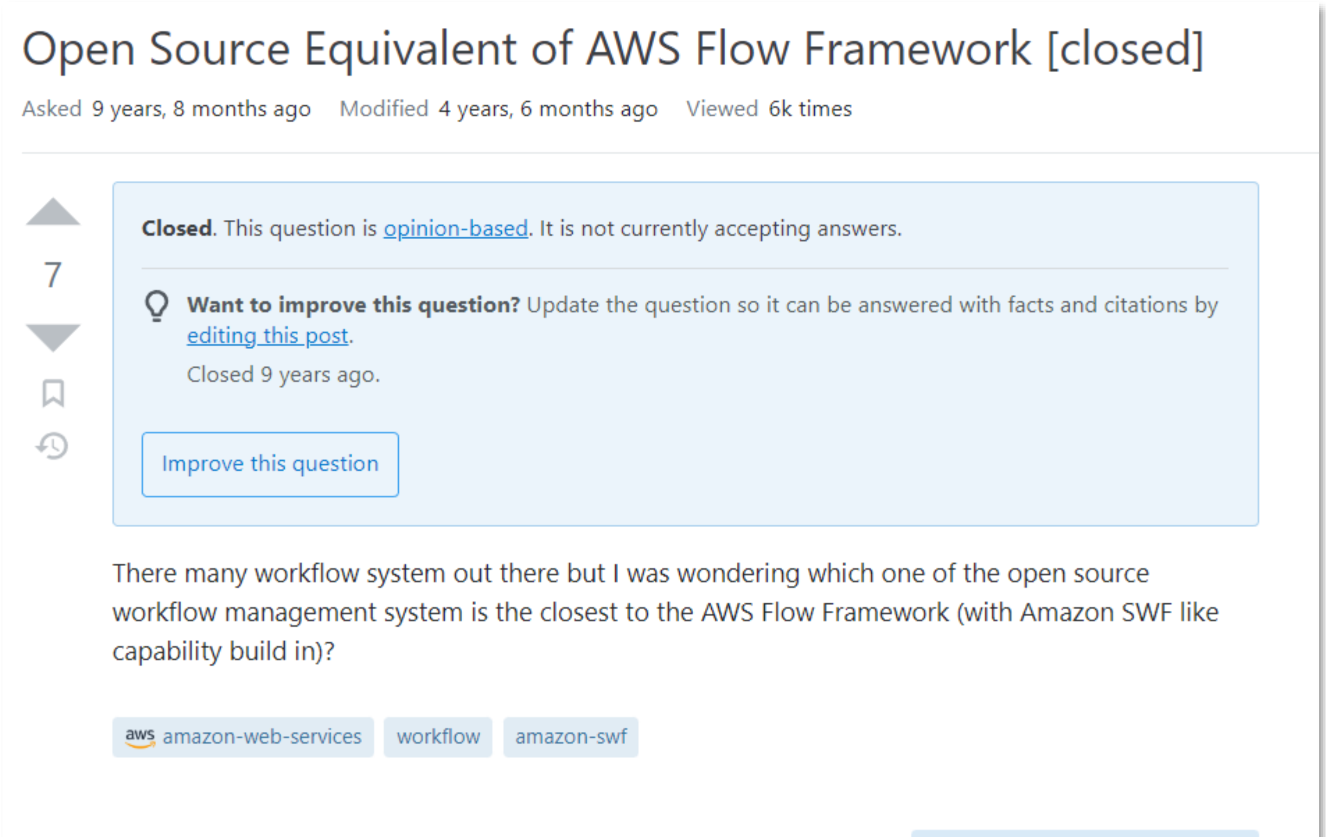
- SWF does not execute any code.
- SWF does not contain the logic of the workflow.
- SWF does not allow you to draw a workflow or a state machine.

So what is it?

SWF is a web service that keeps the **state** of your workflow. That's pretty much it.

What are we using it for?

Our project is based on C#. We are using the AWS API directly (using the .Net SDK).
If you are using Java Or Ruby amazon provider a higher level library for SWF called Flow Framework. For C#, I wrote what I needed myself, or simply used the “low level” API.
Out project processes a large number of files daily, and it was my task to convert our previous batch-based solution to SWF.



Open Source Equivalent of AWS Flow Framework [closed]

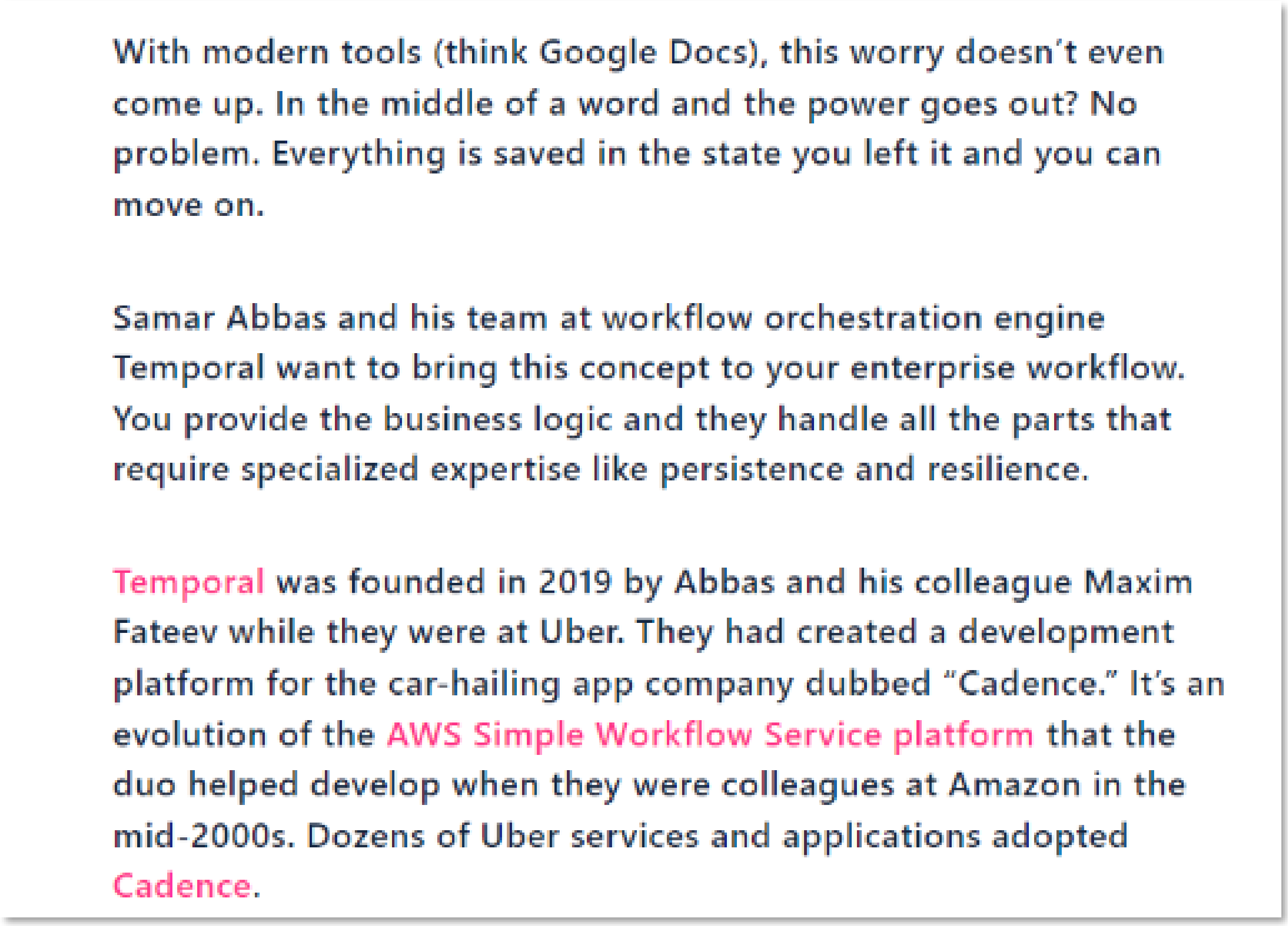
Asked 9 years, 8 months ago Modified 4 years, 6 months ago Viewed 6k times

Closed. This question is opinion-based. It is not currently accepting answers.

Want to improve this question? Update the question so it can be answered with facts and citations by editing this post.

Closed 9 years ago.

There many workflow system out there but I was wondering which one of the open source workflow management system is the closest to the AWS Flow Framework (with Amazon SWF like capability build in)?



With modern tools (think Google Docs), this worry doesn't even come up. In the middle of a word and the power goes out? No problem. Everything is saved in the state you left it and you can move on.

Samar Abbas and his team at workflow orchestration engine Temporal want to bring this concept to your enterprise workflow. You provide the business logic and they handle all the parts that require specialized expertise like persistence and resilience.

Temporal was founded in 2019 by Abbas and his colleague Maxim Fateev while they were at Uber. They had created a development platform for the car-hailing app company dubbed "Cadence." It's an evolution of the AWS Simple Workflow Service platform that the duo helped develop when they were colleagues at Amazon in the mid-2000s. Dozens of Uber services and applications adopted Cadence.

Asuman Dogac (author of the helpful 1998 book). Turkish computer scientist.

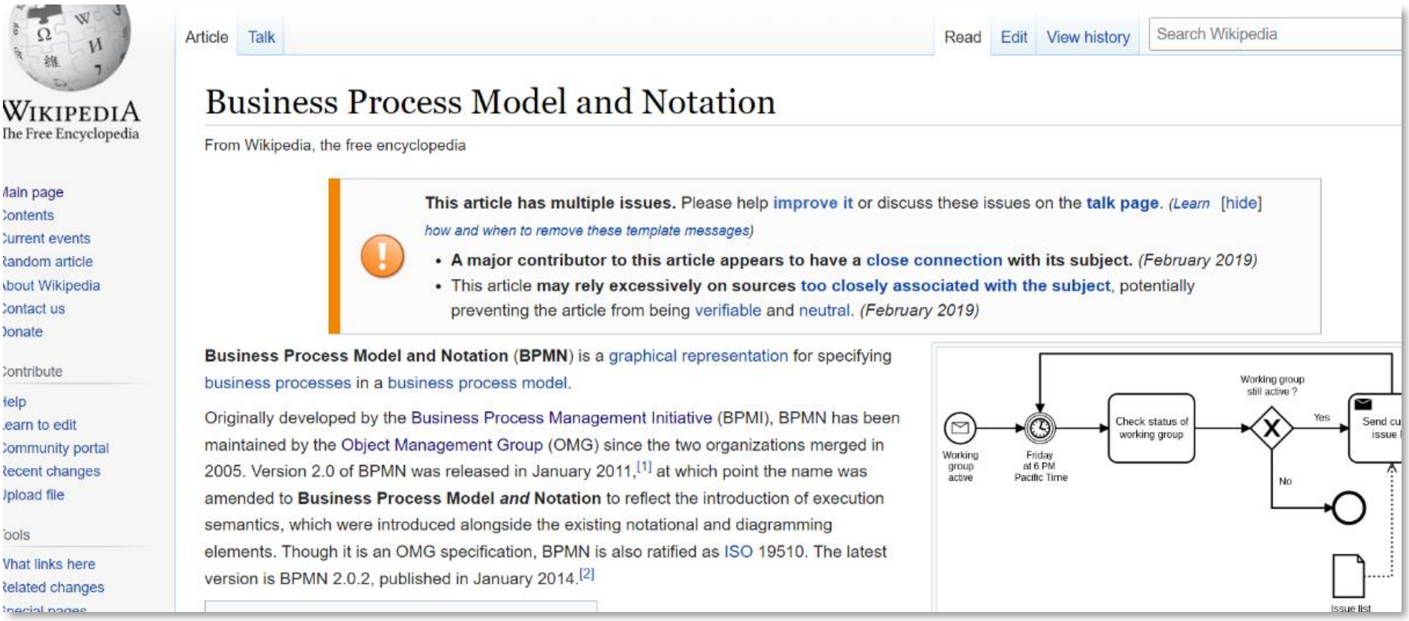
Carl Adam Petri (1926 – 2010)

A German mathematician, he invented the Petri net for the purpose of describing chemical processes.

Richard Soley, the chairman of OMG (the Object Management Group). Why are we interested in the Object Management Group? Because this standards consortium adopted BPMN in 2006 as a standard.

BPMN stands for Business Process Management Notation.

You may also be interested to know that in 2011, OMG formed the Cloud Standards Customer Council (CSCC).



Business Process Model and Notation

From Wikipedia, the free encyclopedia

This article has multiple issues. Please help **improve it** or discuss these issues on the **talk page**. *(Learn how and when to remove these template messages)* [hide]

- **A major contributor to this article appears to have a close connection with its subject.** *(February 2019)*
- This article **may rely excessively on sources too closely associated with the subject**, potentially preventing the article from being verifiable and neutral. *(February 2019)*

Business Process Model and Notation (**BPMN**) is a graphical representation for specifying business processes in a business process model.

Originally developed by the Business Process Management Initiative (BPMI), BPMN has been maintained by the Object Management Group (OMG) since the two organizations merged in 2005. Version 2.0 of BPMN was released in January 2011,[1] at which point the name was amended to **Business Process Model *and* Notation** to reflect the introduction of execution semantics, which were introduced alongside the existing notational and diagramming elements. Though it is an OMG specification, BPMN is also ratified as ISO 19510. The latest version is BPMN 2.0.2, published in January 2014.[2]

TPN

121. ***Phenomenon1*** – the tendency of X to Y.
122. ***Phen2*** – the tendency of X to Y.
123. ***Phen3*** – the tendency of X to Y.
124. ***Phen4*** – the tendency of X to Y.
125. ***Phen5*** – the tendency of X to Y.
126. ***Phen6*** – the tendency of X to Y.
127. ***Phen7*** – the tendency of X to Y.
128. ***Phen8*** – the tendency of X to Y.
129. ***Phen9*** – the tendency of X to Y.
130. ***Phen10*** – the tendency of X to Y.

Glossary

Business logic

Description of what term means here.

Business process

Description of what term means here.

BPM

This stands for Business Process Management.

Decider

Description of what term means here.

Worker

Description of what term means here.

Activity worker

Description of what term means here.

Task

Description of what term means here.

Workflow

“The decider and activities form a "workflow."” [InfoQ article 2012]

Bibliography

I. Official

[Subramanian 2012]

Subramanian, Balan (2012). Amazon Simple Workflow. *AWS Report with Jeff Barr*. 9th Oct 2012. YouTube channel: Amazon Web Services. Available at: <https://www.youtube.com/watch?v=y7Mff1ceypo&ab_channel=AmazonWebServices>

[Subramanian 2012]

Subramanian, Balan (2012). YouTube Channel: Amazon Web Services [Webinar]. Available at: <https://www.youtube.com/watch?v=lBUQiek8Jqk&ab_channel=AmazonWebServices>

[AWS 2013]

AWS. "7 Uses Cases in 7 Minutes Each". AWS Reinvent conference 2013. Las Vegas. Available at: <https://www.youtube.com/watch?v=sXGlQruUrWE&ab_channel=AmazonWebServices>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

[Barr 2012]

Barr, Jeff (2012). Amazon Simple Workflow – Cloud-Based Workflow Management. *AWS News Blog*. Feb 21st 2012. Available at: <https://aws.amazon.com/blogs/aws/amazon-simple-workflow-cloud-based-workflow-management/>

[Vogels 2012]

Vogels, Werner (2012). Expanding the Cloud – The Amazon Simple Workflow Service. *All Things Distributed* [Blog]. Feb 22nd 2012. Available at:
<https://www.allthingsdistributed.com/2012/02/amazon-simple-workflow-service.html>

[Varia 2014]

Automate your big data workflows. Available at:
https://www.youtube.com/watch?v=YQ9Y7HlxGmU&ab_channel=AmazonWebServices

https://www.youtube.com/watch?v=Z_dvXy4AVEE&ab_channel=AmazonWebServices

II. Unofficial

[Seymour 2017]

Seymour, Warren (2017). Simple Workflow Service (SWF) – Ugly Ducking or Beautiful Swan? [AWS South Wales User Group]

Available at: <https://www.youtube.com/watch?v=i_bH-PBeXl8&ab_channel=HackingInsider>

https://stackoverflow.com/questions/14986905/amazon-swf-at-least-one-worker-has-to-be-running-why?rq=1

[Brazeal 2017]

Brazeal, Forrest (2017). Serverless Workflows on AWS. Published on YouTube 18th May 2017. YouTube Channel: Serverlessconf. Available at: <https://www.youtube.com/watch?v=D6qV3bC4rNw&ab_channel=Serverlessconf>

[Lublinksky 2012]

Lublinsky, Boris (2012). Amazon Provides Simple Workflow Service Recipes. *InfoQ*. Nov 16th. Available at: <https://www.infoq.com/news/2012/11/swfrecipes/>

III. Critical

IV. General

[Aalst 2003]

Aalst, Wil van der and Hajo Reijers and Selma Limam. Product-Based Workflow Design. In *Journal of Management Information Systems* 20(1): 229-262. DOI:10.1080/07421222.2003.11045753

[Aalst 1997]

van der Aalst, W.M., Li, K., Olariu, S., Pan, Y., & Stojmenovic, I. (1997). Designing Workflows based on product structures. Journal Name.

[Abollado 2017]

Abollado, J Rojo and Shahab and Bamforth (2017). Challenges and Benefits of Digital Workflow Implementation in Aerospace Manufacturing Engineering.

[Ancona 2020]

Ancona, Adrian (2020). Introduction to Simple Workflow Service (SWF)". Available at: https://ncona.com/2020/07/introduction-to-aws-simple-workflow-service/

[Avram 2012]

Avram, Abel (2012). Is Amazon Getting Ready for PaaS with Simple Workflow Service? *InfoQ*. Available at: <https://www.infoq.com/news/2012/02/Amazon-PaaS-SWF/>

[Becker 2002]

Becker, Jörg and Michael zur Muehlen and Marc Gille (2002). "Workflow Application Architectures: Classification and Characteristics of Workflow-based Information Systems". In Fischer, L. (ed.). Workflow Handbook 2002. Lighthouse Point, FL: Future Strategies. CiteSeerX 10.1.1.24.2311.

[Caverlee 2007]

Caverlee, James et al. Workflow management for enterprise transformation. *Information Knowledge Systems Management* 6: 61-80 (2007).

[Chen 2012]

Chen, Huei-Huang et al (2012). An Overview of Workflow Management System Structures in the supply chain. *Australian Journal of Business and Management Research*. Available at: https://www.ajbmr.com/articlepdf/aus-24-04i4n2a2.pdf

[Davis 2022]

Davis, Brown (2022). **What is Workflow Management? – Detailed Definition & Examples**. *MyTechMag*. July 6th 2022. Available at: <https://www.mytechmag.com/workflow-management/>

[Deloitte]

Strategy execution: What could possible go wrong? Available at: https://www2.deloitte.com/us/en/pages/finance/articles/cfo-insights-solving-for-execution-risk.html

[Dogac 1998] ☆

Dogac, A. et al. (1998). Design and Implementation of a Distributed Workflow Management System: METUFlow. In: Doğaç, A., Kalinichenko, L., Özsu, M.T., Sheth, A. (eds) Workflow Management Systems and Interoperability. NATO ASI Series, vol 164. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-58908-9_4

[FatFinger 2021]

Workflow Software: Your Digital Solution to Industry 4.0 and How to Implement It. Available at: <https://fatfinger.io/workflow-software-your-digital-solution-to-industry-4-0-and-how-to-implement-it/>

[Fischer 2017] ☆

Using AWS's Simple Workflow Service (SWF) with C#. Jan 12th 2017. Available at: <https://www.red-gate.com/simple-talk/development/dotnet-development/using-awss-simple-workflow-service-swf-c/>

[Forrester 2012]

Amazon's New Simple Workflow Service. *Forrester*. Feb 23rd 2012. Available at: <https://www.forrester.com/blogs/amazons-new-simple-workflow-service/>

[Georgakopoulos 1995] ☆

Georgakopoulos, Diimitrios and Mark Hornick and Amit Sheth (1995). An Overview of Workflow Management: From Process Modelling to Workflow Automation Infrastructure. *Distributed and Parallel Databases*, 3, 119-153.

[Globalscape]

Advanced Workflow Engine (AWE) V10. Available at: https://hstechdocs.helpsystems.com/manuals/globalscape/guides/Advanced_Workflow_Engine_v10_User_Guide.pdf

Note on the above bibliography entry:

EFT stands for Enhanced File Transfer. Globalscape is a company which develops software. It is headquartered in San Antonio, Texas. Here is a list of five products which GlobalScape sell: (1) DMZ Gateway (2) CuteFTP (3) Business Activity Monitoring, or BAM (4) Advanced Security Module, or ASM, (5) Web Transfer Client. Another product is the Advanced Workflow Engine (AWE). Globalscape was founded in 1996, as a wholly owned subsidiary of American Telesource Incorporated (ATSI). The firm's original product, released in 1996, was CuteFTP. This was a File Transfer Protocol client application for Windows and Mac platforms.

[Globalscape 2]

Globalscape (2016). How to Create a Workflow in the Advanced Workflow Engine and Trigger CuteFTP. YouTube Channel: Globalscape – Secure Information Exchange Solutions. Available at: <https://www.youtube.com/watch?v=ttv-0ktQRN0&ab_channel=Globalscape-SecureInformationExchangeSolutions>

[Goble 2020]

Goble, Carole et al. FAIR Computational Workflows. Available at: https://direct.mit.edu/dint/article/2/1-2/108/10003/FAIR-Computational-Workflows

Note on the above bibliography entry

Carole Goble is a British academic. She is Professor of Computer Science at the University of Manchester. Goble was given an award in 2008 for her work on Apache Taverna (the Jim Gray e-Science Award).

[Gruber 2009]

Gruber, Horst and Christian Hueber (2009). Profitability Analysis of Workflow Management Systems. IEEE Conference on Commerce and Enterprise Computing (2009). Available at: https://publik.tuwien.ac.at/files/PubDat_183990.pdf

[Hacker 2019 a]

Amazon Simple Workflow compared to Cadence. https://news.ycombinator.com/item?id=19733784

[Hacker 2019 b]

https://news.ycombinator.com/item?id=19732447

[Hacker 2020] ☆

Hacker News [Comment Thread] Available at: https://news.ycombinator.com/item?id=23844177

[Hee 2013]

"Kees van Hee interviewed at AEngD launch". YouTube Channel: TheAEngD. Available at: https://www.youtube.com/watch?v=x8Q_BKSFECw&ab_channel=TheAEngD

[Hee 2000]

Hee, Kees van (2000). Workflow Management: Models, Methods, and Systems. https://doi.org/10.7551/mitpress/7301.001.0001

[HighGear 2021]

What is workflow software and Why do You Need It? *High Gear*. March 11th 2021. Available at: <https://www.highgear.com/blog/what-is-workflow/#:~:text=Workflow%20systems%20ensure%20that%20work,building%20value%20in%20managing%20workflows>

[Jackson 2012]

Jackson, Joab (2012). Amazon launches workflow orchestration service. Available at: <https://www.computerworld.com/article/2493627/amazon-launches-workflow-orchestration-service.html>

[Janakiram 2012]

Amazon Has Got a Winner in the Form of SWF. *YourStory*. Available at: <https://yourstory.com/2012/02/amazon-has-got-a-winner-in-the-form-of-swf>

[Janetschek 2013]

Janetschek, Matthias and Simon Ostermann and Radu Prodan (2013). Bringing Scientific Workflows to Amazon SWF. 2013 39th Euromicro Conference on Software Engineering and Advanced Applications, 2013, pp. 389-396, doi: 10.1109/SEAA.2013.13.

[Jiaqi 2012]

A dozen things to know about AWS Simple Workflow in Eclipse and Maven. Jiaqi's Blog [Blog]. Available at: <https://blog.cyclopsgroup.org/2012/12/a-dozen-things-to-know-about-aws-simple.html>

[Johnson 2016]

Johnson, Chris (2016). AWS Simple Workflow: Overview and Best Practices by Chris Johnson. YouTube Channel: Local Variables. Available at:

https://www.youtube.com/watch?v=3Nni7H5wWBA&ab_channel=LocalVariables

[Juve 2010]

Juve, Gideon et al (2010). Data Sharing Options for Scientific Workflows on Amazon EC2. Available at: https://deelman.isi.edu/wordpress/wp-content/papercite-data/pdf/juve2010sc.pdf

[Kobi 2016]

Two Years with Amazon Simple Workflow (SWF). Available at: https://kobikobi.wordpress.com/2016/06/18/two-years-with-amazon-simple-workflow-swf/

[Lim 2018]

Lim, Michael (2018). Workflow automation: 25 years of tried-and-true success. IBM Blog. Available at: https://www.ibm.com/blogs/cloud-computing/2018/12/18/workflow-automation-25-years/

Beginning with m

[McCarthy 2012]

McCarthy, Jack (2012). Amazon Simplifies Workflow with New Cloud Service. *CRN*. Available at: https://www.crn.com/news/cloud/232601300/amazon-simplifies-workflow-with-new-cloud-service.htm

[Marinescu 2018]

Marinescu, Dan C. (2018). *Cloud Applications. Cloud Computing, 237–279.* doi:10.1016/b978-0-12-812810-7.00010-8

[Monk 2013]

Introduction to AWS SimpleWorkflow Extensions. Part 1 – Hello World Example. Available at: <https://theburningmonk.com/2013/02/introduction-to-aws-simpleworkflow-extensions-part-1-hello-world-example/>

[McFadin 2022]

McFadin, Patrick (2022). The Best Way to Think about Resilience Is Not to. *The New Stack*. Oct 31st 2022. Available at: <https://thenewstack.io/the-best-way-to-think-about-resilience-is-not-to/>

[Miratech 2020]

"What is Workflow Management?" December 14th 2020. Available at: <https://mitratech.com/resource-hub/blog/what-is-workflow-management/>

[Mohan 1998] ☆

Mohan, C. (1998). Recent Trends in Workflow Management Products, Standards and Research. In: Doğaç, A., Kalinichenko, L., Özsu, M.T., Sheth, A. (eds) Workflow Management Systems and Interoperability. NATO ASI Series, vol 164. Springer, Berlin, Heidelberg. Available at: https://doi.org/10.1007/978-3-642-58908-9_17

[Mehta 2012]

Mehta, Chirag (2012). Simple Workflow Service – Amazon Adding One Enterprise Brick at Time. Available at: <https://www.cloudave.com/17358/simple-workflow-service-amazon-adding-one-enterprise-brick-at-time/>

[Mueller 2003]

Mueller, Bob (2003). Case Study: Workflow Management. *Enterprise Systems Journals*. Available at: <https://esj.com/articles/2003/07/15/case-study-workflow-management.aspx?m=1>

Beginning with n

[Nissanka 2017]

Nissanka, Asanka (2017). Building Workflows with Amazon Simple Workflow Versus Step Functions. Second edn 2019. Available at: <https://medium.com/avmconsulting-blog/building-workflows-with-amazon-simple-workflow-service-vs-step-functions-83fdeac35555>

Beginning with o

[Olavsrud 2022]

Olavsrud, Thor and Clint Boulton (2022). What is RPA? A revolution in business process automation. Available at: <https://www.cio.com/article/227908/what-is-rpa-robotic-process-automation-explained.html#:~:text=What%20is%20robotic%20process%20automation,aimed%20at%20automating%20business%20processes>

[Overflow X]

Amazon AWS Simple Workflow Service SWF C# Sample. Question asked April 17th 2013. Available at: <https://stackoverflow.com/questions/16051441/amazon-aws-simple-workflow-service-swf-c-sharp-sample>

[Overflow 2013]

Open Source Equivalent of AWS Flow Framework. Stack Overflow [Forum]. Available at: <https://stackoverflow.com/questions/15236575/open-source-equivalent-of-aws-flow-framework>

[Overflow 1]

https://stackoverflow.com/questions/16158671/aws-whats-the-difference-between-simple-workflow-service-and-data-pipeline

[Patel 2019]

Patel, Ashish (2019). Difference between SQS and SWF. *Medium.* Available at: <https://medium.com/awesome-cloud/aws-difference-between-sqs-and-swf-7a0954999621>

[Peck 2020]

Peck, Holly (2020). What is Workflow Management? (Benefits, Process, Tools). *Clickup.com*. Available at: <https://clickup.com/blog/workflow-management/>

[Poola 2017]

Poola, Deepak and Mohsen Amini Salehi, Kotagiri Ramamohanarao and Rajkumar Buyaa (2017). "A Taxonomy and Survey of Fault-Tolerant Workflow Management Systems in Cloud and Distributed Computing Environments". Chapter 15 in *Software Architecture for Big Data and the Cloud*. 10.1016/B978-0-12-805467-3.00015-6

[Quora 1]

What is the best workflow management software if I'm especially interested in open source solutions and web-based software? Available at: <https://www.quora.com/What-is-the-best-workflow-management-software-if-I%E2%80%99m-especially-interested-in-open-source-solutions-and-web-based-software>

[Quora 2]

What's the best workflow tool? Available at: <https://www.quora.com/Whats-the-best-workflow-tool>

[Quora 3]

Does Amazon internally use the Amazon Simple Workflow service to manage the order life-cycle aka something similar to an Order Management System? Available at: <https://www.quora.com/Does-Amazon-internally-use-Amazon-Simple-Workflow-service-to-manage-the-order-life-cycle-aka-something-similar-to-an-Order-Management-System>

[Rao 2011]

Rao, Leena (2011). Joel Spolsky's Trello is a Simple Workflow and List Manager for Groups. *TechCrunch*. Available at: <https://techcrunch.com/2011/09/13/joel-spolskys-trello-is-a-simple-workflow-and-list-manager-for-groups/>

[Reddit 1]

"Airflow vs AWS?" Question asked in 2020. Available at: https://www.reddit.com/r/dataengineering/comments/gq9bax/airflow_vs_aws/

[Reddit 2016]

Can someone please explain how SWF works using a real-life example? Available at: <https://www.reddit.com/r/aws/comments/657yac/can_someone_please_explain_how_swf_works_using_a/>

[Rigby 2016]

Rigby, Darrell and Jeff Sutherland and Hirotaka Takeuchi (2016). Embracing Agile. *Harvard Business Review* [Magazine]. Available at: https://hbr.org/2016/05/embracing-agile

[Rodriguez 2017] ☆

Rodriguez, Maria and Rajkumar Buyya (2017). "Scientific Workflow Management System for Clouds". Chapter 18 in *Software Architecture for Big Data and the Cloud*. Available at: https://doi.org/10.1016/B978-0-12-805467-3.00018-1

[Rossi 2012]

Rossi, Ben (2012). AWS launches cloud-workflow automation service. Feb 22nd 2012. Available at: <https://www.information-age.com/aws-launches-cloud-workflow-automation-service-26966/>

[Sambandam 2019]

Sambandam, Suresh (2019). The Past and Future of Workflow Automation. March 29th 2019. Available at: <https://workflowotg.com/the-past-and-future-of-workflow-automation/>

[Schmidt 2007]

Brahe, Steen and Kjeld Schmidt (2007). The Story of a Working Workflow Management System. GROUP'07 - Proceedings of the 2007 International ACM Conference on Supporting Group Work. 249-258. Available at: https://doi.org/10.1145/1316624.1316661

[Scotti 2012]

Scotti, Anthony (2012). *128 bit*. Available at: https://128bit.io/2012/07/29/amazon-simple-workflow-services/

[Silva 2017]

Silva, Rafael Ferreira da and Rosa Filgueira et al. (2017). **A Characterization of Workflow Management Systems for Extreme-Scale Applications**. Future Generation Computer Systems. 75: 228-238. Available at: https://doi.org/10.1016/j.future.2017.02.026

[Stack Share]

"Alternatives to Amazon SWF". Available at: <https://stackshare.io/amazon-swf/alternatives>

[Tavaxy]

https://tavaxy.org/

[WebProNews 2012]

Amazon: Simple Workflow Service Webinar March 13th. *WebProNews*. Feb 12th 2012. Available at: <https://www.webpronews.com/amazon-simple-workflow-service-webinar-march-13th/>

[Wikipedia 1]

"Petri Net". Wikipedia [Online]. Available at: https://en.wikipedia.org/wiki/Petri_net

[Wikipedia 2]

Carl Adam Petri. *Wikipedia* [Online]. Available at: <https://en.wikipedia.org/wiki/Carl_Adam_Petri>

[Wikipedia 3]

"Workflow Engine". *Wikipedia* [Online]. Available at: <https://en.wikipedia.org/wiki/Workflow_engine>

[Wikipedia 4]

"Workflow management system". *Wikipedia* [Online]. Available at: https://en.wikipedia.org/wiki/Workflow_management_system

[Wikipedia 5]

"Workflow application". *Wikipedia* [Online]. Available at: https://en.wikipedia.org/wiki/Workflow_application

[Wikipedia 6]

Business Process Model and Notation. Available at: <https://en.wikipedia.org/wiki/Business_Process_Model_and_Notation>

[Wikipedia 7]

"JBPM". Available at: <https://en.wikipedia.org/wiki/JBPM>

[Wikipedia 8]

Object Management Group (OMG). Available at: https://en.wikipedia.org/wiki/Object_Management_Group#History

[Wired 2012]

Amazon Blurs Cloud Lines with Simple Workflow Service. *Wired*. Available at: <https://www.wired.com/insights/2012/02/simple-workflow-service/>

Beginning with x

Beginning with y

Beginning with z

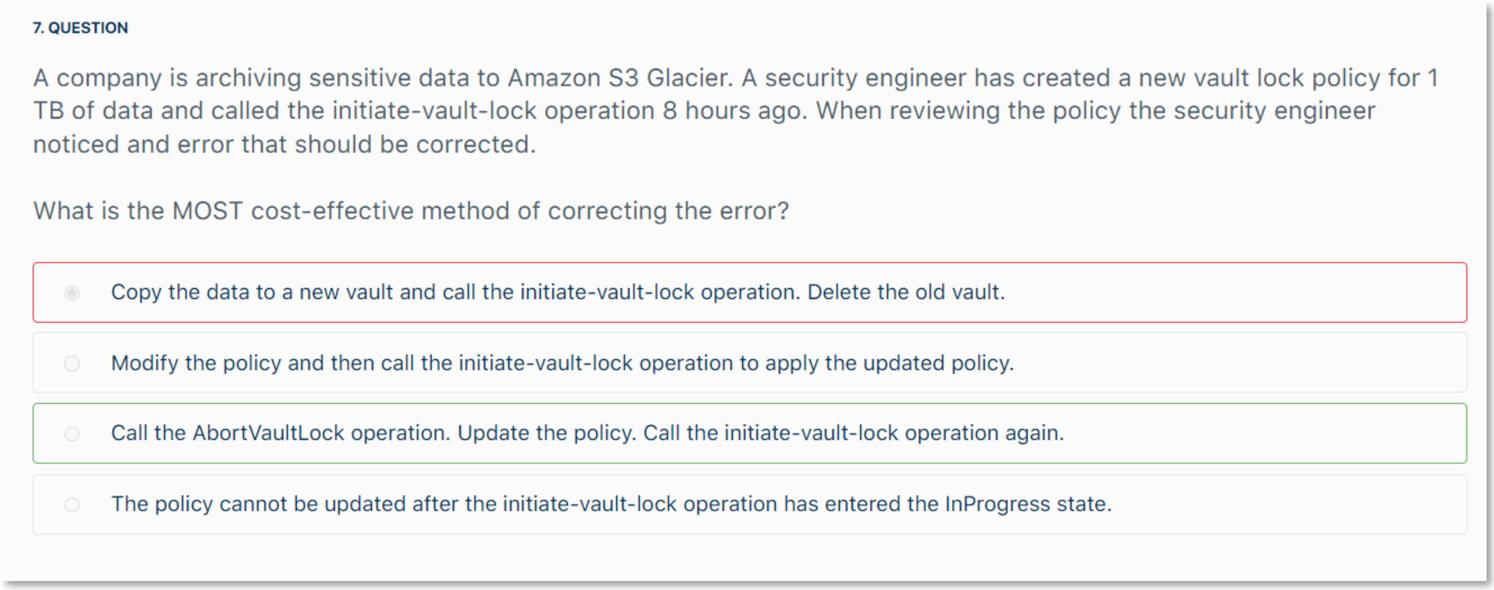
[Zhao 2015]

Zhao, Yong and Youfu Li and Ioan Raicu and Wenhong Tian and Heng Liu. Enabling scalable scientific workflow management in the cloud. *Future Generation Computer Systems*, 46, 3-16. DOI: 10.1016/j.future.2014.10.023

Glacier

←

https://www.nbcnews.com/science/environment/worlds-glaciers-contain-less-ice-thought-researchers-find-rcna15258



7. QUESTION

A company is archiving sensitive data to Amazon S3 Glacier. A security engineer has created a new vault lock policy for 1 TB of data and called the initiate-vault-lock operation 8 hours ago. When reviewing the policy the security engineer noticed and error that should be corrected.

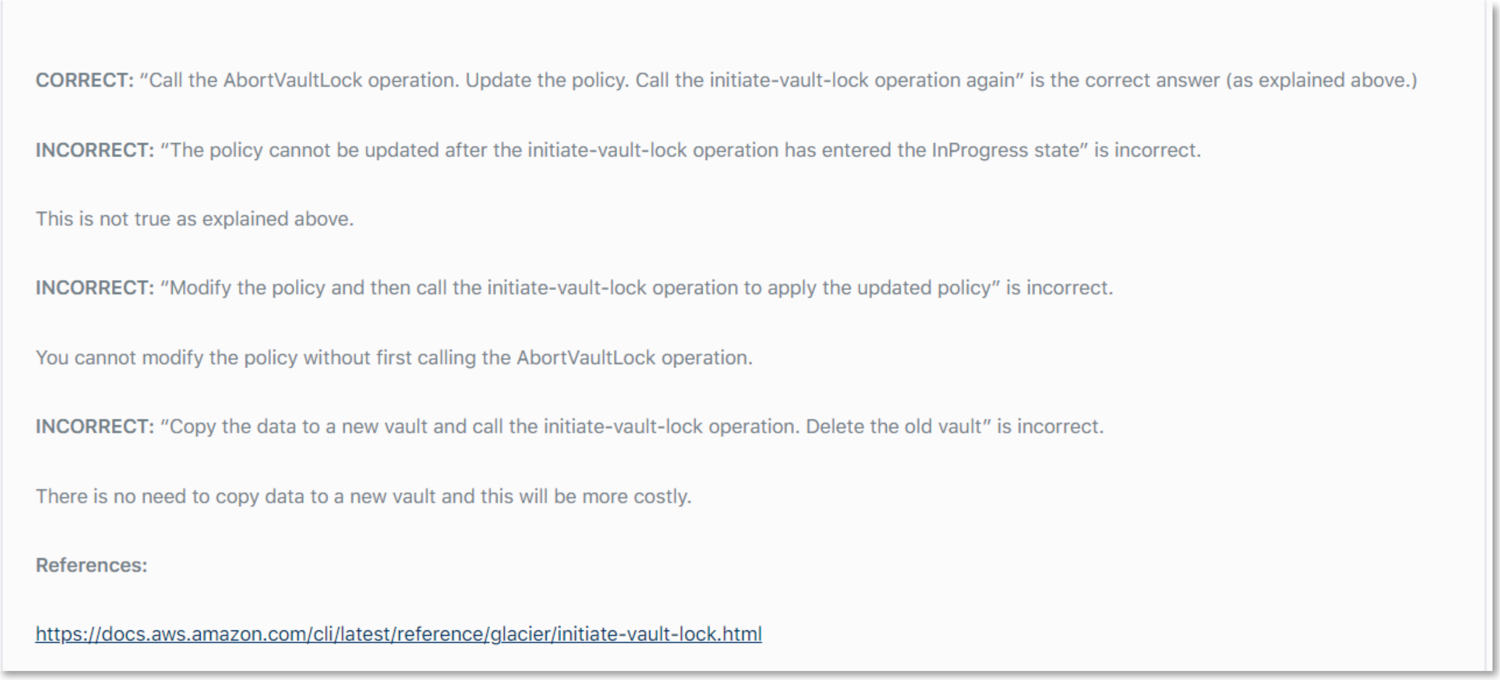
What is the MOST cost-effective method of correcting the error?

● Copy the data to a new vault and call the initiate-vault-lock operation. Delete the old vault.

○ Modify the policy and then call the initiate-vault-lock operation to apply the updated policy.

○ Call the AbortVaultLock operation. Update the policy. Call the initiate-vault-lock operation again.

○ The policy cannot be updated after the initiate-vault-lock operation has entered the InProgress state.



CORRECT: "Call the AbortVaultLock operation. Update the policy. Call the initiate-vault-lock operation again" is the correct answer (as explained above.)

INCORRECT: "The policy cannot be updated after the initiate-vault-lock operation has entered the InProgress state" is incorrect.

This is not true as explained above.

INCORRECT: "Modify the policy and then call the initiate-vault-lock operation to apply the updated policy" is incorrect.

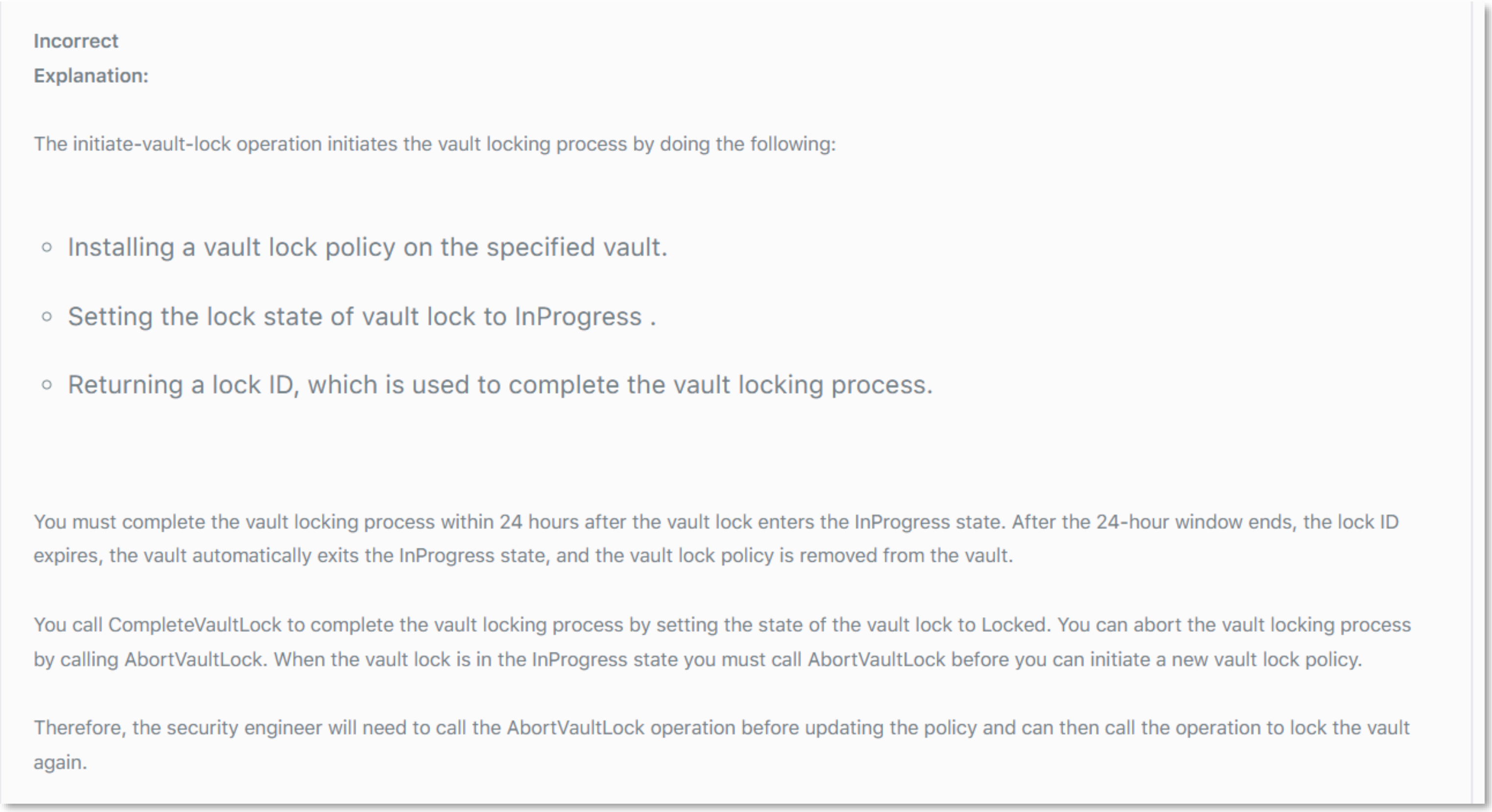
You cannot modify the policy without first calling the AbortVaultLock operation.

INCORRECT: "Copy the data to a new vault and call the initiate-vault-lock operation. Delete the old vault" is incorrect.

There is no need to copy data to a new vault and this will be more costly.

References:

https://docs.aws.amazon.com/cli/latest/reference/glacier/initiate-vault-lock.html



Incorrect

Explanation:

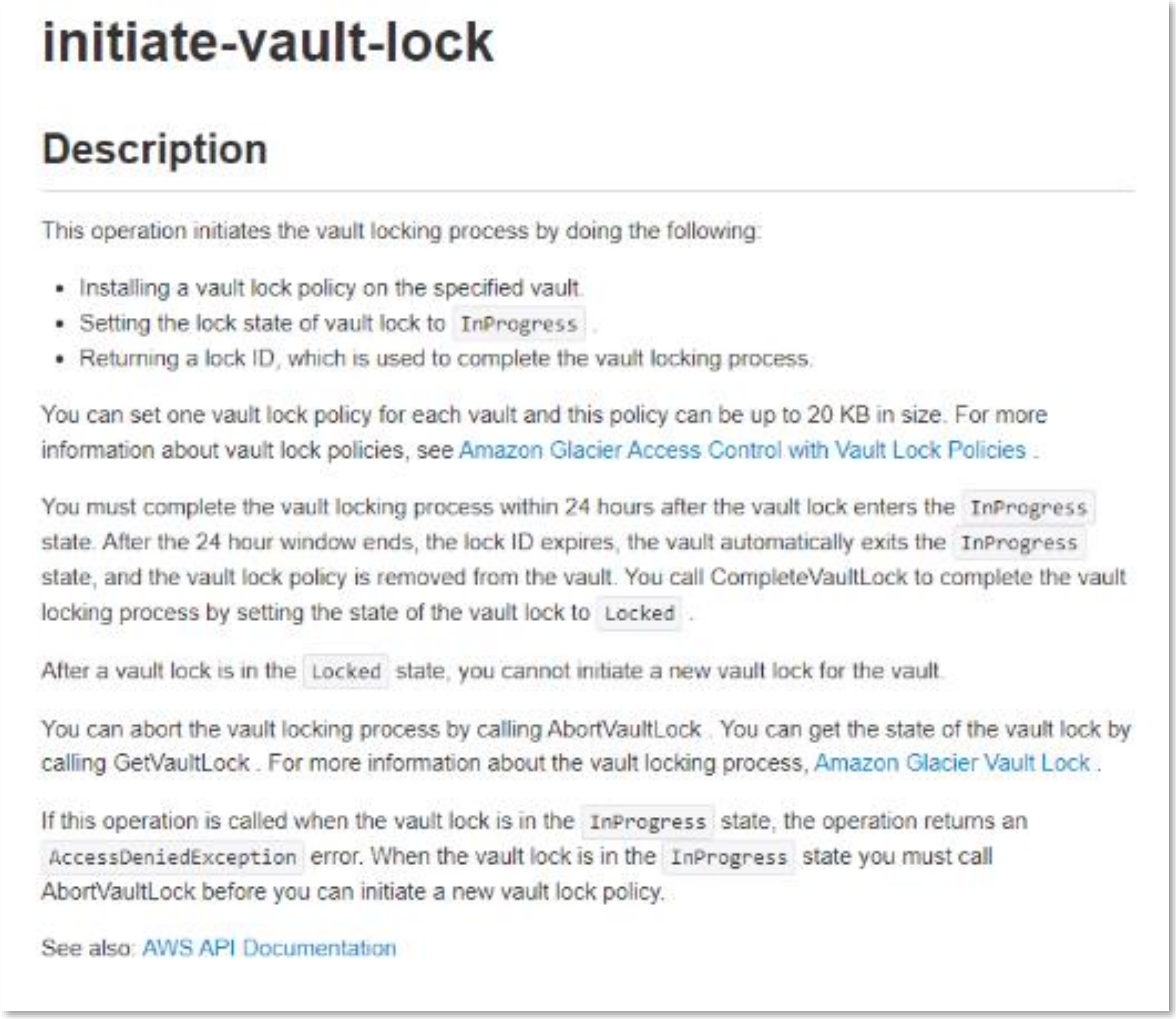
The initiate-vault-lock operation initiates the vault locking process by doing the following:

- Installing a vault lock policy on the specified vault.
- Setting the lock state of vault lock to InProgress .
- Returning a lock ID, which is used to complete the vault locking process.

You must complete the vault locking process within 24 hours after the vault lock enters the InProgress state. After the 24-hour window ends, the lock ID expires, the vault automatically exits the InProgress state, and the vault lock policy is removed from the vault.

You call CompleteVaultLock to complete the vault locking process by setting the state of the vault lock to Locked. You can abort the vault locking process by calling AbortVaultLock. When the vault lock is in the InProgress state you must call AbortVaultLock before you can initiate a new vault lock policy.

Therefore, the security engineer will need to call the AbortVaultLock operation before updating the policy and can then call the operation to lock the vault again.



initiate-vault-lock

Description

This operation initiates the vault locking process by doing the following:

- Installing a vault lock policy on the specified vault.
- Setting the lock state of vault lock to `InProgress` .
- Returning a lock ID, which is used to complete the vault locking process.

You can set one vault lock policy for each vault and this policy can be up to 20 KB in size. For more information about vault lock policies, see Amazon Glacier Access Control with Vault Lock Policies .

You must complete the vault locking process within 24 hours after the vault lock enters the `InProgress` state. After the 24 hour window ends, the lock ID expires, the vault automatically exits the `InProgress` state, and the vault lock policy is removed from the vault. You call CompleteVaultLock to complete the vault locking process by setting the state of the vault lock to `Locked` .

After a vault lock is in the `Locked` state, you cannot initiate a new vault lock for the vault.

You can abort the vault locking process by calling AbortVaultLock . You can get the state of the vault lock by calling GetVaultLock . For more information about the vault locking process, Amazon Glacier Vault Lock .

If this operation is called when the vault lock is in the `InProgress` state, the operation returns an `AccessDeniedException` error. When the vault lock is in the `InProgress` state you must call AbortVaultLock before you can initiate a new vault lock policy.

See also: AWS API Documentation

TPN

131. ***Phenomenon1*** – the tendency of X to Y.
132. ***Phen2*** – the tendency of X to Y.
133. ***Phen3*** – the tendency of X to Y.
134. ***Phen4*** – the tendency of X to Y.
135. ***Phen5*** – the tendency of X to Y.
136. ***Phen6*** – the tendency of X to Y.
137. ***Phen7*** – the tendency of X to Y.
138. ***Phen8*** – the tendency of X to Y.
139. ***Phen9*** – the tendency of X to Y.
140. ***Phen10*** – the tendency of X to Y.

Review Questions

Glossary

Term1
Description of what term means here.

Term2
Description of what term means here.

Term3
Description of what term means here.

Bibliography

I. Official

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

II. Unofficial

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

IV. General

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

CloudSearch

NOTE

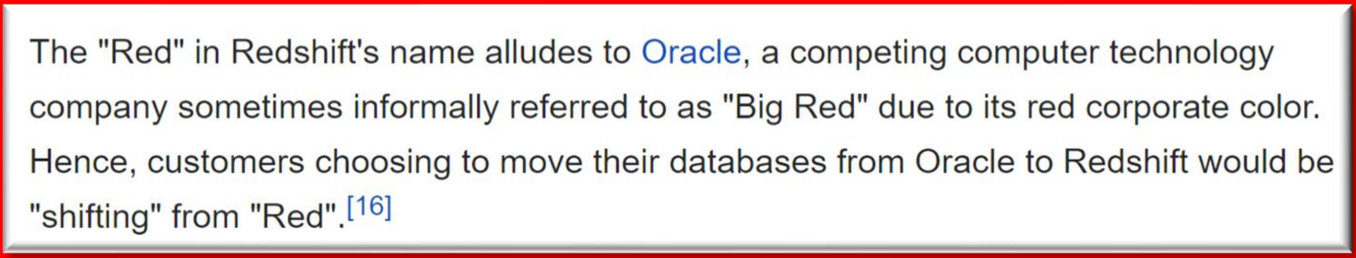
Other search services

A number of other search services are going to come after this. In 2015, AWS announce the Amazon Elasticsearch Service. Then in September 2021, they announced the Amazon OpenSearch Service. Note that there is another search service, known as Amazon Kendra. This allows enterprises (large organisations) to implement an internal search engine, to search their large volumes of documents. Kendra was announced in 2019. Kendra is a nice, distinct name, but how can we get on top of the other names?

Well, like some young librarian settling into his new job, and getting better and better at searching for books for the local citizens, he started off quite withdrawn, pursuing his passion for CS (computer science). But he starts to make friends, and eventually springs out of his shell (elastic). He becomes quite an amiable and open person in the end. So, we can quite easily remember the sequence: **C**loud**S**earch (CS), **E***lasticsearch*, and finally **O**pen**S**earch [2012, 2015, 2021].

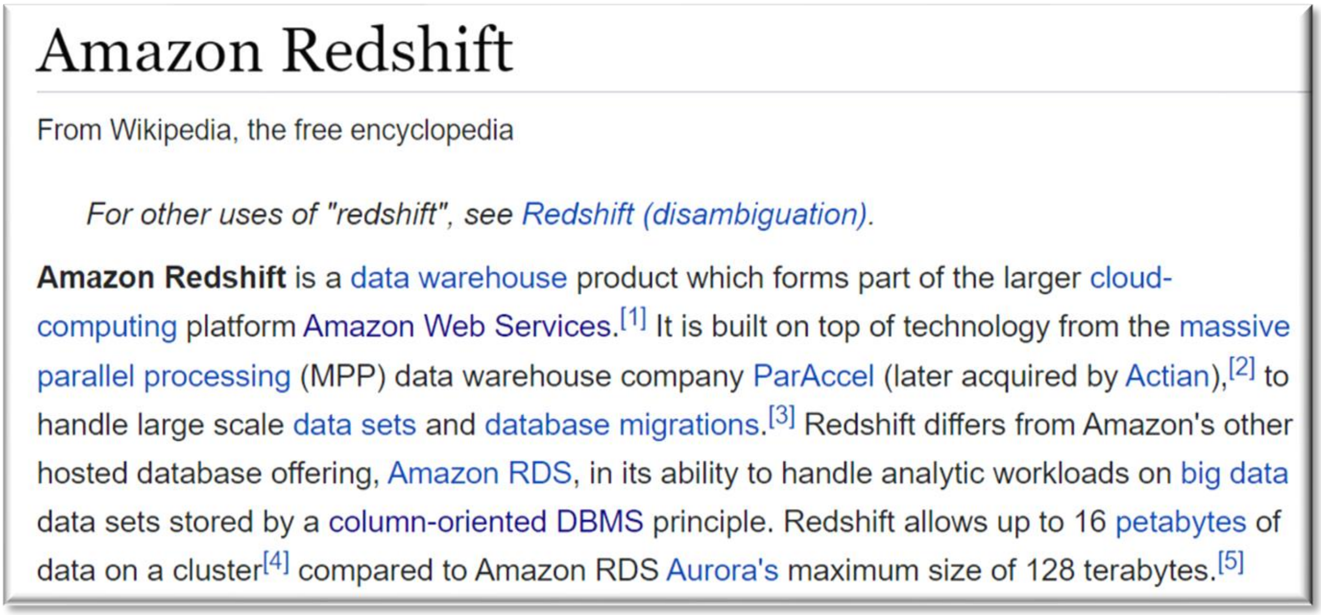
Jon Handler explaining Amazon CloudSearch in 2013

REDSHIFT



The "Red" in Redshift's name alludes to Oracle, a competing computer technology company sometimes informally referred to as "Big Red" due to its red corporate color. Hence, customers choosing to move their databases from Oracle to Redshift would be "shifting" from "Red".[16]

Redshift

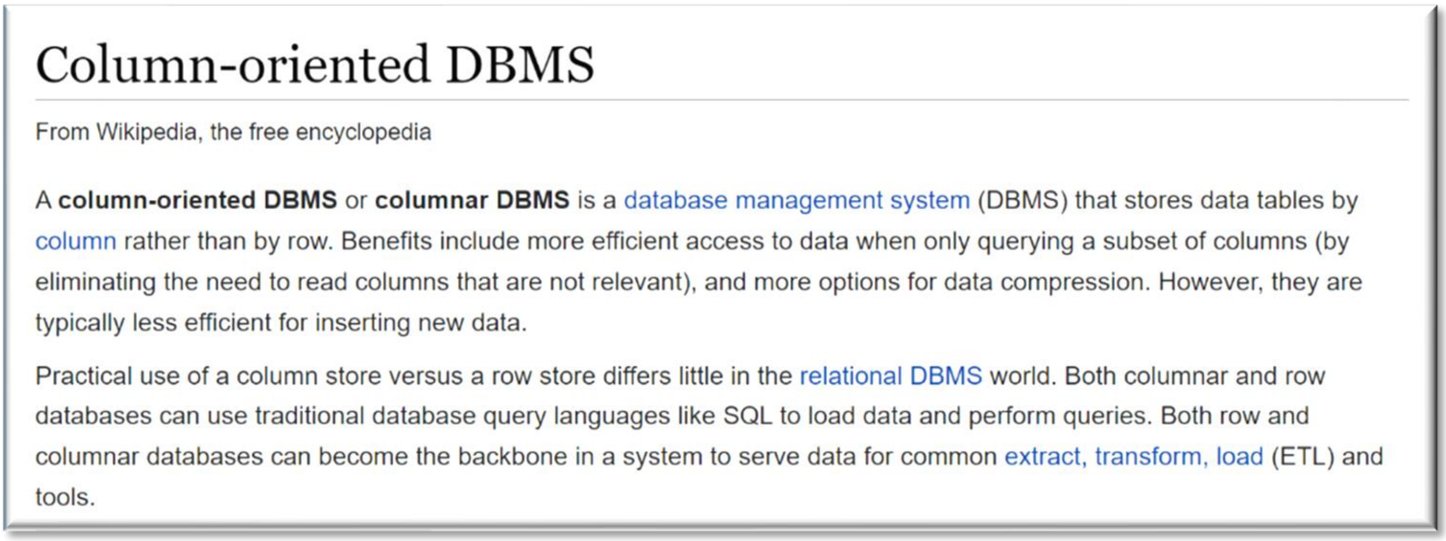


Amazon Redshift

From Wikipedia, the free encyclopedia

For other uses of "redshift", see Redshift (disambiguation).

Amazon Redshift is a data warehouse product which forms part of the larger cloud-computing platform Amazon Web Services.[1] It is built on top of technology from the massive parallel processing (MPP) data warehouse company ParAccel (later acquired by Actian),[2] to handle large scale data sets and database migrations.[3] Redshift differs from Amazon's other hosted database offering, Amazon RDS, in its ability to handle analytic workloads on big data data sets stored by a column-oriented DBMS principle. Redshift allows up to 16 petabytes of data on a cluster[4] compared to Amazon RDS Aurora's maximum size of 128 terabytes.[5]

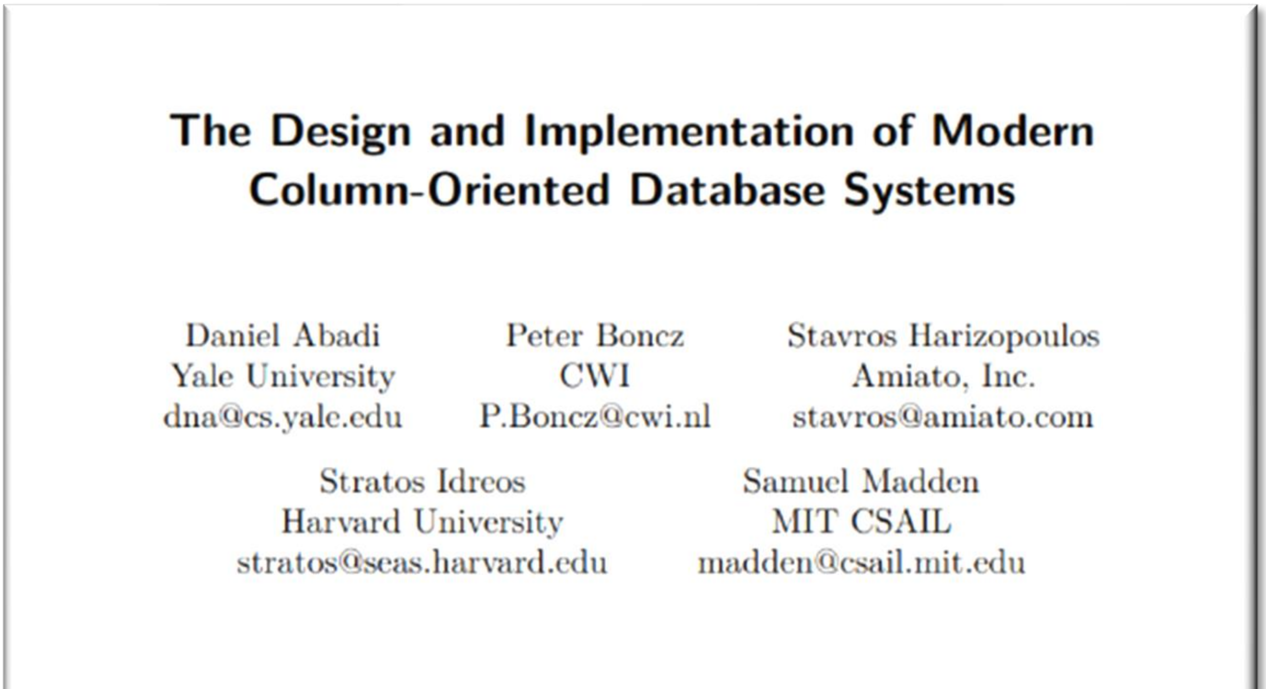


Column-oriented DBMS

From Wikipedia, the free encyclopedia

A **column-oriented DBMS** or **columnar DBMS** is a database management system (DBMS) that stores data tables by column rather than by row. Benefits include more efficient access to data when only querying a subset of columns (by eliminating the need to read columns that are not relevant), and more options for data compression. However, they are typically less efficient for inserting new data.

Practical use of a column store versus a row store differs little in the relational DBMS world. Both columnar and row databases can use traditional database query languages like SQL to load data and perform queries. Both row and columnar databases can become the backbone in a system to serve data for common extract, transform, load (ETL) and tools.



The Design and Implementation of Modern Column-Oriented Database Systems

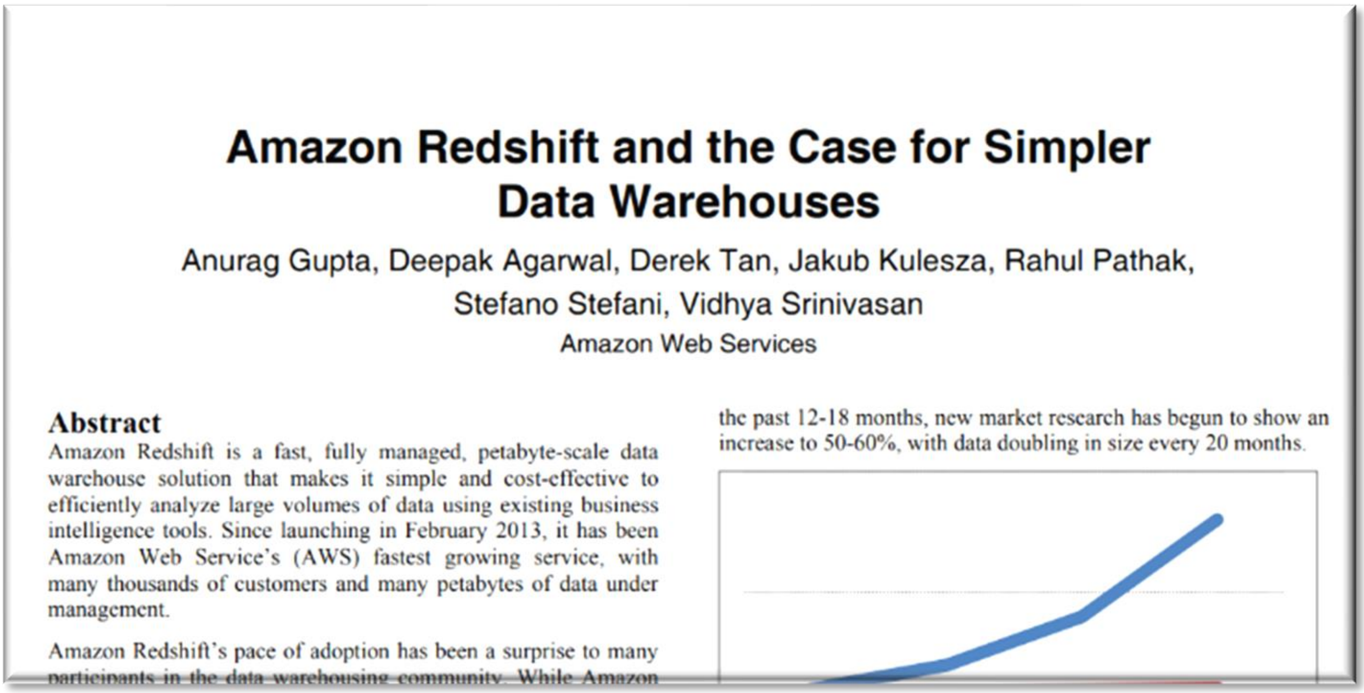
Daniel Abadi
Yale University
dna@cs.yale.edu

Peter Boncz
CWI
P.Boncz@cwi.nl

Stavros Harizopoulos
Amiato, Inc.
stavros@amiato.com

Stratos Idreos
Harvard University
stratos@seas.harvard.edu

Samuel Madden
MIT CSAIL
madden@csail.mit.edu



Amazon Redshift and the Case for Simpler Data Warehouses

Anurag Gupta, Deepak Agarwal, Derek Tan, Jakub Kulesza, Rahul Pathak,
Stefano Stefani, Vidhya Srinivasan
Amazon Web Services

Abstract

Amazon Redshift is a fast, fully managed, petabyte-scale data warehouse solution that makes it simple and cost-effective to efficiently analyze large volumes of data using existing business intelligence tools. Since launching in February 2013, it has been Amazon Web Service's (AWS) fastest growing service, with many thousands of customers and many petabytes of data under management.

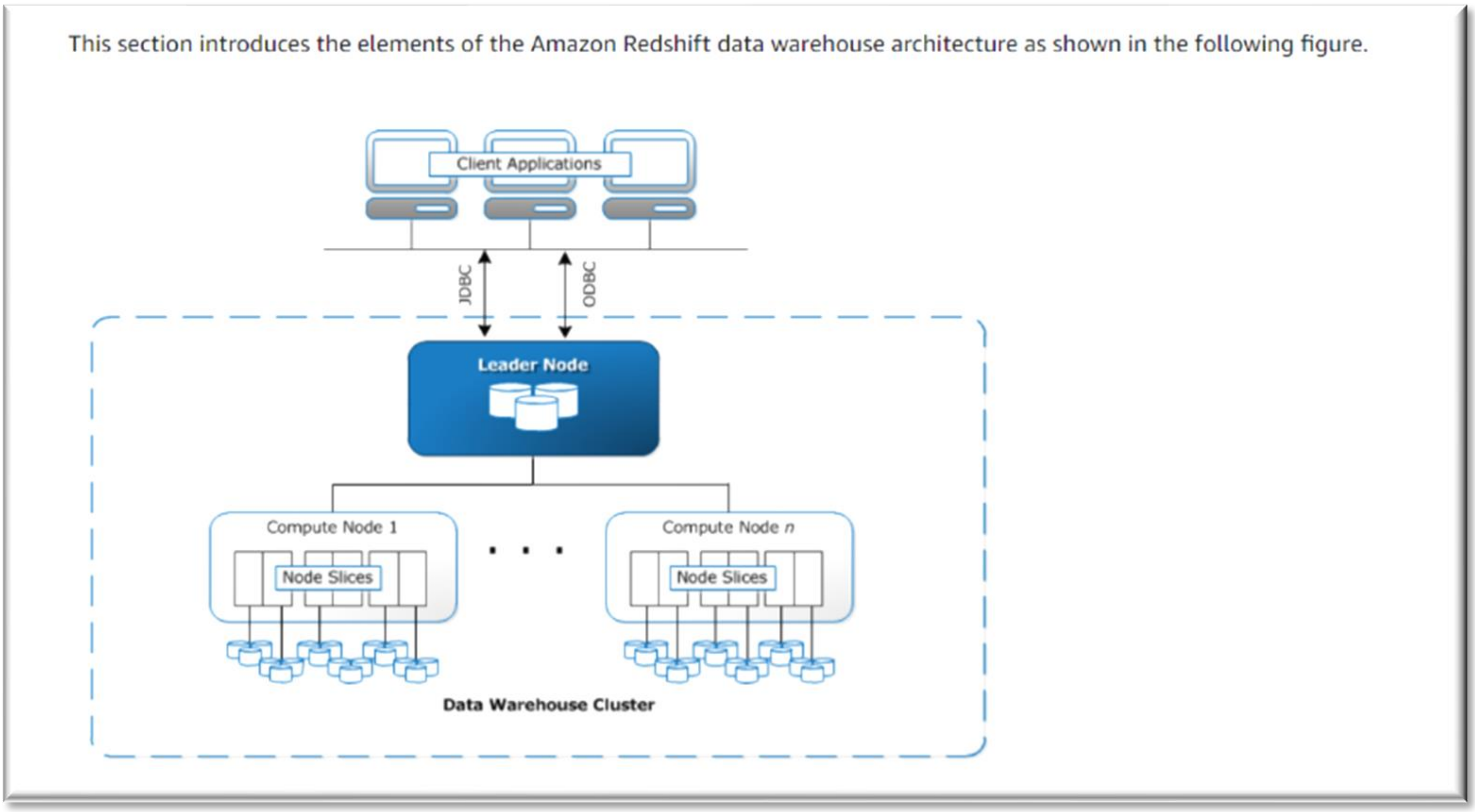
Amazon Redshift's pace of adoption has been a surprise to many participants in the data warehousing community. While Amazon

the past 12-18 months, new market research has begun to show an increase to 50-60%, with data doubling in size every 20 months.

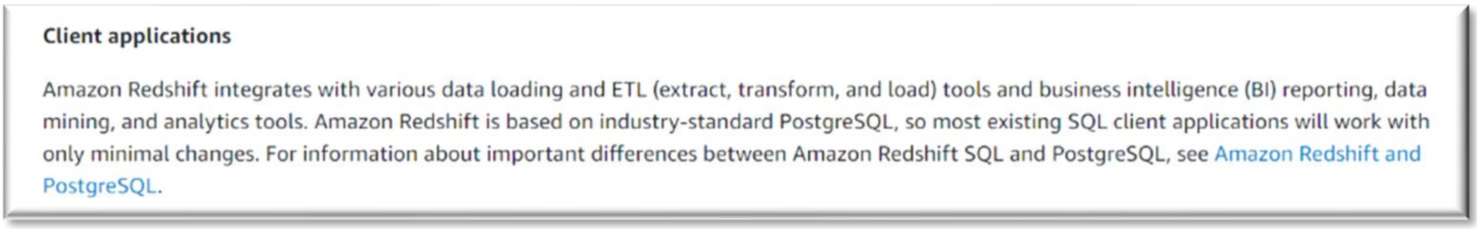
What do I need to know about compute nodes?

What do I need to know about leader nodes?

What do I need to know about Data distribution styles?

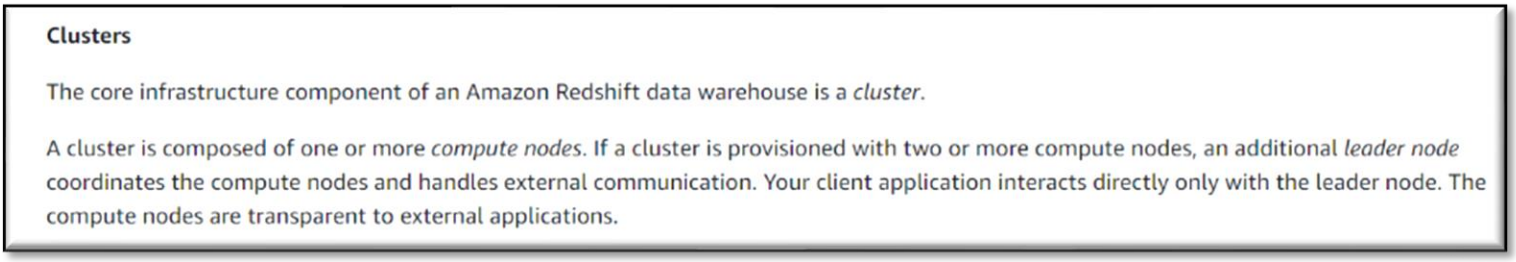


This section introduces the elements of the Amazon Redshift data warehouse architecture as shown in the following figure.



Client applications

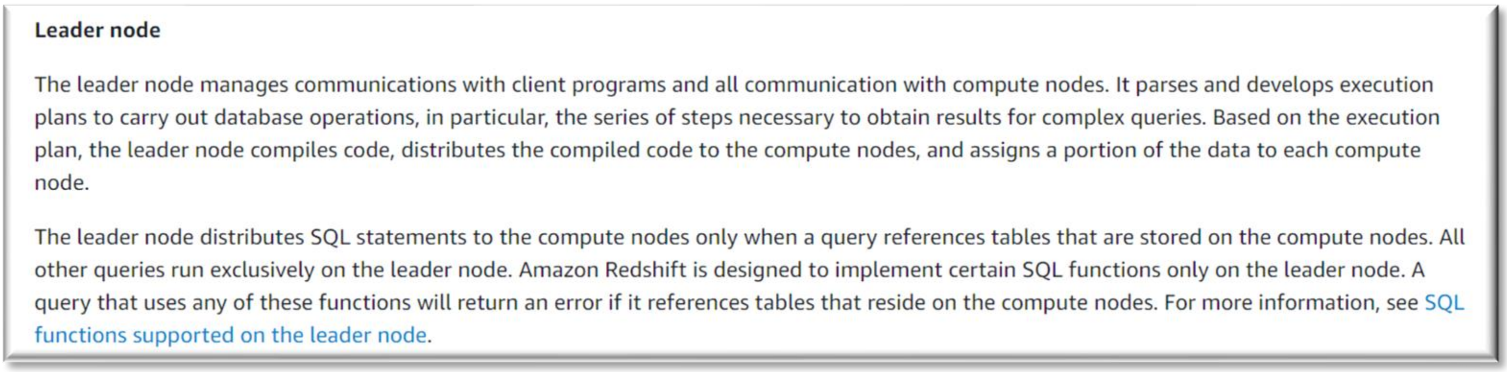
Amazon Redshift integrates with various data loading and ETL (extract, transform, and load) tools and business intelligence (BI) reporting, data mining, and analytics tools. Amazon Redshift is based on industry-standard PostgreSQL, so most existing SQL client applications will work with only minimal changes. For information about important differences between Amazon Redshift SQL and PostgreSQL, see Amazon Redshift and PostgreSQL.



Clusters

The core infrastructure component of an Amazon Redshift data warehouse is a *cluster*.

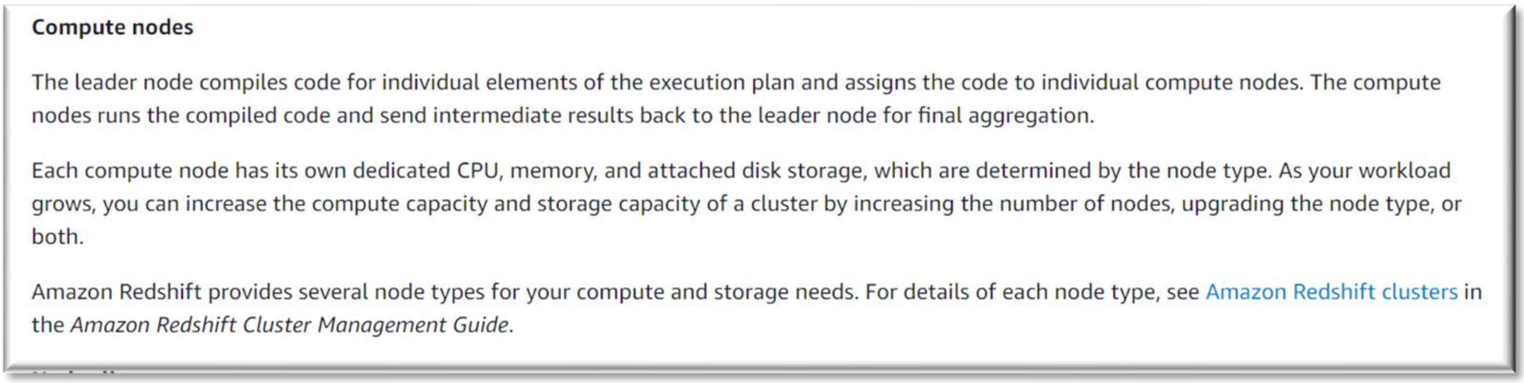
A cluster is composed of one or more *compute nodes*. If a cluster is provisioned with two or more compute nodes, an additional *leader node* coordinates the compute nodes and handles external communication. Your client application interacts directly only with the leader node. The compute nodes are transparent to external applications.



Leader node

The leader node manages communications with client programs and all communication with compute nodes. It parses and develops execution plans to carry out database operations, in particular, the series of steps necessary to obtain results for complex queries. Based on the execution plan, the leader node compiles code, distributes the compiled code to the compute nodes, and assigns a portion of the data to each compute node.

The leader node distributes SQL statements to the compute nodes only when a query references tables that are stored on the compute nodes. All other queries run exclusively on the leader node. Amazon Redshift is designed to implement certain SQL functions only on the leader node. A query that uses any of these functions will return an error if it references tables that reside on the compute nodes. For more information, see SQL functions supported on the leader node.

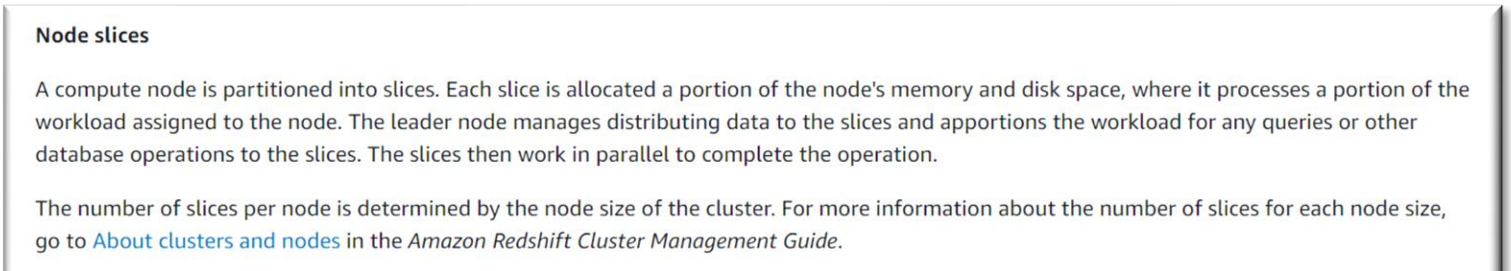


Compute nodes

The leader node compiles code for individual elements of the execution plan and assigns the code to individual compute nodes. The compute nodes runs the compiled code and send intermediate results back to the leader node for final aggregation.

Each compute node has its own dedicated CPU, memory, and attached disk storage, which are determined by the node type. As your workload grows, you can increase the compute capacity and storage capacity of a cluster by increasing the number of nodes, upgrading the node type, or both.

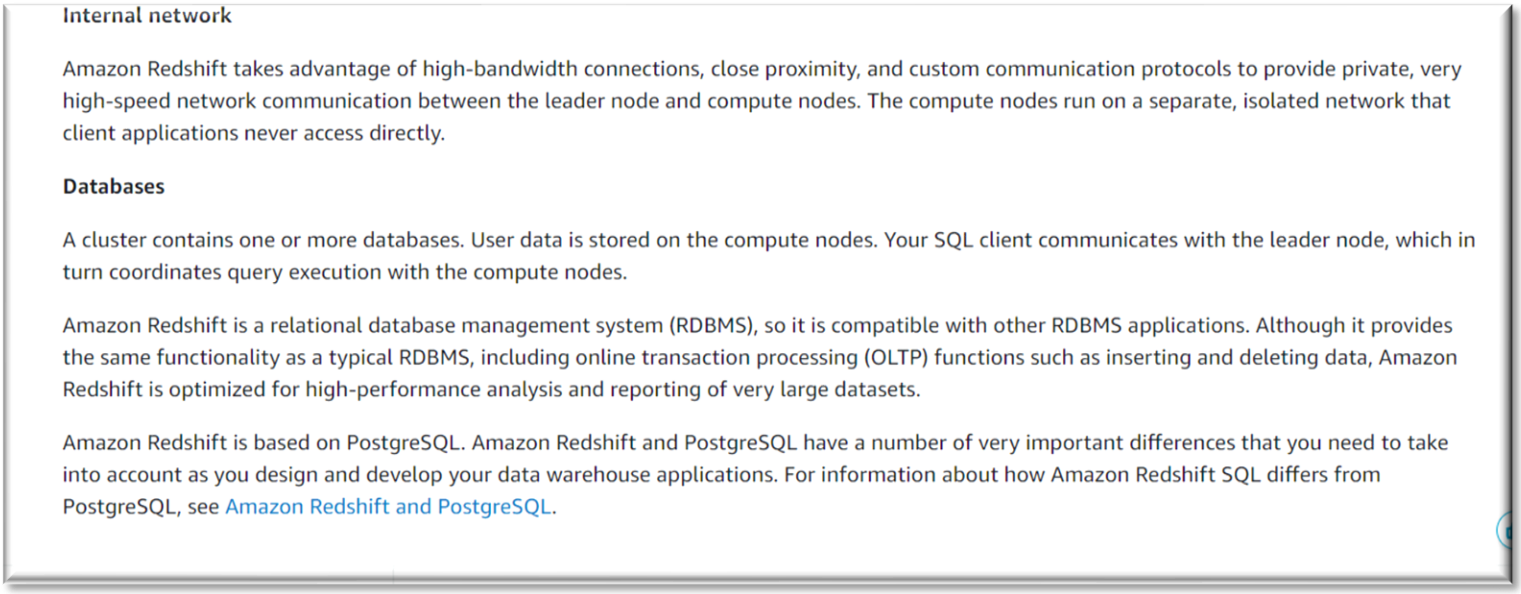
Amazon Redshift provides several node types for your compute and storage needs. For details of each node type, see Amazon Redshift clusters in the *Amazon Redshift Cluster Management Guide*.



Node slices

A compute node is partitioned into slices. Each slice is allocated a portion of the node's memory and disk space, where it processes a portion of the workload assigned to the node. The leader node manages distributing data to the slices and apportions the workload for any queries or other database operations to the slices. The slices then work in parallel to complete the operation.

The number of slices per node is determined by the node size of the cluster. For more information about the number of slices for each node size, go to About clusters and nodes in the *Amazon Redshift Cluster Management Guide*.



Internal network

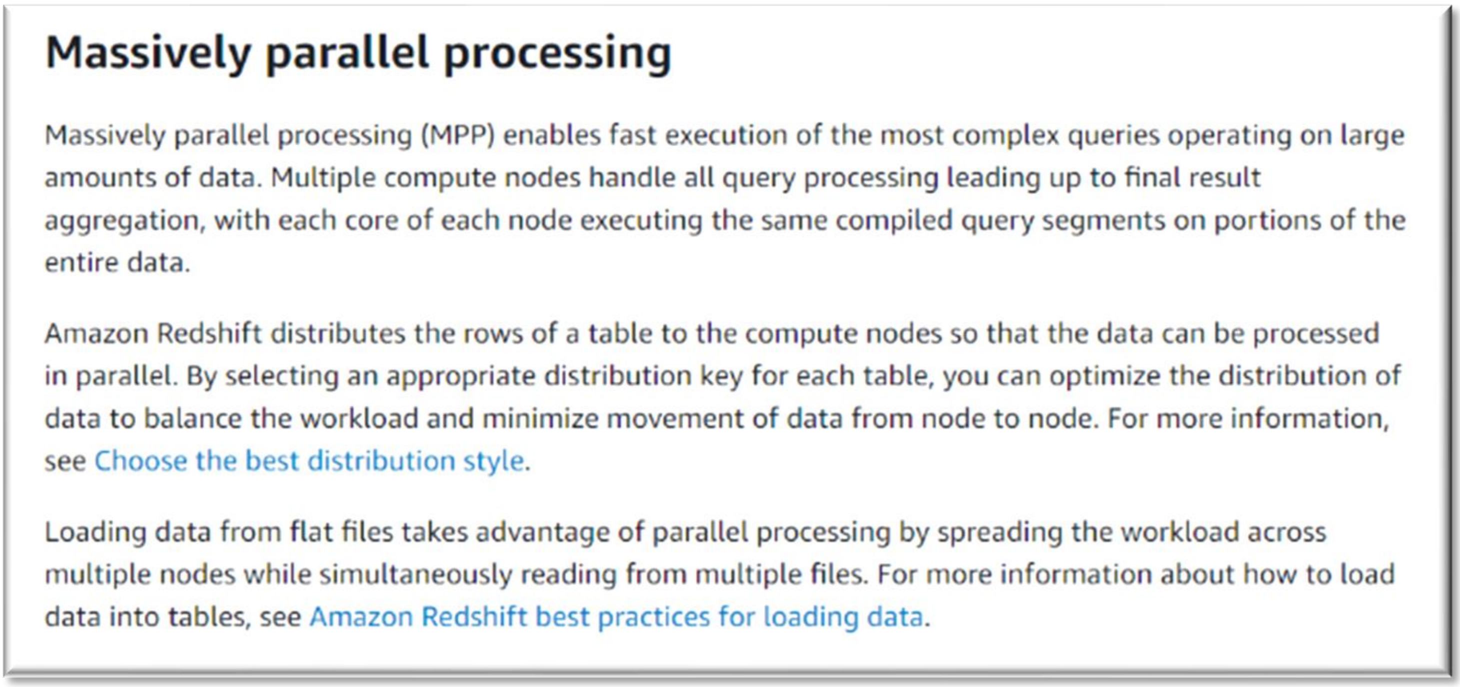
Amazon Redshift takes advantage of high-bandwidth connections, close proximity, and custom communication protocols to provide private, very high-speed network communication between the leader node and compute nodes. The compute nodes run on a separate, isolated network that client applications never access directly.

Databases

A cluster contains one or more databases. User data is stored on the compute nodes. Your SQL client communicates with the leader node, which in turn coordinates query execution with the compute nodes.

Amazon Redshift is a relational database management system (RDBMS), so it is compatible with other RDBMS applications. Although it provides the same functionality as a typical RDBMS, including online transaction processing (OLTP) functions such as inserting and deleting data, Amazon Redshift is optimized for high-performance analysis and reporting of very large datasets.

Amazon Redshift is based on PostgreSQL. Amazon Redshift and PostgreSQL have a number of very important differences that you need to take into account as you design and develop your data warehouse applications. For information about how Amazon Redshift SQL differs from PostgreSQL, see Amazon Redshift and PostgreSQL.



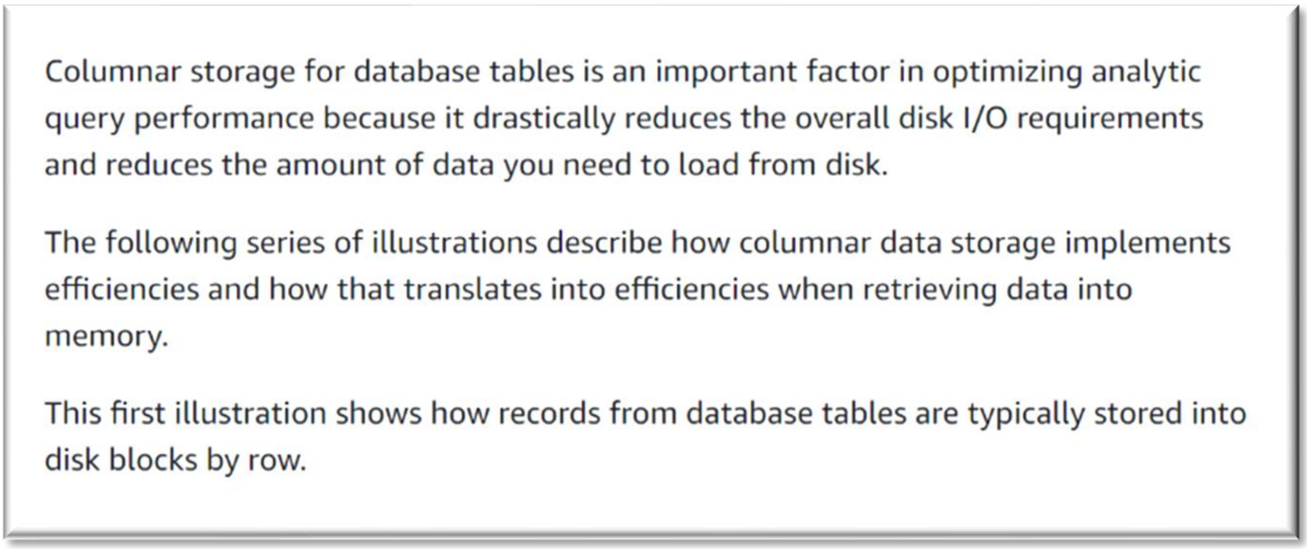
Massively parallel processing

Massively parallel processing (MPP) enables fast execution of the most complex queries operating on large amounts of data. Multiple compute nodes handle all query processing leading up to final result aggregation, with each core of each node executing the same compiled query segments on portions of the entire data.

Amazon Redshift distributes the rows of a table to the compute nodes so that the data can be processed in parallel. By selecting an appropriate distribution key for each table, you can optimize the distribution of data to balance the workload and minimize movement of data from node to node. For more information, see Choose the best distribution style.

Loading data from flat files takes advantage of parallel processing by spreading the workload across multiple nodes while simultaneously reading from multiple files. For more information about how to load data into tables, see Amazon Redshift best practices for loading data.

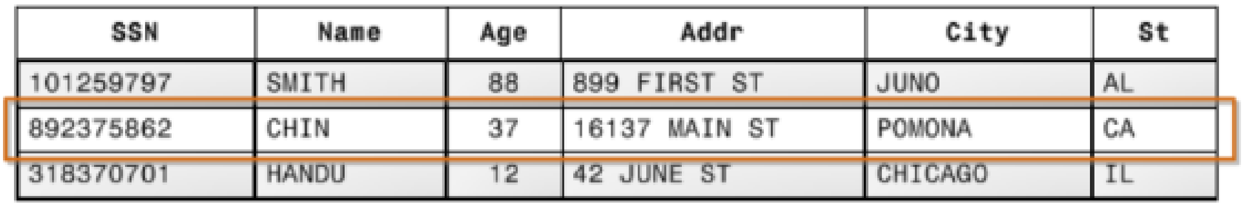
What on earth is “columnar storage”?



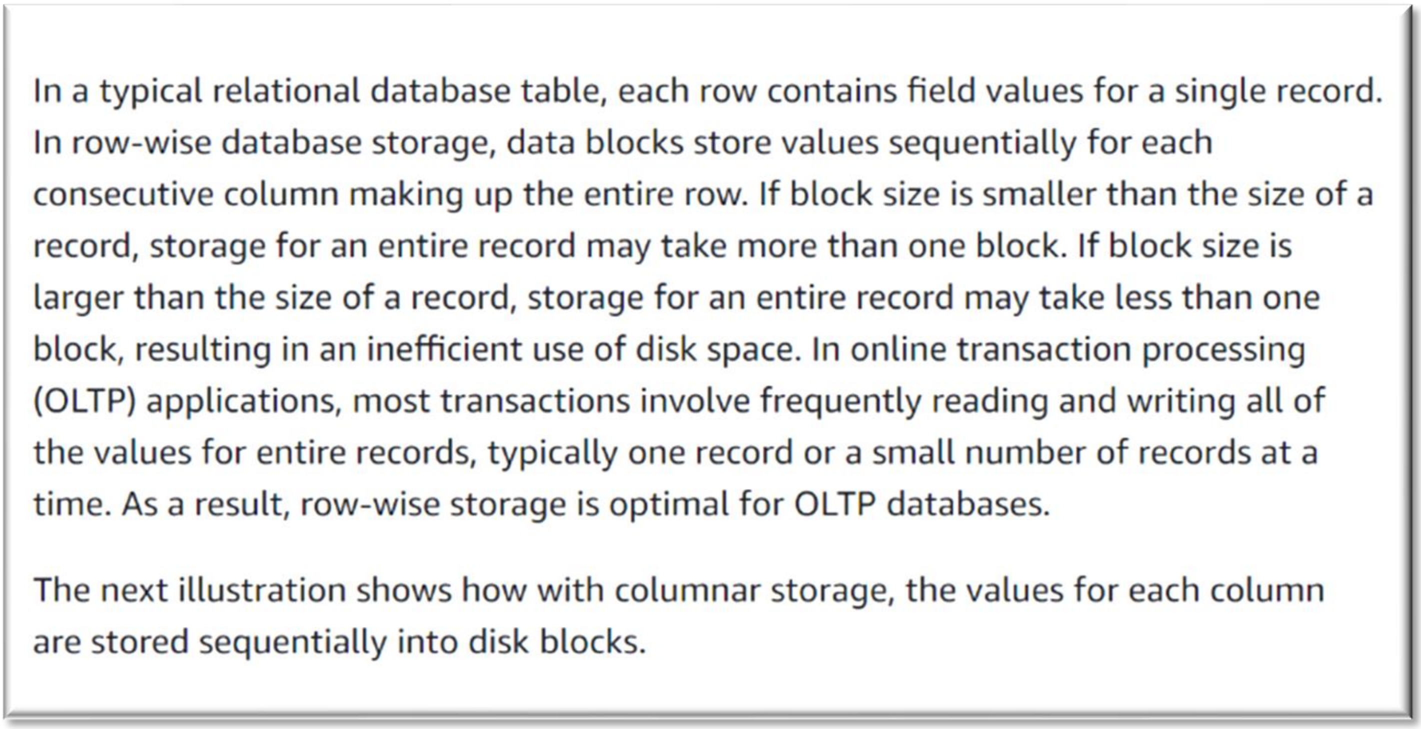
Columnar storage for database tables is an important factor in optimizing analytic query performance because it drastically reduces the overall disk I/O requirements and reduces the amount of data you need to load from disk.

The following series of illustrations describe how columnar data storage implements efficiencies and how that translates into efficiencies when retrieving data into memory.

This first illustration shows how records from database tables are typically stored into disk blocks by row.

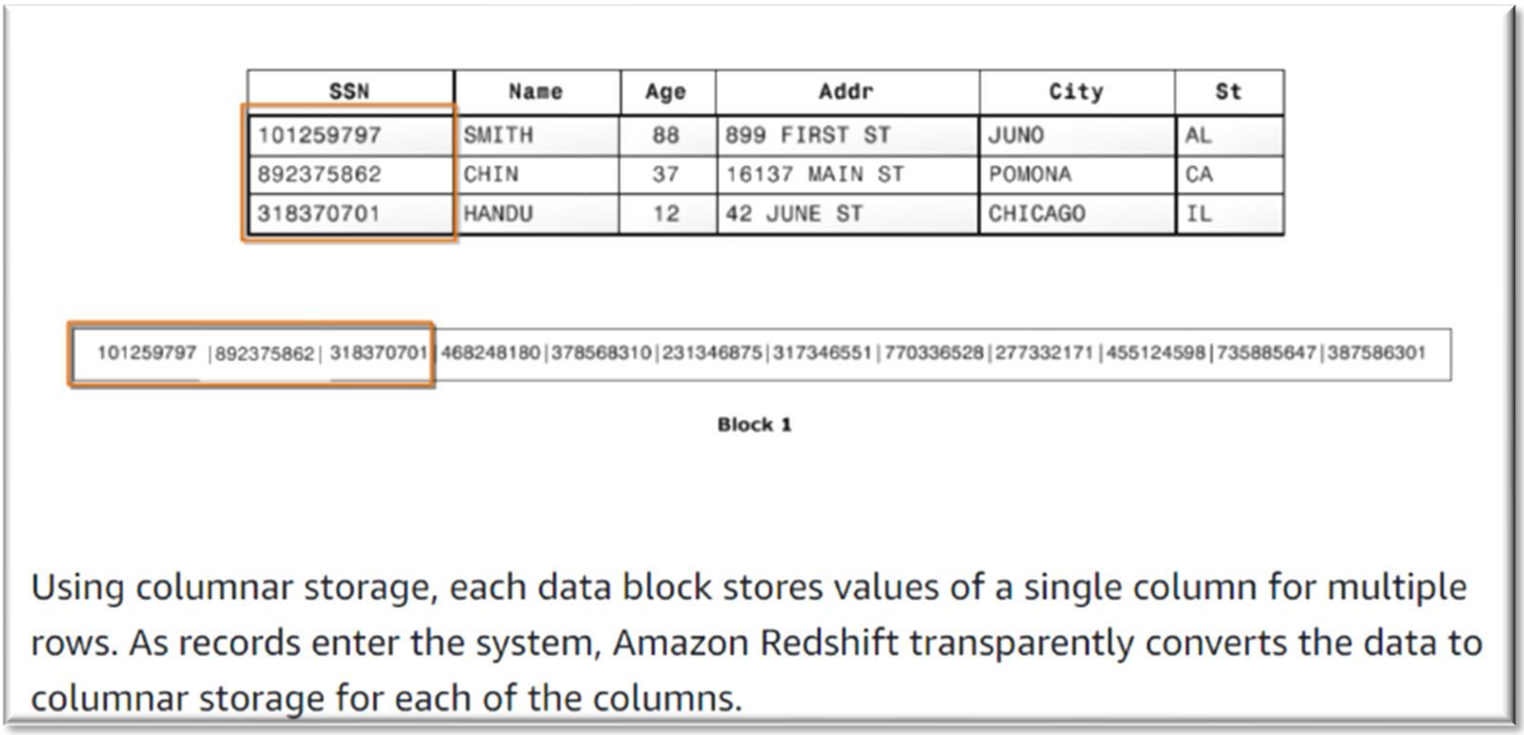


SSN	Name	Age	Addr	City	St
101259797	SMITH	88	899 FIRST ST	JUNO	AL
892375862	CHIN	37	16137 MAIN ST	POMONA	CA
318370701	HANDU	12	42 JUNE ST	CHICAGO	IL



In a typical relational database table, each row contains field values for a single record. In row-wise database storage, data blocks store values sequentially for each consecutive column making up the entire row. If block size is smaller than the size of a record, storage for an entire record may take more than one block. If block size is larger than the size of a record, storage for an entire record may take less than one block, resulting in an inefficient use of disk space. In online transaction processing (OLTP) applications, most transactions involve frequently reading and writing all of the values for entire records, typically one record or a small number of records at a time. As a result, row-wise storage is optimal for OLTP databases.

The next illustration shows how with columnar storage, the values for each column are stored sequentially into disk blocks.

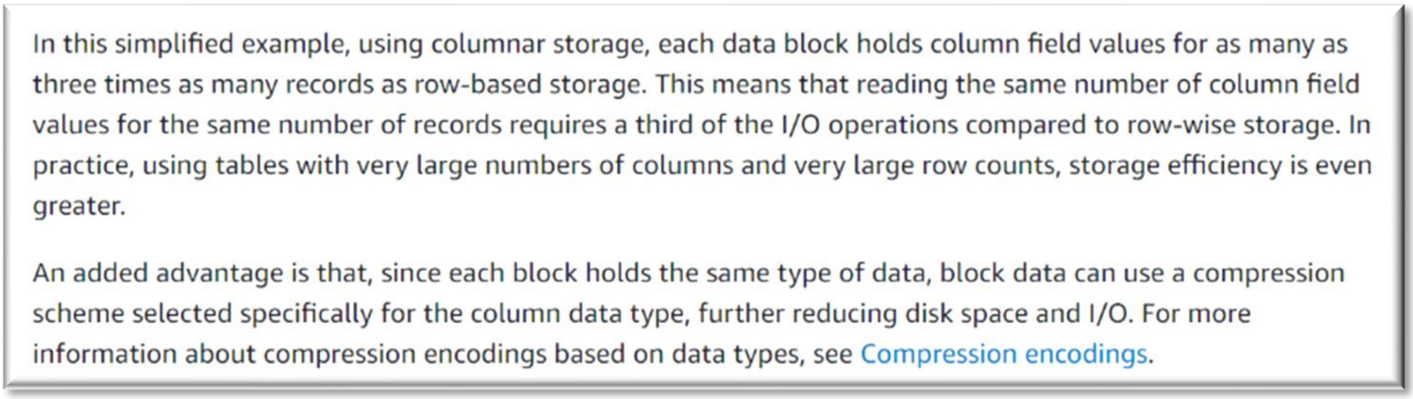


SSN	Name	Age	Addr	City	St
101259797	SMITH	88	899 FIRST ST	JUNO	AL
892375862	CHIN	37	16137 MAIN ST	POMONA	CA
318370701	HANDU	12	42 JUNE ST	CHICAGO	IL

101259797 |892375862| 318370701|468248180|378568310|231346875|317346551|770336528|277332171|455124598|735885647|387586301

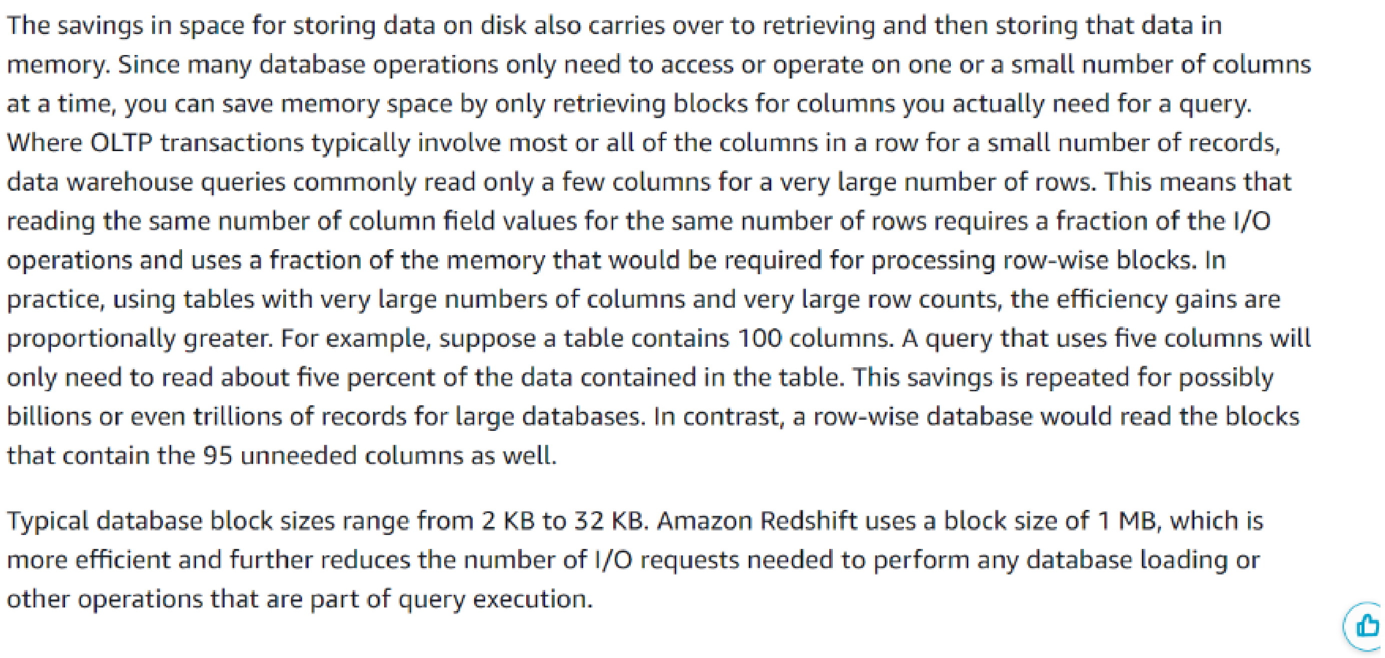
Block 1

Using columnar storage, each data block stores values of a single column for multiple rows. As records enter the system, Amazon Redshift transparently converts the data to columnar storage for each of the columns.



In this simplified example, using columnar storage, each data block holds column field values for as many as three times as many records as row-based storage. This means that reading the same number of column field values for the same number of records requires a third of the I/O operations compared to row-wise storage. In practice, using tables with very large numbers of columns and very large row counts, storage efficiency is even greater.

An added advantage is that, since each block holds the same type of data, block data can use a compression scheme selected specifically for the column data type, further reducing disk space and I/O. For more information about compression encodings based on data types, see Compression encodings.

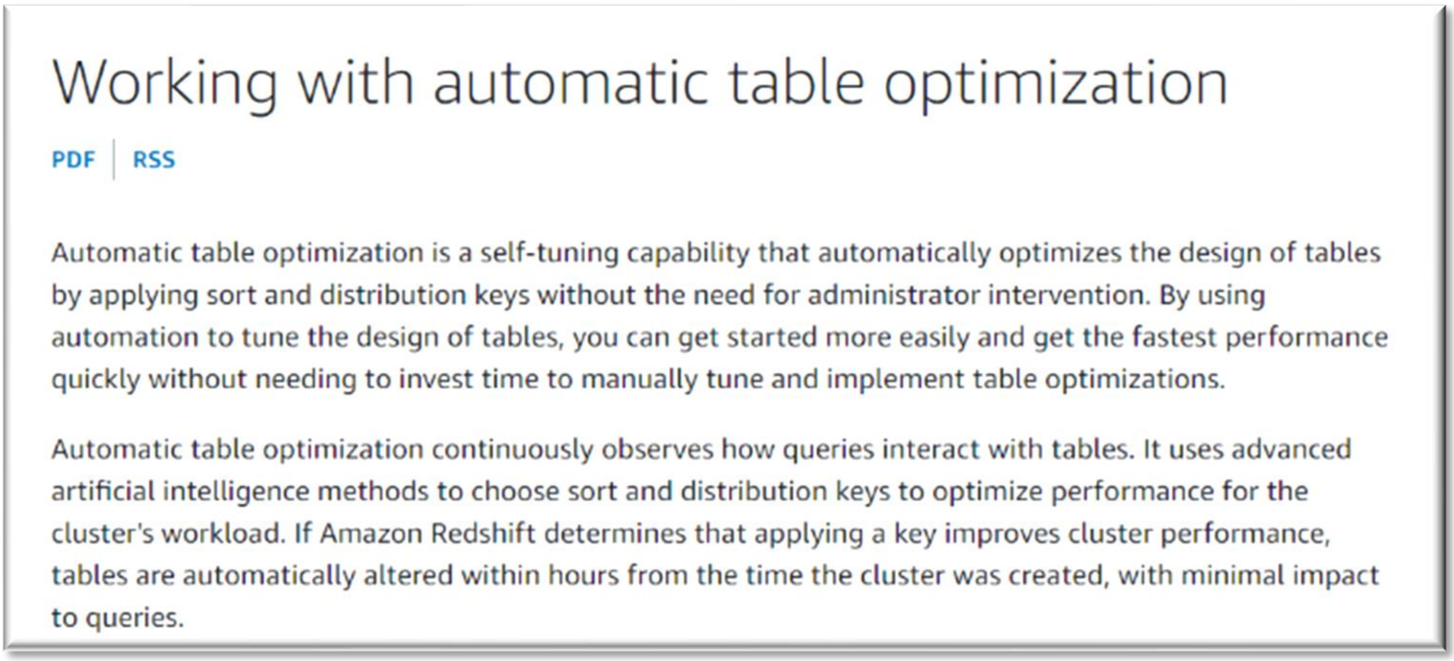


The savings in space for storing data on disk also carries over to retrieving and then storing that data in memory. Since many database operations only need to access or operate on one or a small number of columns at a time, you can save memory space by only retrieving blocks for columns you actually need for a query. Where OLTP transactions typically involve most or all of the columns in a row for a small number of records, data warehouse queries commonly read only a few columns for a very large number of rows. This means that reading the same number of column field values for the same number of rows requires a fraction of the I/O operations and uses a fraction of the memory that would be required for processing row-wise blocks. In practice, using tables with very large numbers of columns and very large row counts, the efficiency gains are proportionally greater. For example, suppose a table contains 100 columns. A query that uses five columns will only need to read about five percent of the data contained in the table. This savings is repeated for possibly billions or even trillions of records for large databases. In contrast, a row-wise database would read the blocks that contain the 95 unneeded columns as well.

Typical database block sizes range from 2 KB to 32 KB. Amazon Redshift uses a block size of 1 MB, which is more efficient and further reduces the number of I/O requests needed to perform any database loading or other operations that are part of query execution.

What on earth is "automatic table optimization"?

AWS write:

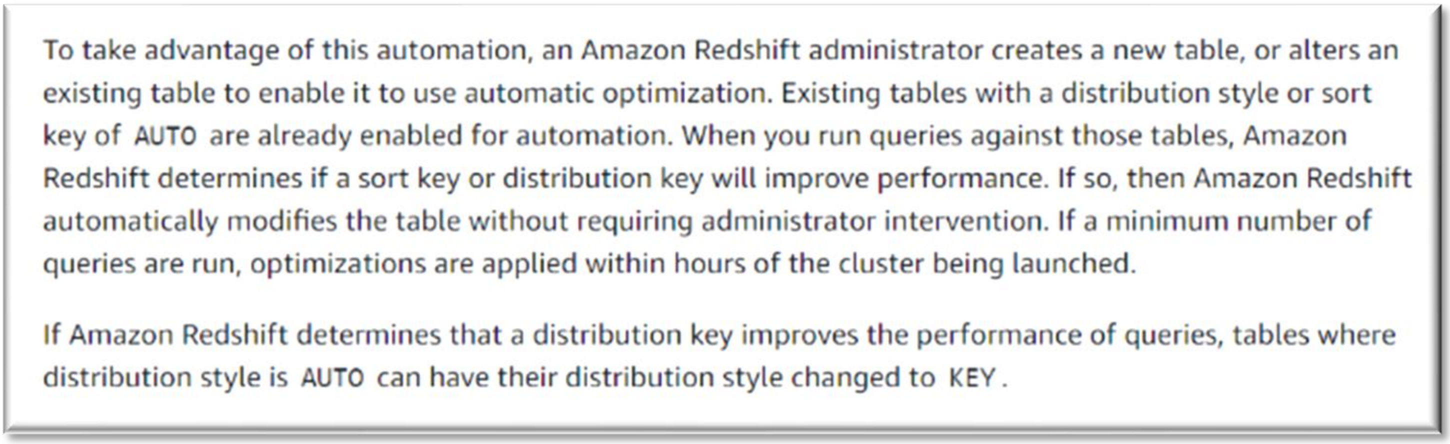


Working with automatic table optimization

PDF | RSS

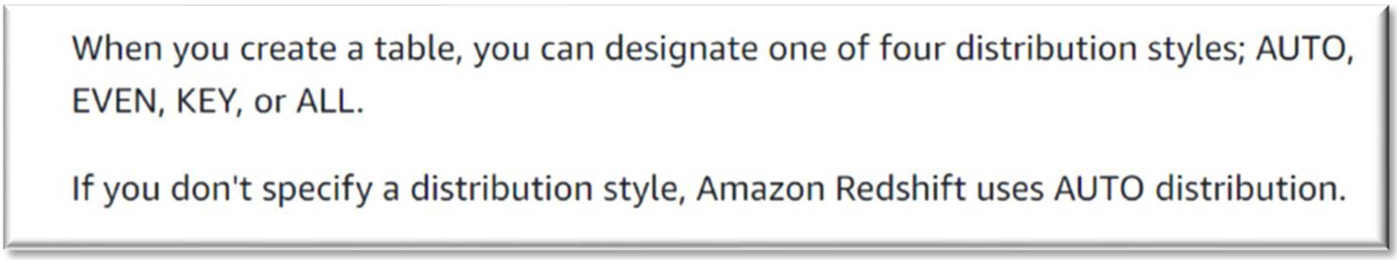
Automatic table optimization is a self-tuning capability that automatically optimizes the design of tables by applying sort and distribution keys without the need for administrator intervention. By using automation to tune the design of tables, you can get started more easily and get the fastest performance quickly without needing to invest time to manually tune and implement table optimizations.

Automatic table optimization continuously observes how queries interact with tables. It uses advanced artificial intelligence methods to choose sort and distribution keys to optimize performance for the cluster's workload. If Amazon Redshift determines that applying a key improves cluster performance, tables are automatically altered within hours from the time the cluster was created, with minimal impact to queries.



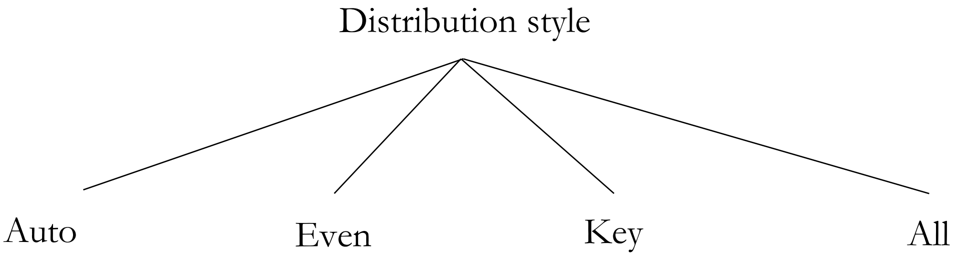
To take advantage of this automation, an Amazon Redshift administrator creates a new table, or alters an existing table to enable it to use automatic optimization. Existing tables with a distribution style or sort key of `AUTO` are already enabled for automation. When you run queries against those tables, Amazon Redshift determines if a sort key or distribution key will improve performance. If so, then Amazon Redshift automatically modifies the table without requiring administrator intervention. If a minimum number of queries are run, optimizations are applied within hours of the cluster being launched.

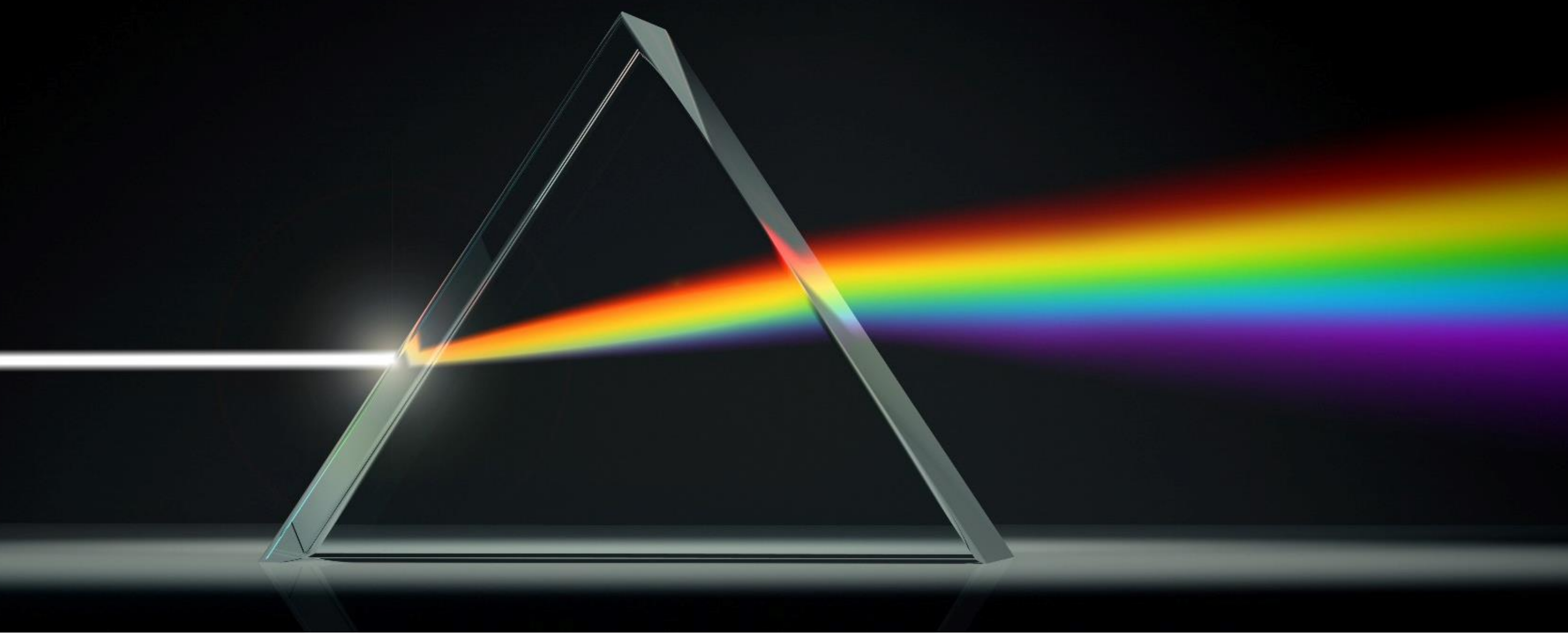
If Amazon Redshift determines that a distribution key improves the performance of queries, tables where distribution style is `AUTO` can have their distribution style changed to `KEY`.



When you create a table, you can designate one of four distribution styles; AUTO, EVEN, KEY, or ALL.

If you don't specify a distribution style, Amazon Redshift uses AUTO distribution.





Redshift Spectrum

Werner Vogels announcing Redshift Spectrum on stage, in San Francisco (2017)

TPN

141. ***Phenomenon1*** – the tendency of X to Y.
142. ***Phen2*** – the tendency of X to Y.
143. ***Phen3*** – the tendency of X to Y.
144. ***Phen4*** – the tendency of X to Y.
145. ***Phen5*** – the tendency of X to Y.
146. ***Phen6*** – the tendency of X to Y.
147. ***Phen7*** – the tendency of X to Y.
148. ***Phen8*** – the tendency of X to Y.
149. ***Phen9*** – the tendency of X to Y.
150. ***Phen10*** – the tendency of X to Y.

Review Questions

Glossary

ODBC

Open Database Connectivity.

JDBC

Java Database Connectivity.

EVEN Distribution

Description of what term means here.

KEY Distribution

Description of what term means here.

ALL Distribution

Description of what term means here.

Bibliography

I. Official

[Surname1]
Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]
Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

II. Unofficial

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

https://dev.to/alexantra/the-r-a-g-redshift-analyst-guide-understanding-the-query-plan-explain-360d

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

IV. General

[Surname1]

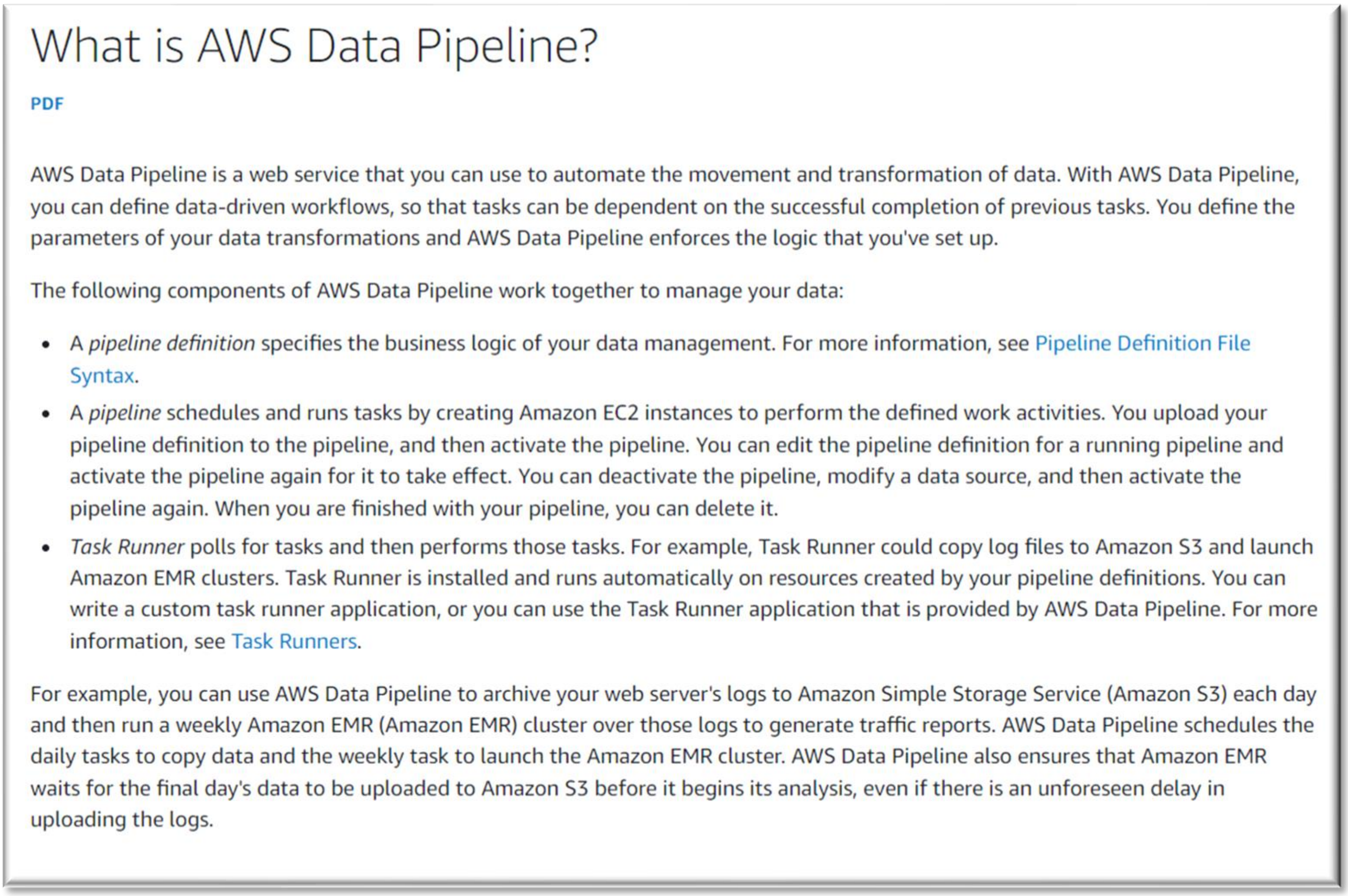
Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

Data Pipeline

Kathryn Shih explaining Data Pipeline in Las Vegas, in November 2012



What is AWS Data Pipeline?

PDF

AWS Data Pipeline is a web service that you can use to automate the movement and transformation of data. With AWS Data Pipeline, you can define data-driven workflows, so that tasks can be dependent on the successful completion of previous tasks. You define the parameters of your data transformations and AWS Data Pipeline enforces the logic that you've set up.

The following components of AWS Data Pipeline work together to manage your data:

- A *pipeline definition* specifies the business logic of your data management. For more information, see Pipeline Definition File Syntax.
- A *pipeline* schedules and runs tasks by creating Amazon EC2 instances to perform the defined work activities. You upload your pipeline definition to the pipeline, and then activate the pipeline. You can edit the pipeline definition for a running pipeline and activate the pipeline again for it to take effect. You can deactivate the pipeline, modify a data source, and then activate the pipeline again. When you are finished with your pipeline, you can delete it.
- *Task Runner* polls for tasks and then performs those tasks. For example, Task Runner could copy log files to Amazon S3 and launch Amazon EMR clusters. Task Runner is installed and runs automatically on resources created by your pipeline definitions. You can write a custom task runner application, or you can use the Task Runner application that is provided by AWS Data Pipeline. For more information, see Task Runners.

For example, you can use AWS Data Pipeline to archive your web server's logs to Amazon Simple Storage Service (Amazon S3) each day and then run a weekly Amazon EMR (Amazon EMR) cluster over those logs to generate traffic reports. AWS Data Pipeline schedules the daily tasks to copy data and the weekly task to launch the Amazon EMR cluster. AWS Data Pipeline also ensures that Amazon EMR waits for the final day's data to be uploaded to Amazon S3 before it begins its analysis, even if there is an unforeseen delay in uploading the logs.

Text

NOTE

Similar names

To those new to the service, the expression “data pipeline” denotes a very general idea: moving data from one place to another. Once you become more familiar with the service you will learn about how *precisely* it helps with the movement of data. But it remains the case that the *name* of the service denotes a very general idea. Arguably, the whole of the IT industry is about moving data from one place to another.

A name such as “data pipeline” will therefore become problematic. You will confuse it with similar names which start with the word ‘data’. In 2018, AWS announced the “DataSync” product; in 2019, they announced “Data Exchange”. These names *also* suggest the general idea of moving data from one place to another. We also see other services which end in “pipeline”. In 2015, *Code*Pipeline was announced.

Without being aware of this phenomenon, it can be quite annoying. One day you will be surprised to learn that there are in fact two services—CodePipeline and DataPipeline—and you had unconsciously been treating them as the same service. And we need to get names right, because we partly learn about services by hearing comments on them, in a piecemeal manner, accumulating our conception of a service. If we have been conflating two names (such as the two “Pipelines”), then we have been ‘filing’ our findings in the wrong cabinet. We have to re-arrange our map of things in our head. We cannot simply focus on the service-in-itself (paying no attention to these silly, similar names), because people use the names. And we need to know which service is being mapped to, so that we can increase our knowledge of it.

There are other examples of this phenomenon. There is Amazon Connect and Direct Connect. I start to be sympathetic with unique names, such as Fargate, Glacier, and Macie. The phenomenon can be formulated as follows.

(1) First, there is some compound name (such as Data Pipeline), which consists of more than one word. One, or perhaps both, of the constituent words is then used for other services. For example, we get Data Exchange, and Code Pipeline.

(2) Second, the whole expression (e.g., Data Pipelines) also *denotes* a highly general idea. This exacerbates the superficial, linguistic similarity in (1).

Both parts are necessary. No one gets confused by the expressions "river bank" and "Lloyds bank" even though these expressions have a word in common. Each denotes a distinct, specific idea. We have the side of a river and a financial institution.

So, how should we respond to this phenomenon? Now that we know it exists, our aim is to minimise the extend to which it can cause confusion. I do not think the solution is simply to ignore names. It is true that if we comes to use the expression "data pipeline" a lot, it becomes harder to conflate it with Code Pipeline. But this is sort of begging the question. We *aren't* familiar with the services. Obviously, those who helped design the CodePipeline service never read about DataPipeline and believe—for a moment—that they are reading about CodePipeline. But this is because it *means* a lot to them. We are students. We are *un*familiar, so we don't have this luxury of habituation (of having *inhabited* the terms).

My solution essentially involves making the exact letters involved in the expression indispensable to the explanation of the service itself. In other words, it's impossible to understand the service itself without thinking of the specific expression used to name it.

We're *couching* our explanation of the service in the exact name used. Sometimes, what we're doing is almost like telling a story. The idea is that the story *explains* the name. Stories are harder to forget. But it's also the fact that things which have an explanation are understood, and things which are understood stay with us more permanently. Humans do not remember arbitrary strings of characters (in an instance ID, for example). Humans do not remember arbitrary things.

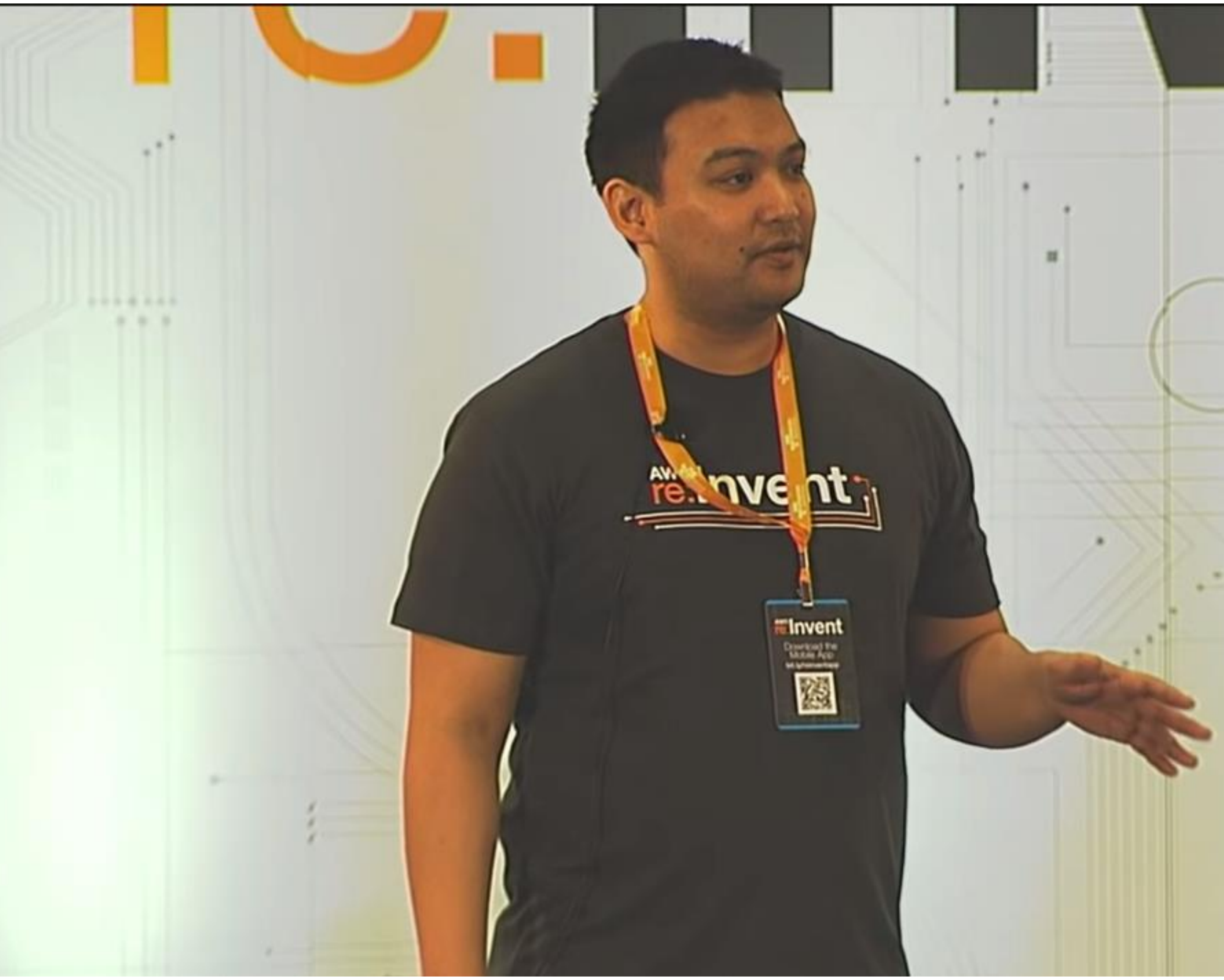
Of course, some services have no reason for the name. They were arbitrarily chosen. So, what we're doing is perhaps inventing a fiction. But it is fiction which means we reliably produce true statements (or get the exact name of the service, in our case). Do not be repelled by the word "fiction". The scientific world is packed with models that are fictional but allow us to reliable make the right statements and do the right things.

Command Line Interface (CLI)

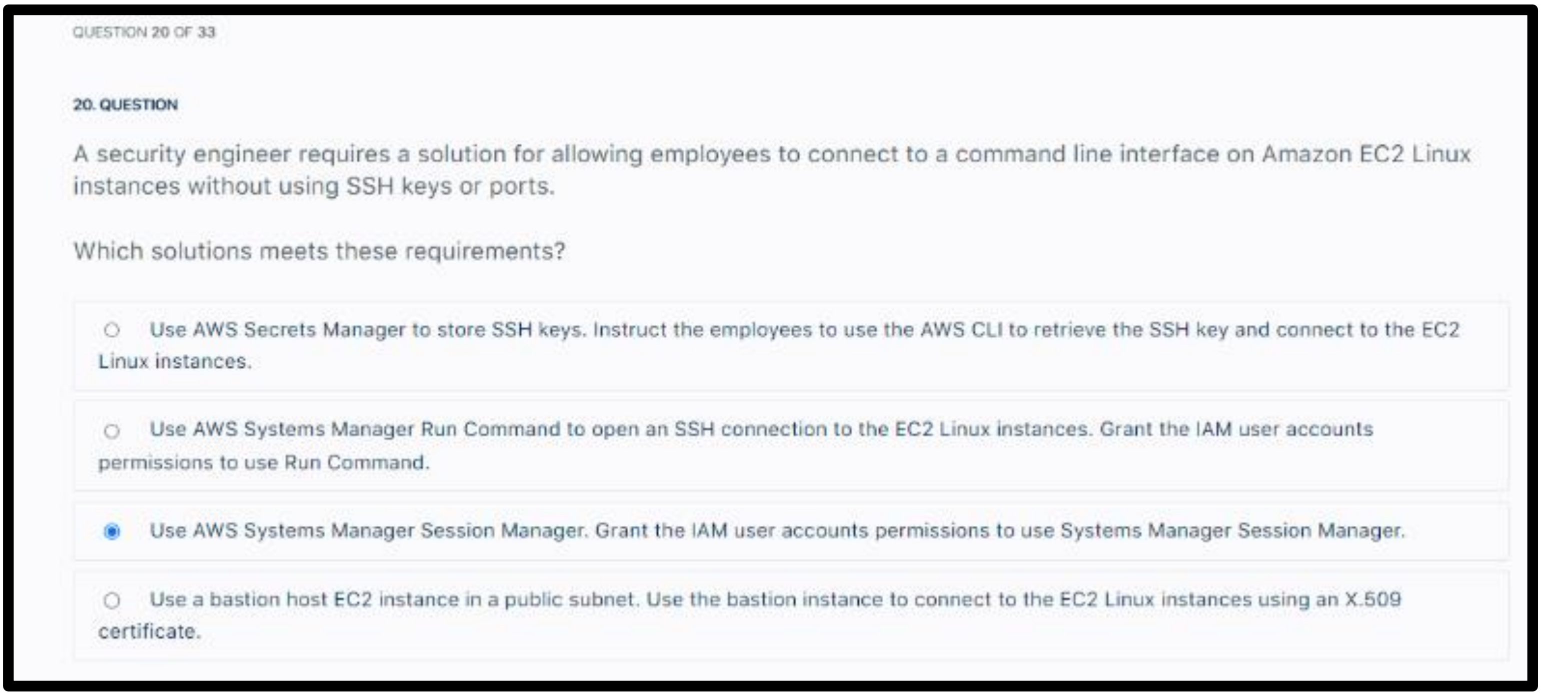
Questions I had:

1. What is the meaning of two dashes before something?
2. What about when there is one dash before a word?
3. What does it mean when there are no dashes before something?

When you put two dashes, it's like you're saying "GET READY". Thump the table with each word: GET READY (dash, dash). Why get ready? Because you're giving the name of a *parameter*. Parameter as in "determinable", or something which can *take* various values. And what you're going to do next is set the value of that parameter. You let the CLI know you're specifying a value by using a *single* dash.



James Saryerwinnie delivering his great presentation “becoming an AWS Command Line Expert”



QUESTION 20 OF 33

20. QUESTION

A security engineer requires a solution for allowing employees to connect to a command line interface on Amazon EC2 Linux instances without using SSH keys or ports.

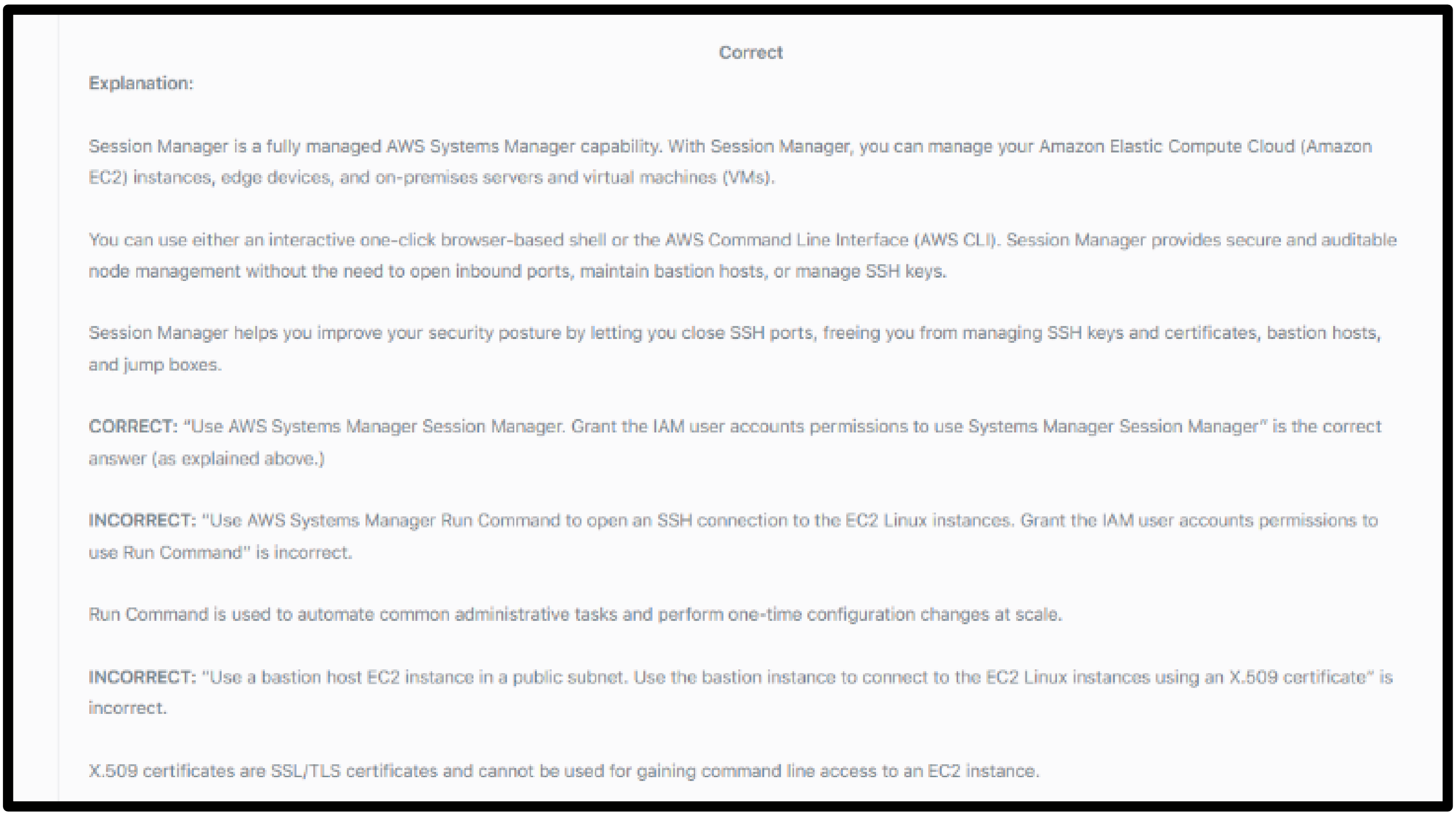
Which solutions meets these requirements?

○ Use AWS Secrets Manager to store SSH keys. Instruct the employees to use the AWS CLI to retrieve the SSH key and connect to the EC2 Linux instances.

○ Use AWS Systems Manager Run Command to open an SSH connection to the EC2 Linux instances. Grant the IAM user accounts permissions to use Run Command.

● Use AWS Systems Manager Session Manager. Grant the IAM user accounts permissions to use Systems Manager Session Manager.

○ Use a bastion host EC2 instance in a public subnet. Use the bastion instance to connect to the EC2 Linux instances using an X.509 certificate.



Correct

Explanation:

Session Manager is a fully managed AWS Systems Manager capability. With Session Manager, you can manage your Amazon Elastic Compute Cloud (Amazon EC2) instances, edge devices, and on-premises servers and virtual machines (VMs).

You can use either an interactive one-click browser-based shell or the AWS Command Line Interface (AWS CLI). Session Manager provides secure and auditable node management without the need to open inbound ports, maintain bastion hosts, or manage SSH keys.

Session Manager helps you improve your security posture by letting you close SSH ports, freeing you from managing SSH keys and certificates, bastion hosts, and jump boxes.

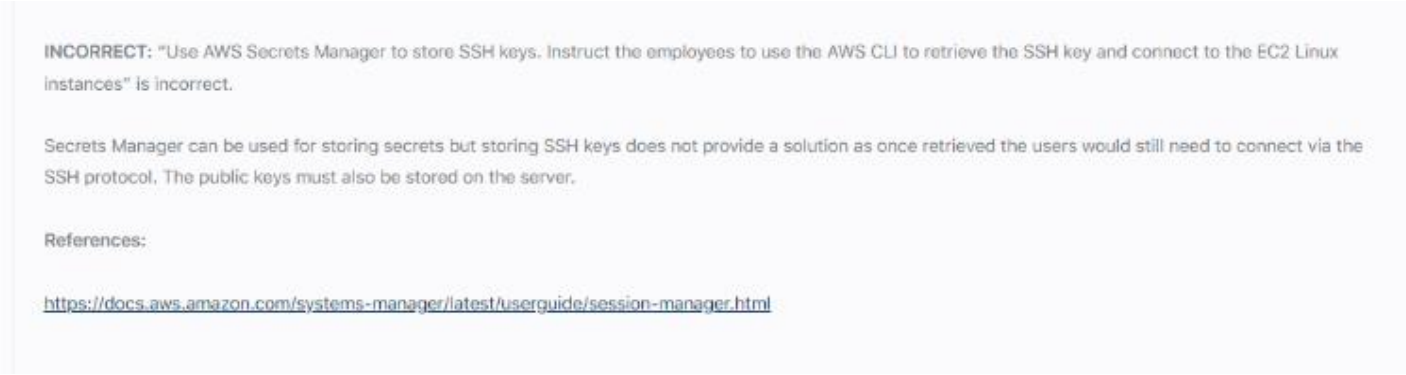
CORRECT: "Use AWS Systems Manager Session Manager. Grant the IAM user accounts permissions to use Systems Manager Session Manager" is the correct answer (as explained above.)

INCORRECT: "Use AWS Systems Manager Run Command to open an SSH connection to the EC2 Linux instances. Grant the IAM user accounts permissions to use Run Command" is incorrect.

Run Command is used to automate common administrative tasks and perform one-time configuration changes at scale.

INCORRECT: "Use a bastion host EC2 instance in a public subnet. Use the bastion instance to connect to the EC2 Linux instances using an X.509 certificate" is incorrect.

X.509 certificates are SSL/TLS certificates and cannot be used for gaining command line access to an EC2 instance.

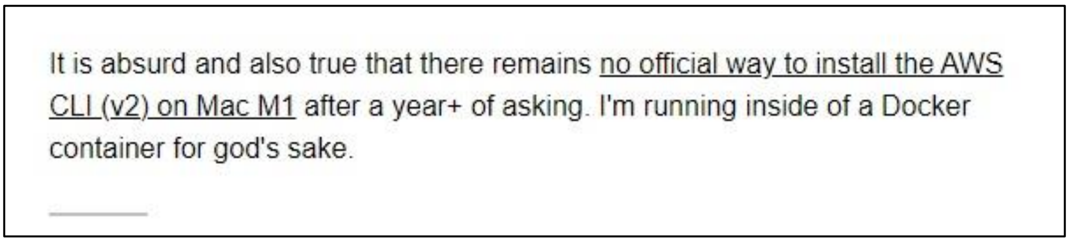


INCORRECT: "Use AWS Secrets Manager to store SSH keys. Instruct the employees to use the AWS CLI to retrieve the SSH key and connect to the EC2 Linux instances" is incorrect.

Secrets Manager can be used for storing secrets but storing SSH keys does not provide a solution as once retrieved the users would still need to connect via the SSH protocol. The public keys must also be stored on the server.

References:

https://docs.aws.amazon.com/systems-manager/latest/userguide/session-manager.html



It is absurd and also true that there remains no official way to install the AWS CLI (v2) on Mac M1 after a year+ of asking. I'm running inside of a Docker container for god's sake.

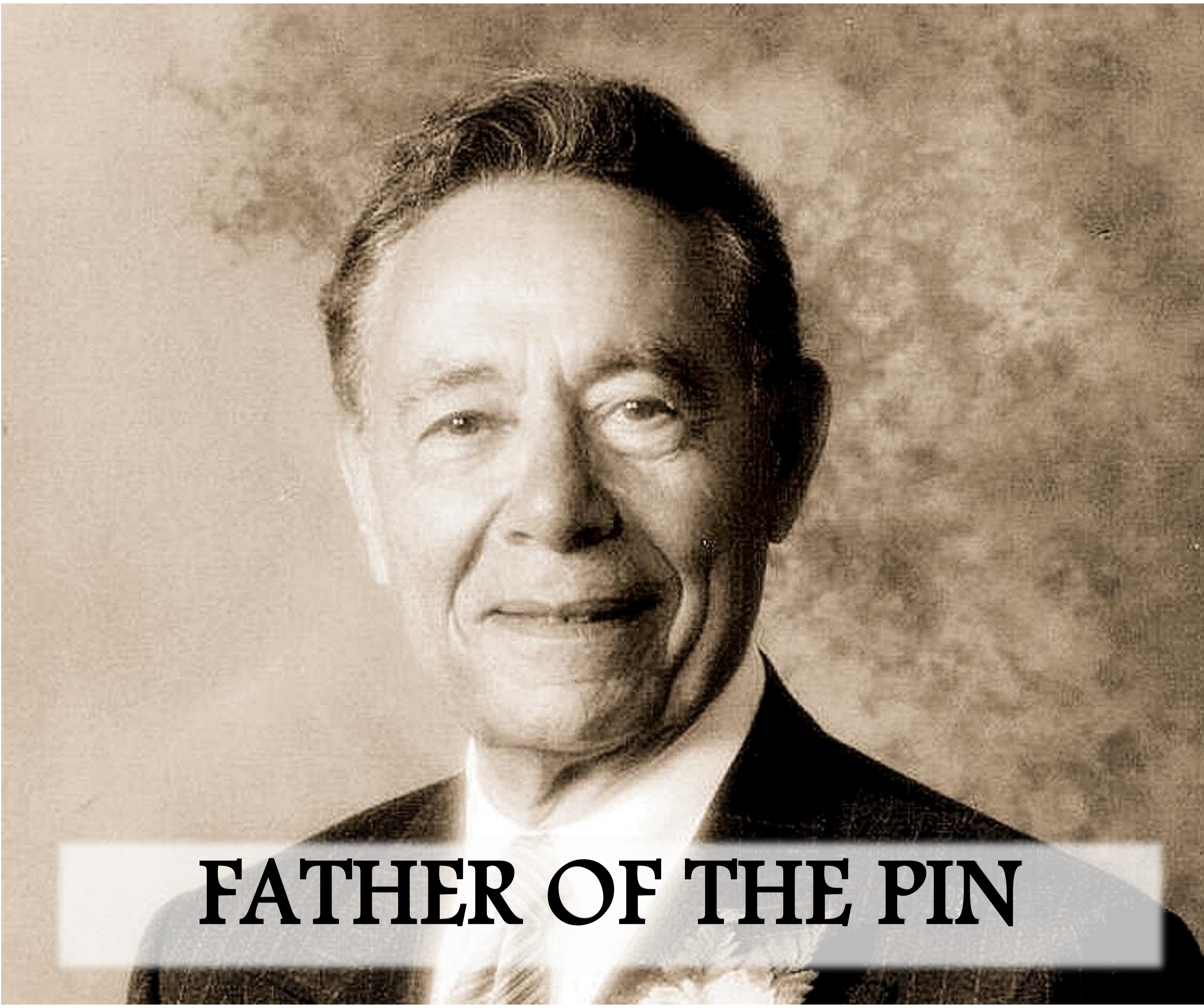
Email from Corey Quinn on June 26th 2023 https://github.com/aws/aws-cli/issues/7252?ck_subscriber_id=1560524742

Management Console Mobile Application

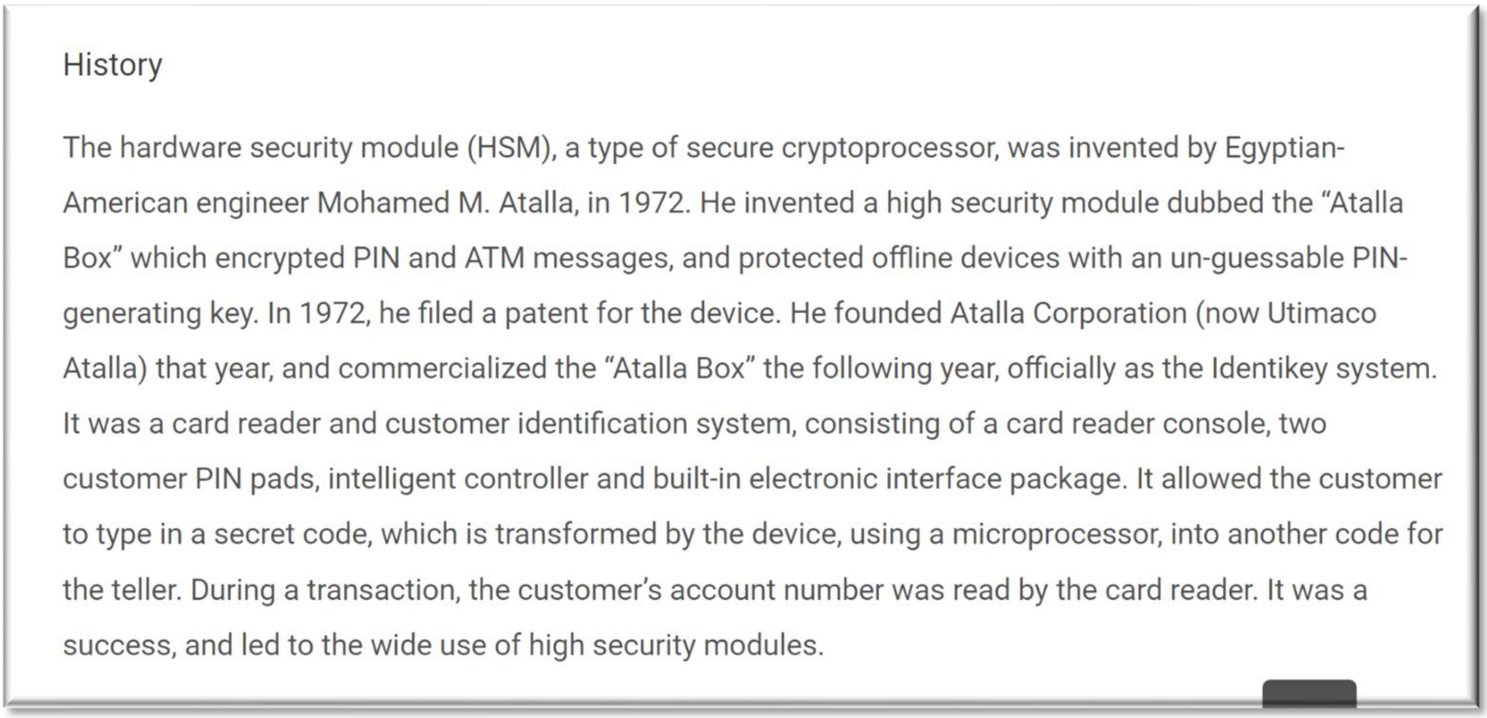
Elastic Transcoder

OpsWorks

←

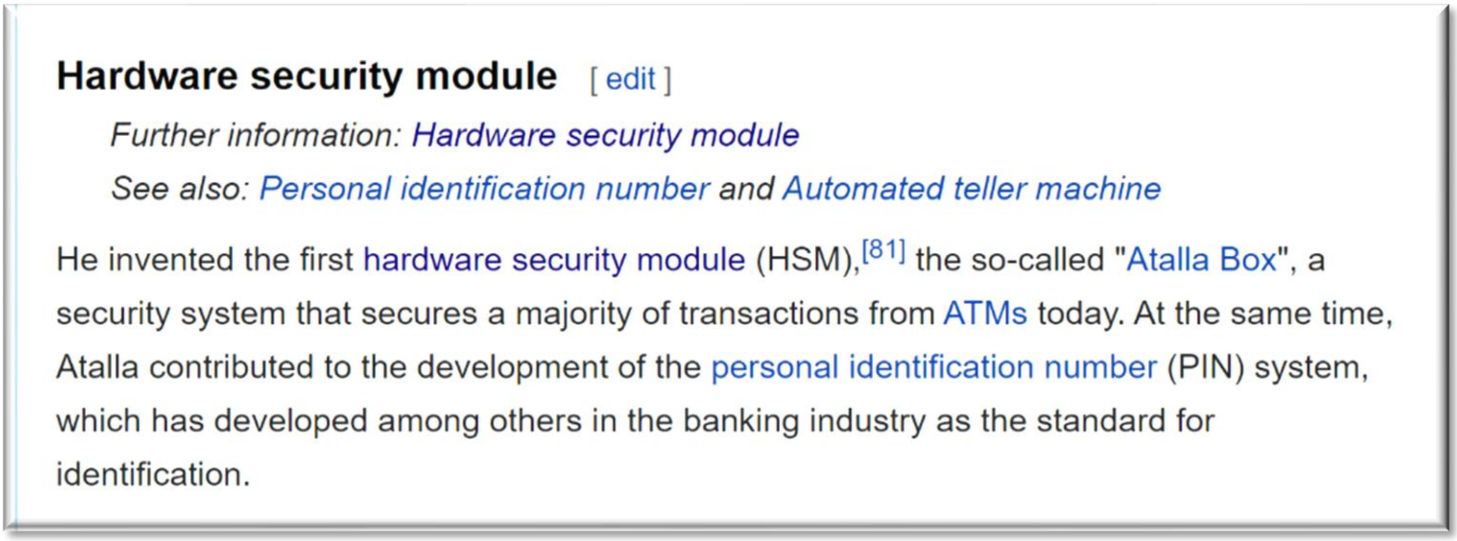


FATHER OF THE PIN



History

The hardware security module (HSM), a type of secure cryptoprocessor, was invented by Egyptian-American engineer Mohamed M. Atalla, in 1972. He invented a high security module dubbed the “Atalla Box” which encrypted PIN and ATM messages, and protected offline devices with an un-guessable PIN-generating key. In 1972, he filed a patent for the device. He founded Atalla Corporation (now Utimaco Atalla) that year, and commercialized the “Atalla Box” the following year, officially as the Identikey system. It was a card reader and customer identification system, consisting of a card reader console, two customer PIN pads, intelligent controller and built-in electronic interface package. It allowed the customer to type in a secret code, which is transformed by the device, using a microprocessor, into another code for the teller. During a transaction, the customer’s account number was read by the card reader. It was a success, and led to the wide use of high security modules.

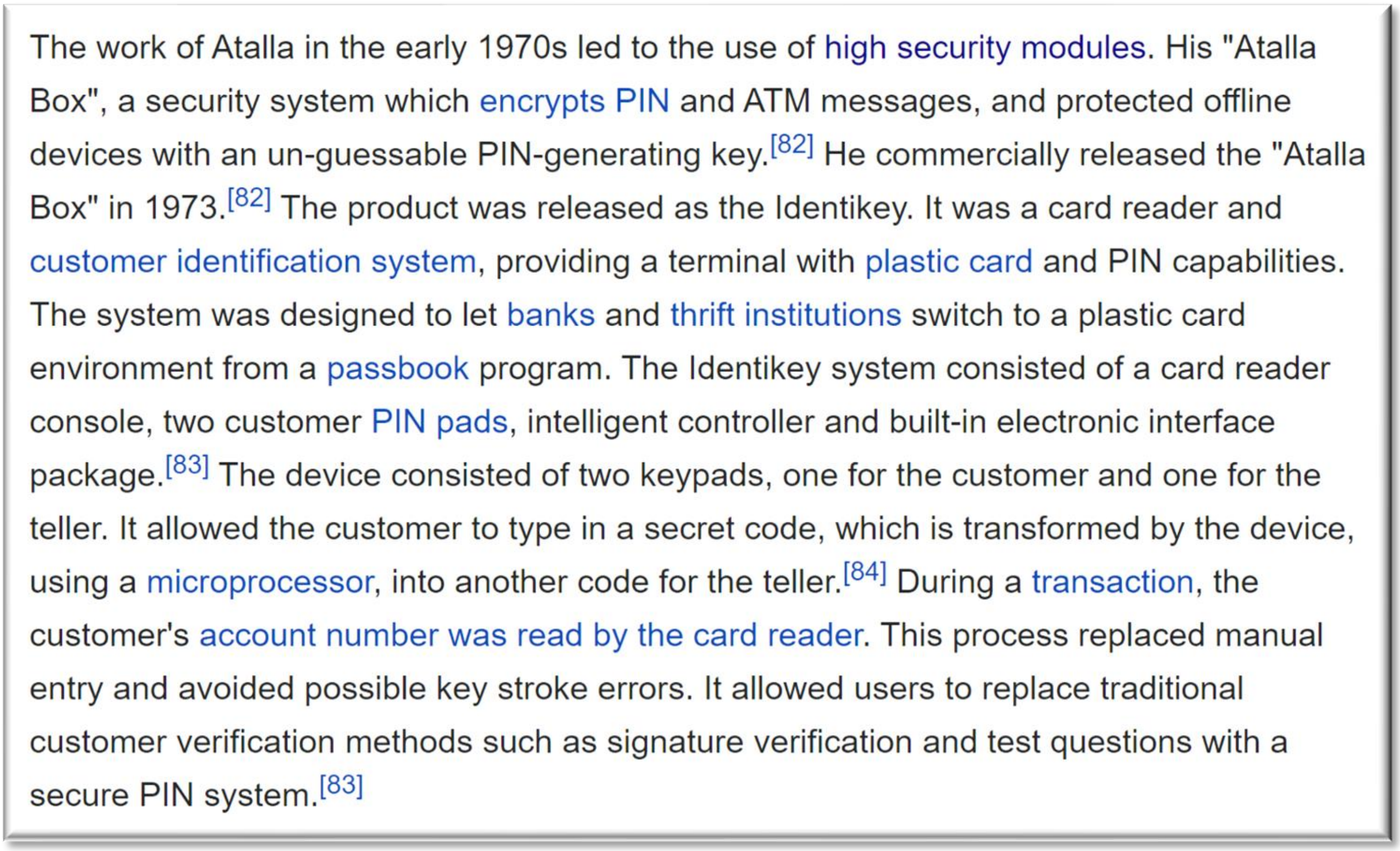


Hardware security module [edit]

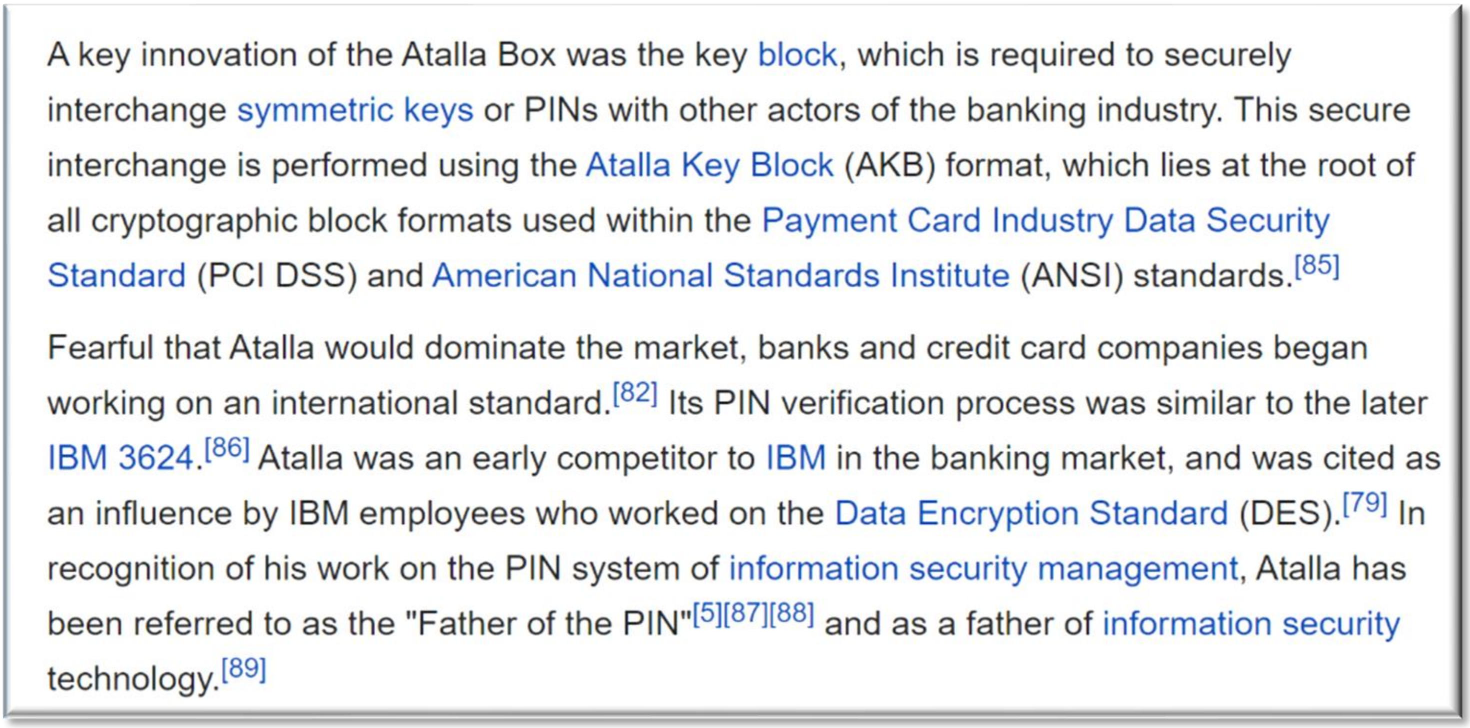
Further information: Hardware security module

See also: Personal identification number and Automated teller machine

He invented the first hardware security module (HSM),[81] the so-called "Atalla Box", a security system that secures a majority of transactions from ATMs today. At the same time, Atalla contributed to the development of the personal identification number (PIN) system, which has developed among others in the banking industry as the standard for identification.

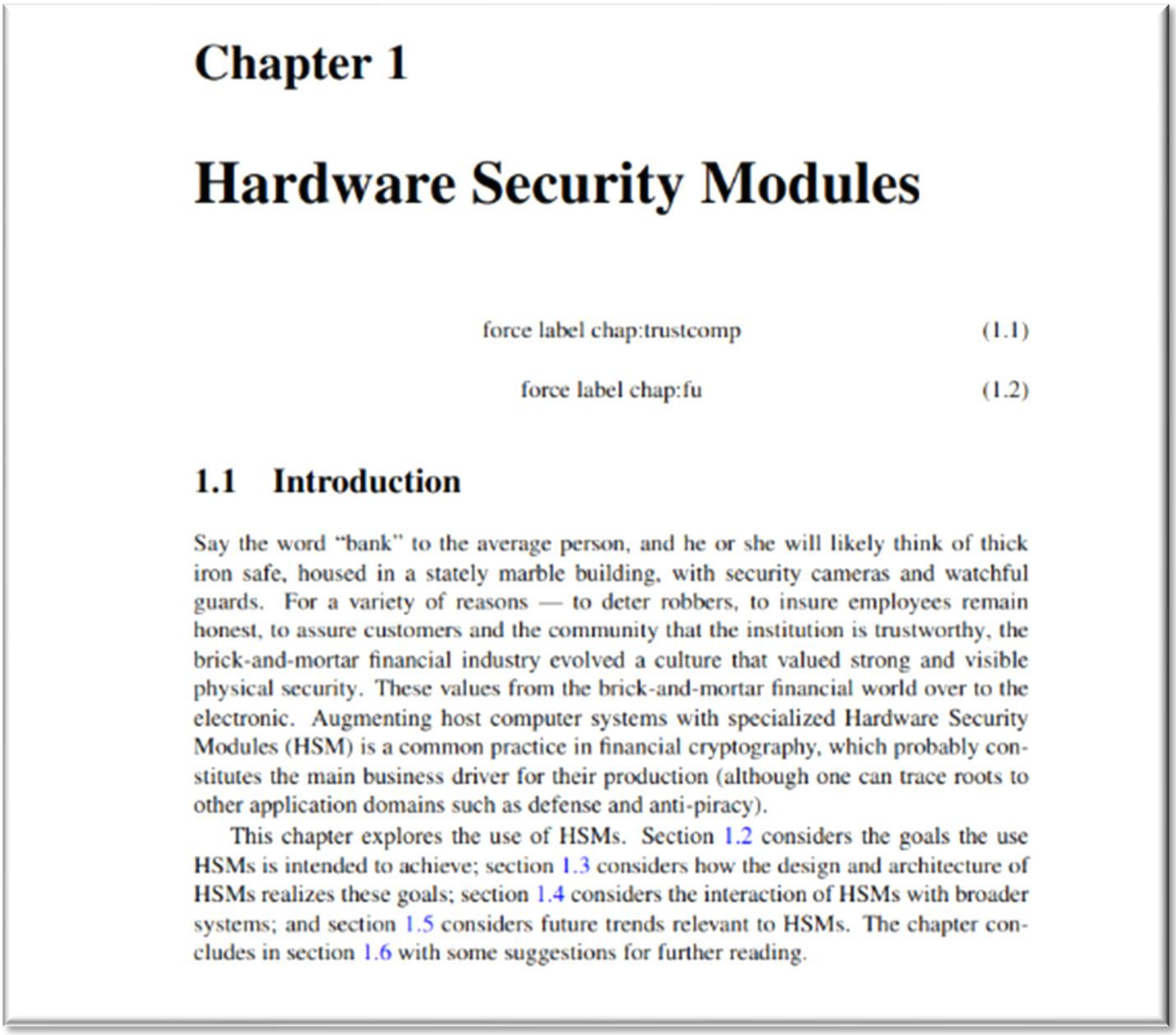


The work of Atalla in the early 1970s led to the use of high security modules. His "Atalla Box", a security system which encrypts PIN and ATM messages, and protected offline devices with an un-guessable PIN-generating key.[82] He commercially released the "Atalla Box" in 1973.[82] The product was released as the Identikey. It was a card reader and customer identification system, providing a terminal with plastic card and PIN capabilities. The system was designed to let banks and thrift institutions switch to a plastic card environment from a passbook program. The Identikey system consisted of a card reader console, two customer PIN pads, intelligent controller and built-in electronic interface package.[83] The device consisted of two keypads, one for the customer and one for the teller. It allowed the customer to type in a secret code, which is transformed by the device, using a microprocessor, into another code for the teller.[84] During a transaction, the customer's account number was read by the card reader. This process replaced manual entry and avoided possible key stroke errors. It allowed users to replace traditional customer verification methods such as signature verification and test questions with a secure PIN system.[83]



A key innovation of the Atalla Box was the key block, which is required to securely interchange symmetric keys or PINs with other actors of the banking industry. This secure interchange is performed using the Atalla Key Block (AKB) format, which lies at the root of all cryptographic block formats used within the Payment Card Industry Data Security Standard (PCI DSS) and American National Standards Institute (ANSI) standards.[85]

Fearful that Atalla would dominate the market, banks and credit card companies began working on an international standard.[82] Its PIN verification process was similar to the later IBM 3624.[86] Atalla was an early competitor to IBM in the banking market, and was cited as an influence by IBM employees who worked on the Data Encryption Standard (DES).[79] In recognition of his work on the PIN system of information security management, Atalla has been referred to as the "Father of the PIN"[5][87][88] and as a father of information security technology.[89]



Chapter 1

Hardware Security Modules

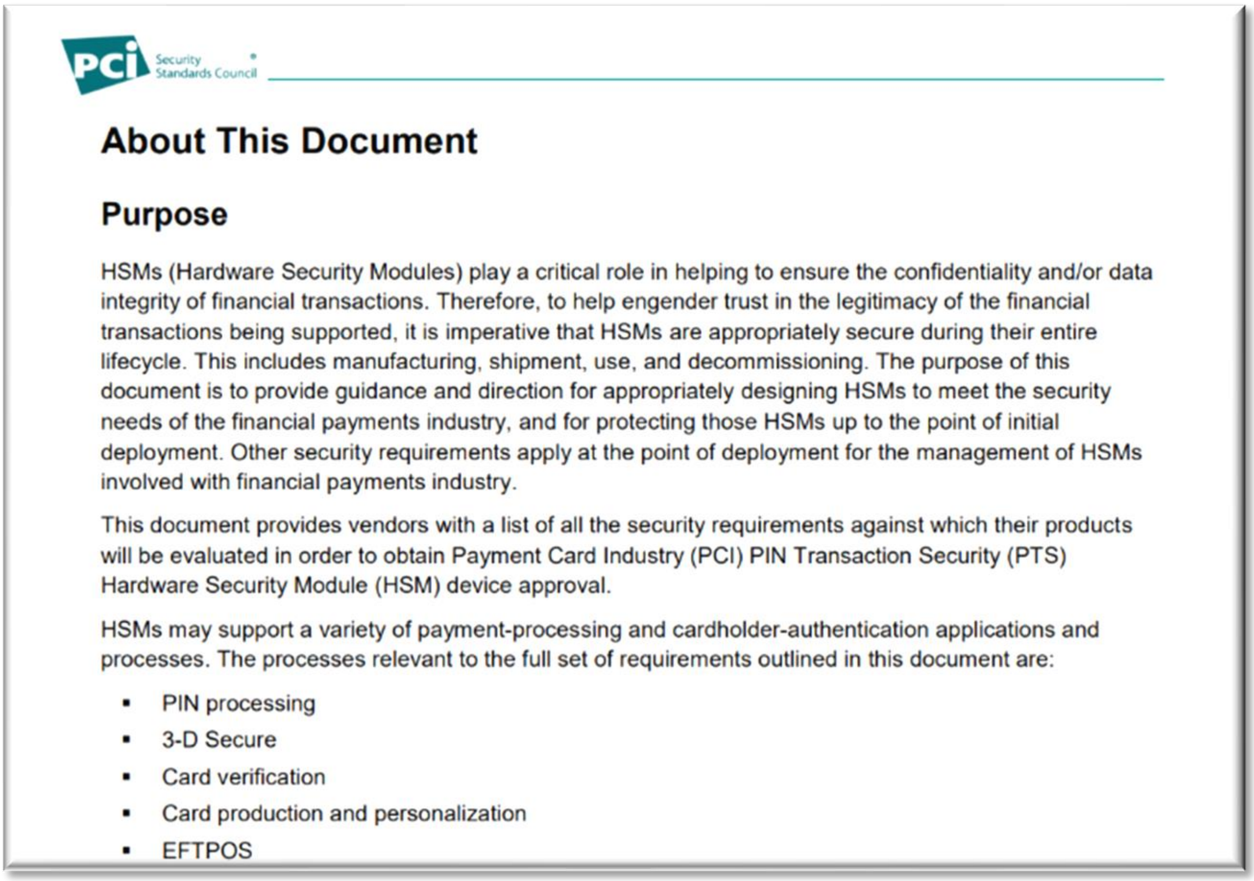
force label chap:trustcomp (1.1)

force label chap:fu (1.2)

1.1 Introduction

Say the word "bank" to the average person, and he or she will likely think of thick iron safe, housed in a stately marble building, with security cameras and watchful guards. For a variety of reasons — to deter robbers, to insure employees remain honest, to assure customers and the community that the institution is trustworthy, the brick-and-mortar financial industry evolved a culture that valued strong and visible physical security. These values from the brick-and-mortar financial world over to the electronic. Augmenting host computer systems with specialized Hardware Security Modules (HSM) is a common practice in financial cryptography, which probably constitutes the main business driver for their production (although one can trace roots to other application domains such as defense and anti-piracy).

This chapter explores the use of HSMs. Section 1.2 considers the goals the use HSMs is intended to achieve; section 1.3 considers how the design and architecture of HSMs realizes these goals; section 1.4 considers the interaction of HSMs with broader systems; and section 1.5 considers future trends relevant to HSMs. The chapter concludes in section 1.6 with some suggestions for further reading.



About This Document

Purpose

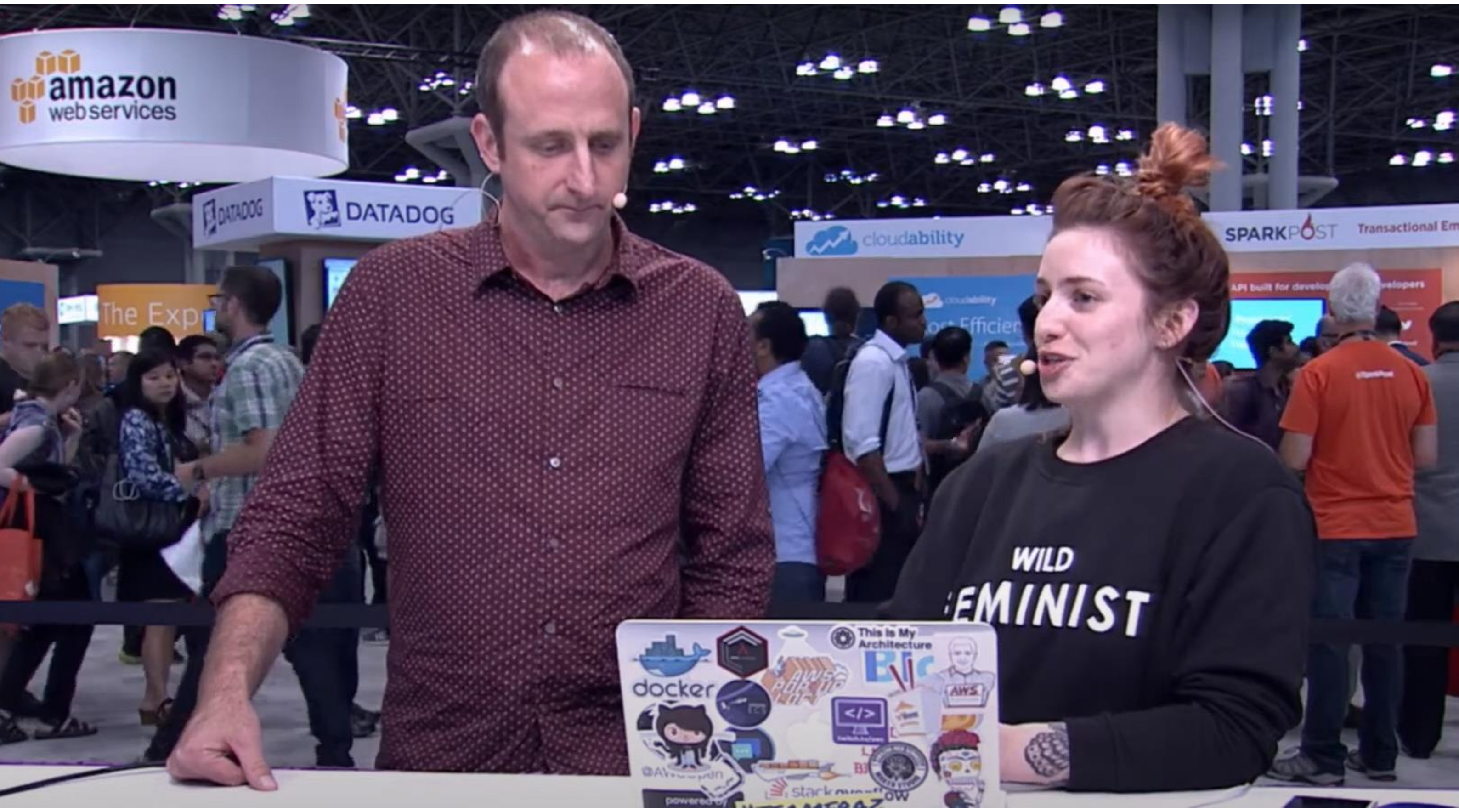
HSMs (Hardware Security Modules) play a critical role in helping to ensure the confidentiality and/or data integrity of financial transactions. Therefore, to help engender trust in the legitimacy of the financial transactions being supported, it is imperative that HSMs are appropriately secure during their entire lifecycle. This includes manufacturing, shipment, use, and decommissioning. The purpose of this document is to provide guidance and direction for appropriately designing HSMs to meet the security needs of the financial payments industry, and for protecting those HSMs up to the point of initial deployment. Other security requirements apply at the point of deployment for the management of HSMs involved with financial payments industry.

This document provides vendors with a list of all the security requirements against which their products will be evaluated in order to obtain Payment Card Industry (PCI) PIN Transaction Security (PTS) Hardware Security Module (HSM) device approval.

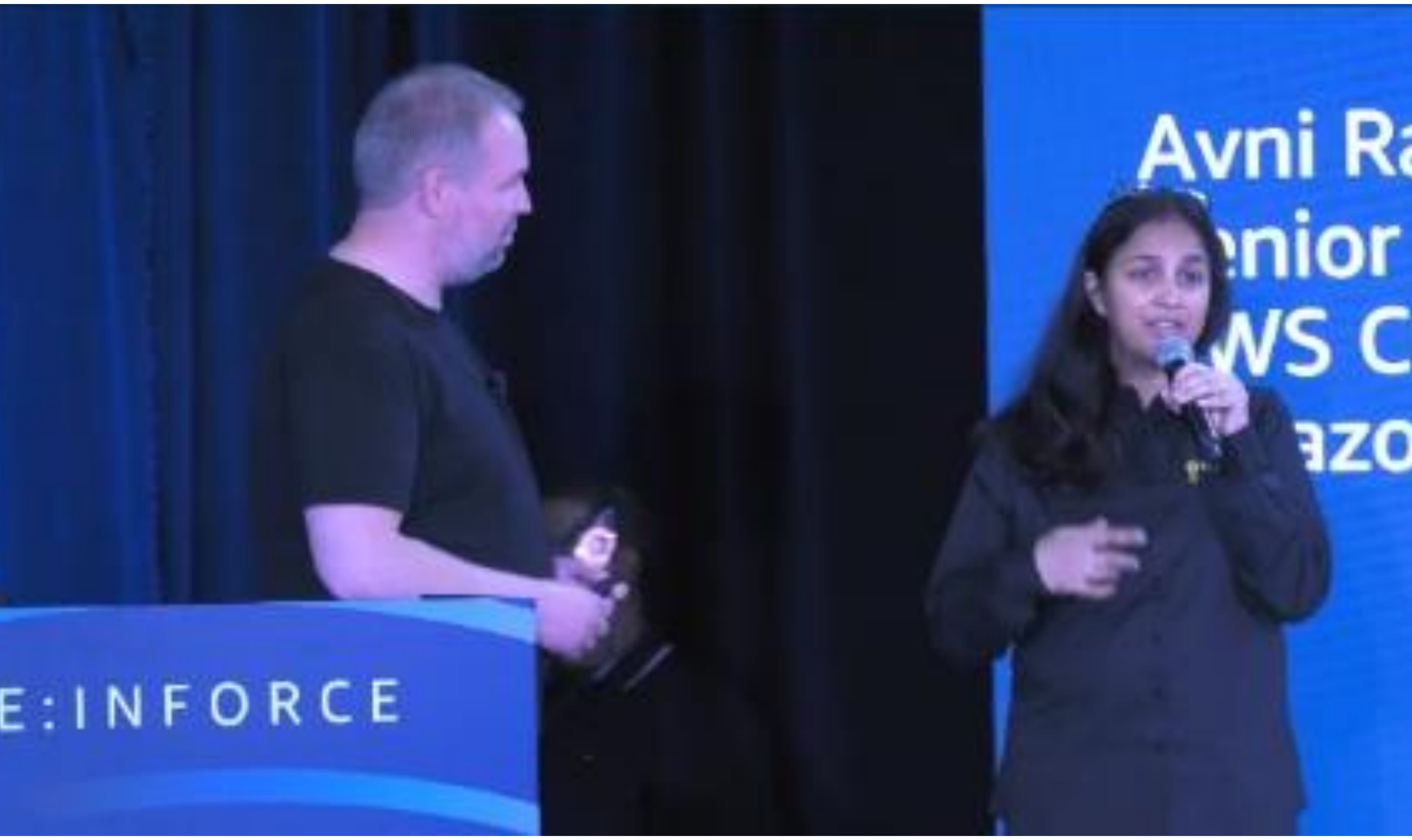
HSMs may support a variety of payment-processing and cardholder-authentication applications and processes. The processes relevant to the full set of requirements outlined in this document are:

- PIN processing
- 3-D Secure
- Card verification
- Card production and personalization
- EFTPOS

Todd Cignetti speaking in 2014. This was 12th November in Las Vegas, at the Reinvent conference.

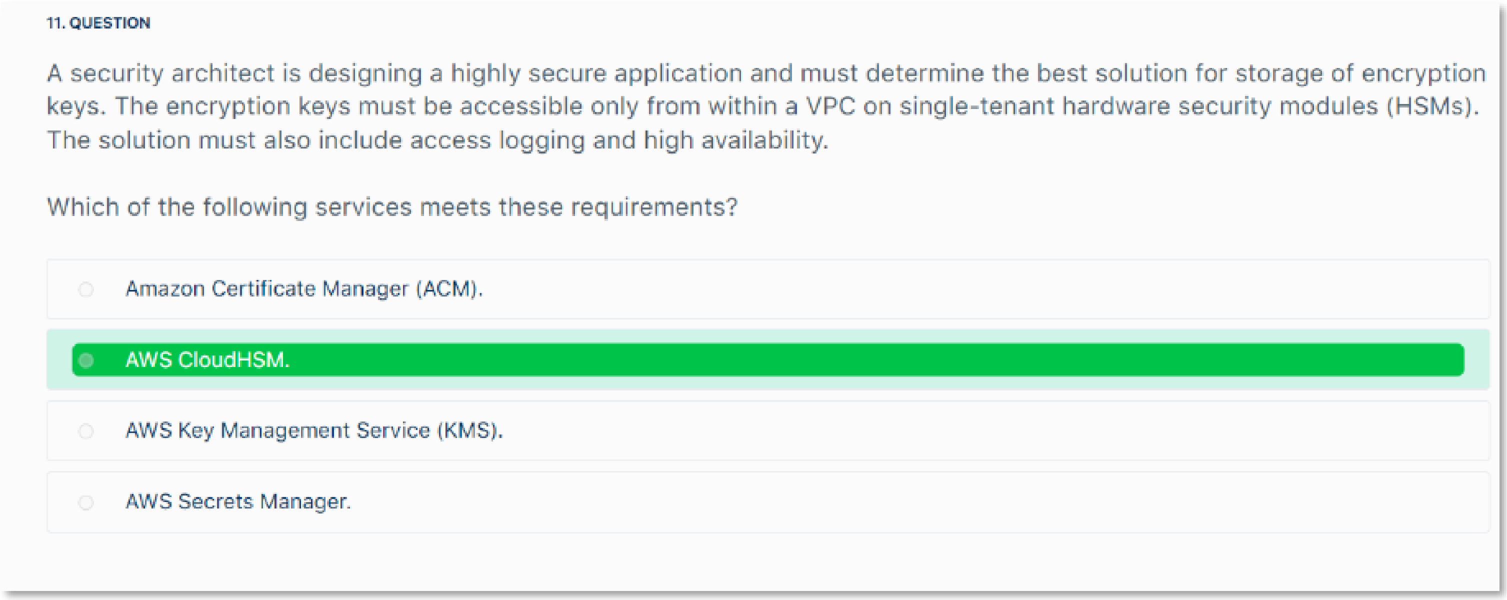


Bill Shinn and Abby Fuller. The new release of AWS CloudHSM in 2017



"Squigg" and Avni Rambhia explaining AWS CloudHSM in Summer 2019.

Stephen Quigg speaking at AWS Reinvent 2019 (SEC 305-R1). This conference took place between 2nd Dec and 6th Dec in Las Vegas (2019).



11. QUESTION

A security architect is designing a highly secure application and must determine the best solution for storage of encryption keys. The encryption keys must be accessible only from within a VPC on single-tenant hardware security modules (HSMs). The solution must also include access logging and high availability.

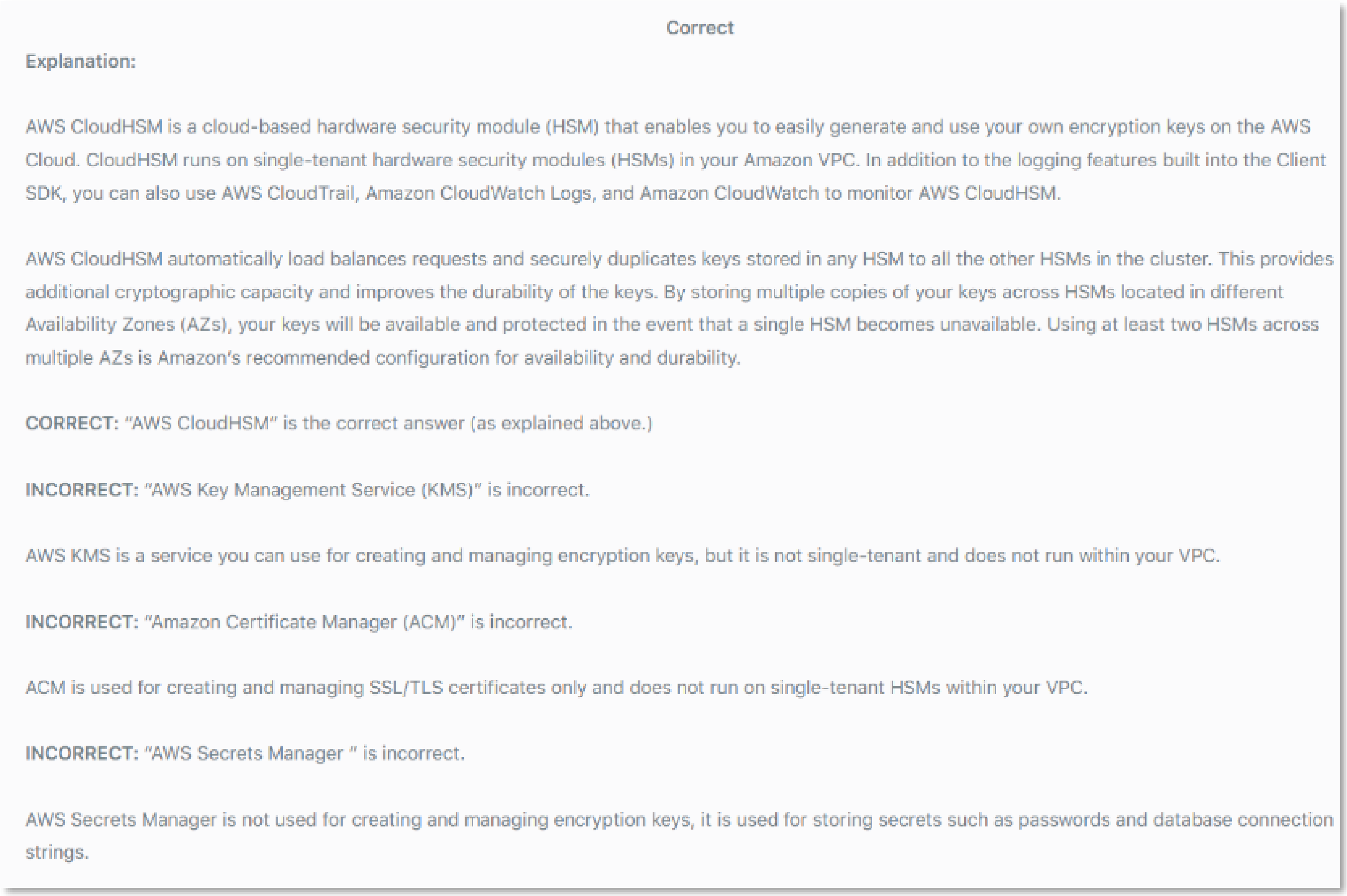
Which of the following services meets these requirements?

○ Amazon Certificate Manager (ACM).

● AWS CloudHSM.

○ AWS Key Management Service (KMS).

○ AWS Secrets Manager.



Correct

Explanation:

AWS CloudHSM is a cloud-based hardware security module (HSM) that enables you to easily generate and use your own encryption keys on the AWS Cloud. CloudHSM runs on single-tenant hardware security modules (HSMs) in your Amazon VPC. In addition to the logging features built into the Client SDK, you can also use AWS CloudTrail, Amazon CloudWatch Logs, and Amazon CloudWatch to monitor AWS CloudHSM.

AWS CloudHSM automatically load balances requests and securely duplicates keys stored in any HSM to all the other HSMs in the cluster. This provides additional cryptographic capacity and improves the durability of the keys. By storing multiple copies of your keys across HSMs located in different Availability Zones (AZs), your keys will be available and protected in the event that a single HSM becomes unavailable. Using at least two HSMs across multiple AZs is Amazon's recommended configuration for availability and durability.

CORRECT: "AWS CloudHSM" is the correct answer (as explained above.)

INCORRECT: "AWS Key Management Service (KMS)" is incorrect.

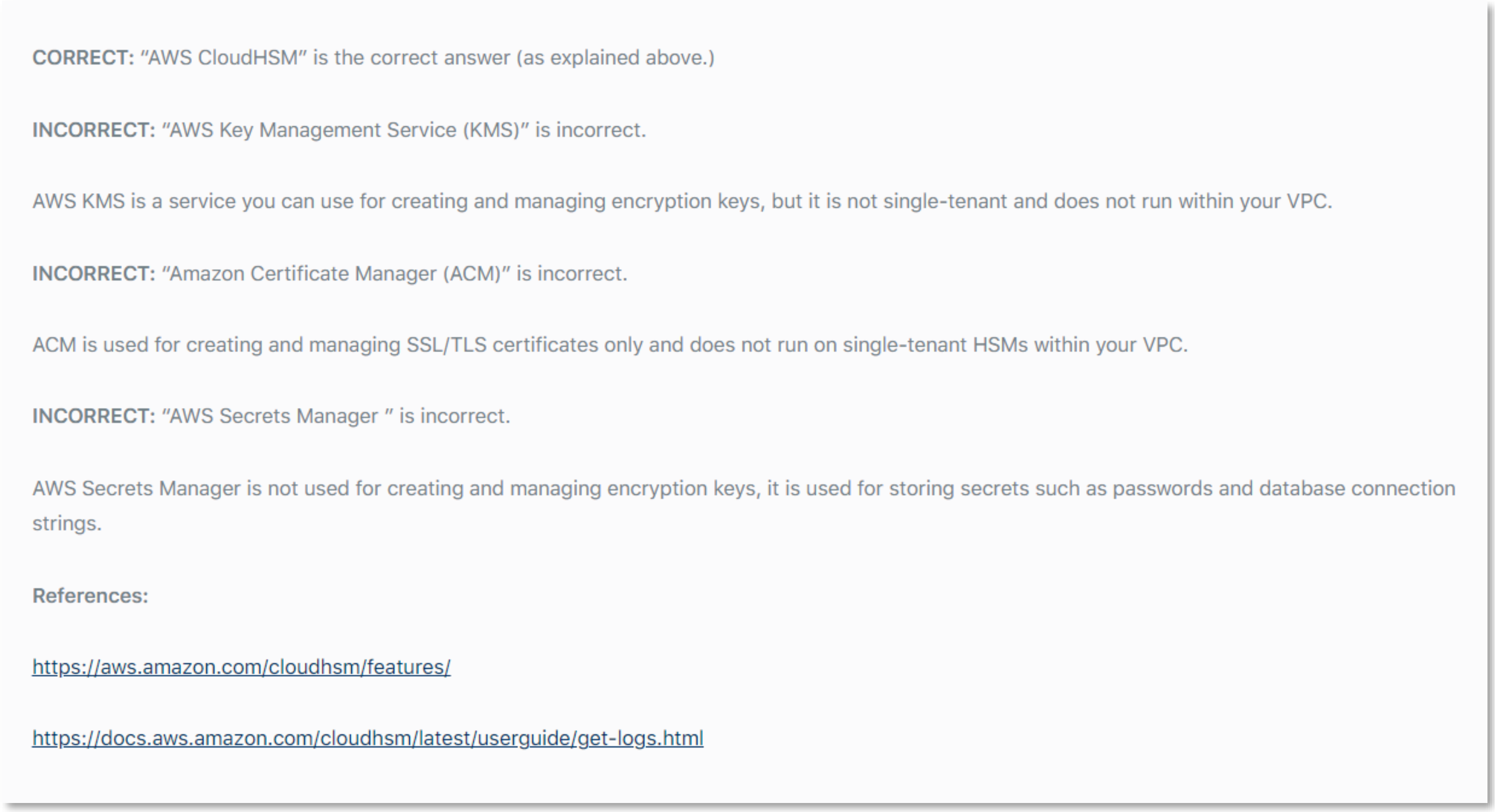
AWS KMS is a service you can use for creating and managing encryption keys, but it is not single-tenant and does not run within your VPC.

INCORRECT: "Amazon Certificate Manager (ACM)" is incorrect.

ACM is used for creating and managing SSL/TLS certificates only and does not run on single-tenant HSMs within your VPC.

INCORRECT: "AWS Secrets Manager " is incorrect.

AWS Secrets Manager is not used for creating and managing encryption keys, it is used for storing secrets such as passwords and database connection strings.



CORRECT: "AWS CloudHSM" is the correct answer (as explained above.)

INCORRECT: "AWS Key Management Service (KMS)" is incorrect.

AWS KMS is a service you can use for creating and managing encryption keys, but it is not single-tenant and does not run within your VPC.

INCORRECT: "Amazon Certificate Manager (ACM)" is incorrect.

ACM is used for creating and managing SSL/TLS certificates only and does not run on single-tenant HSMs within your VPC.

INCORRECT: "AWS Secrets Manager " is incorrect.

AWS Secrets Manager is not used for creating and managing encryption keys, it is used for storing secrets such as passwords and database connection strings.

References:

https://aws.amazon.com/cloudhsm/features/

https://docs.aws.amazon.com/cloudhsm/latest/userguide/get-logs.html

Review Questions

Glossary

ECDSA

Elliptic Curve Digital Signature Algorithm.

ttt

Java Database Connectivity.

ttt

Description of what term means here.

ttt

Description of what term means here.

ttt

Description of what term means here.

Bibliography

I. Official

[AWS 2022]

AWS. “What is AWS CloudHSM?” 4st Aug 2022. YouTube. Available at: <https://www.youtube.com/watch?v=BLnuUtjJNLE&ab_channel=AmazonWebServices>.

[Avni 2017]

Rambhia, Avni (2017). CloudHSM: Secure Scalable Key Storage in AWS – 2017 Online Tech Talks. 25th Oct 2017. YouTube. Available at: <https://www.youtube.com/watch?v=hEVks207ALM&ab_channel=AWSOnlineTechTalks>.

[Shinn 2017]

Shinn, Bill (2017). Live from the NY Summit. YouTube. Available at: <https://www.youtube.com/watch?v=0fpPESoFCew&ab_channel=AmazonWebServices>

[Sayyed 2018]

Sayyed, Shafreen (2018). AWS Security by Design. 10th May 2018. YouTube. Available at: <https://www.youtube.com/watch?v=I1SwoKxB13c&ab_channel=AmazonWebServices>

[Cignetti 2013]

Cignetti, Todd and Ken Beer and Jason Chan (2013). Encryption and Key Management in AWS. Reinvent conference 2013 (SEC 304). 15th November 2013. [Channel: Amazon Web Services] Available at: https://www.youtube.com/watch?v=-v7rsCWUC1I&ab_channel=AmazonWebServices

[Cignetti 2014]

Cignetti, Todd and Ken Beer (2014). Encryption and Key Management in AWS. Reinvent conference (SEC 301). 12th November 2014. [Channel: Amazon Web Services]. Available at: <https://www.youtube.com/watch?v=bqIYI3mDsd4&ab_channel=AmazonWebServices>

[Quigg 2019a]

Quigg, Stephen and Avni Rambhia (2019). Achieving Security Goals with AWS CloudHSM. AWS Reinvent Conference (SDD333). Available at: https://www.youtube.com/watch?v=_gezaWmwzYY&t=7s&ab_channel=AmazonWebServices

[Quigg 2019b]

Quigg, Stephen and Avni Rambhia (2019). AWS Cryptography Services: Selecting the Right Tool for the Job. Reinvent Conference. December 2019. (SEC305-R1). Available at: https://www.youtube.com/watch?v=Vox-PDRHIUs&ab_channel=AWSEvents

[Chowdhary]

Chowdhary, Ankush (2017). Understanding AWS CloudHSM and AWS WAF. AWS Public Sector Summit, Singapore. Available at: <https://www.youtube.com/watch?v=IMxImFoVpmI&ab_channel=AmazonWebServices>

[Samaha 2015]

Samaha, Camil (2015). Protecting your data with AWS KMS and AWS CloudHSM. AWS Government, Education and Nonprofits Symposium. 25-26 July 2015. Available at: <https://www.youtube.com/watch?v=hCxkWGAWJCM&ab_channel=AmazonWebServices>

II. Unofficial

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:
<URL here>.

[BitLovers]

AWS KMS Vs CloudHSM. *Bit's Lovers*. Available at: https://www.bitslovers.com/aws-kms-vs-cloudhsm/#:~:text=The%20difference%20between%20KMS%20and,not%20exclusive%20only%20for%20you.

[Cignetti 2013]

Cignetti, Todd (2013). Why CloudHSM can revolutionize AWS. YouTube Channel: LASCON. Available at: <https://www.youtube.com/watch?v=BgBecjA-QlM&ab_channel=LASCON>

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

IV. General

https://www.cardlogix.com/glossary/hardware-security-module-hsm/

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

https://www.youtube.com/watch?v=RL9IDuYL7BI&ab_channel=Cryptosense

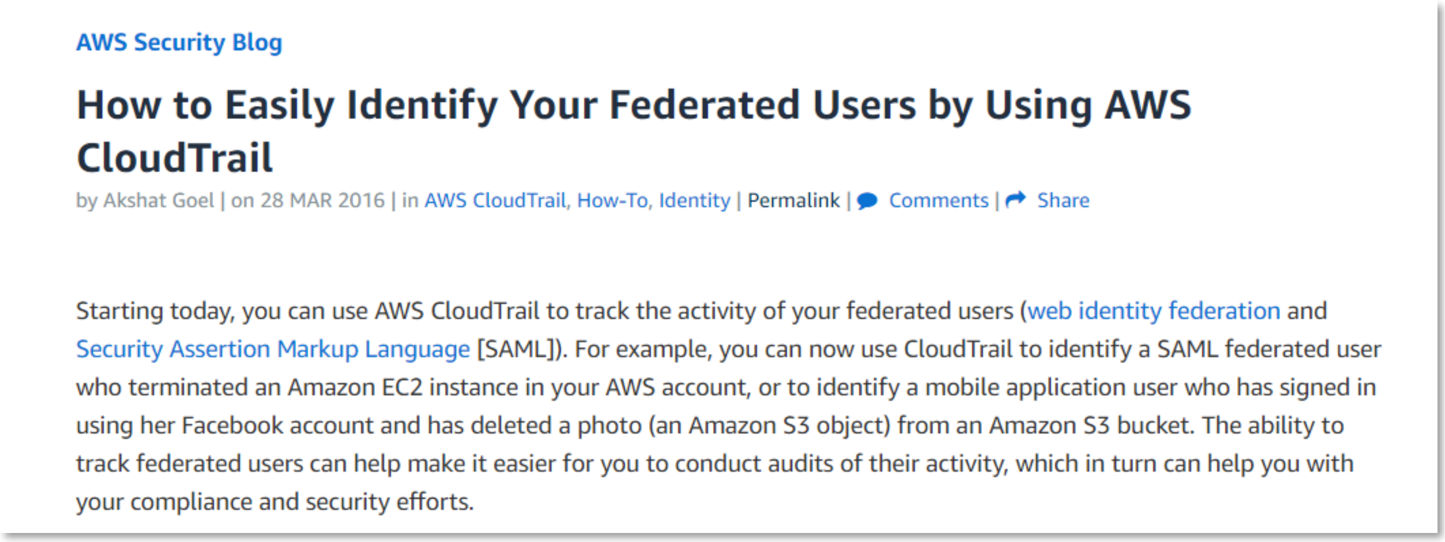
https://www.cs.dartmouth.edu/~sws/pubs/hsm-draft.pdf

AppStream

CloudTrail

Sivakanth Mundru explaining the idea of CloudTrail at Reinvent 2013

CloudTrail



AWS Security Blog

How to Easily Identify Your Federated Users by Using AWS CloudTrail

by Akshat Goel | on 28 MAR 2016 | in AWS CloudTrail, How-To, Identity | Permalink | Comments | Share

Starting today, you can use AWS CloudTrail to track the activity of your federated users (web identity federation and Security Assertion Markup Language [SAML]). For example, you can now use CloudTrail to identify a SAML federated user who terminated an Amazon EC2 instance in your AWS account, or to identify a mobile application user who has signed in using her Facebook account and has deleted a photo (an Amazon S3 object) from an Amazon S3 bucket. The ability to track federated users can help make it easier for you to conduct audits of their activity, which in turn can help you with your compliance and security efforts.

A 2016 article by Akshat Goel. In this article, Goel shows how CloudTrail can be used to identity an employee (Bob) who terminated an EC2 instance

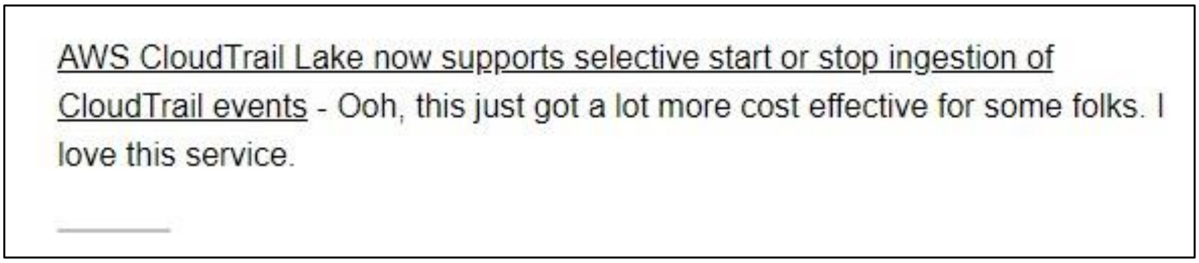
AWS CloudTrail Lake now supports selective start or stop ingestion of CloudTrail events

Posted On: Jun 5, 2023

AWS CloudTrail Lake now provides the ability to selectively start or stop ingestion of CloudTrail events into your CloudTrail Lake event data store. This capability enables you to collect events only for a specific time window for troubleshooting or security analysis without having to delete or recreate the event data store. When you stop ingestion, the event data store continues to retain ingested events based on its retention period. For audit purposes, CloudTrail generates events that capture the start and stop ingestion activity.

You can enable this feature in the CloudTrail console, by using the AWS Software Development Kits (SDKs), or AWS Command Line Interface (CLI). This feature is available in the following AWS Regions: US East (N. Virginia), US East (Ohio), US West (N. California), US West (Oregon), Canada (Central), Africa (Cape Town), Asia Pacific (Hong Kong), Asia Pacific (Hyderabad), Asia Pacific (Jakarta), Asia Pacific (Melbourne), Asia Pacific (Mumbai), Asia Pacific (Osaka), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), Europe (London), Europe (Milan), Europe (Paris), Europe (Stockholm), Middle East (Bahrain), Middle East (UAE), South America (São Paulo), AWS GovCloud (US-East), and AWS GovCloud (US-West).

To get started, see Working with CloudTrail Lake in the CloudTrail User Guide.



AWS CloudTrail Lake now supports selective start or stop ingestion of CloudTrail events - Ooh, this just got a lot more cost effective for some folks. I love this service.

Email from Corey Quinn on June 12th 2023

Bibliography

I. Official

[AWS 2023]

AWS CloudTrail Lake now supports selective start of stop ingestion of CloudTrail events. [Announcement]. Available at: <https://aws.amazon.com/about-aws/whats-new/2023/06/aws-cloudtrail-lake-start-stop-ingestion-cloudtrail-events/?ck_subscriber_id=1560524742>

[AWS 2024]

AWS CloudTrail Lake announces AI-powered natural language query generation. June 11th 2024. Available at: <https://aws.amazon.com/about-aws/whats-new/2024/06/aws-cloudtrail-lake-ai-powered-query-generation-preview/?ck_subscriber_id=1560524742>

II. Unofficial

III. Critical

IV. General

WorkSpaces

Desktop and Application Streaming

Empowering your workforce with Amazon WorkSpaces services and Microsoft 365

by Dilip Kumar | on 01 AUG 2023 | in Amazon WorkSpaces, Amazon WorkSpaces Core, Announcements, Desktop & Application Streaming, End User Computing | Permalink | Share

Tens of thousands of customers use Amazon WorkSpaces services as their end-user computing service to provide secure, scalable, and cost-effective virtual desktops with all the tools required for users to do their jobs. Customers use all sorts of applications on their WorkSpaces virtual desktops, from single-purpose web apps to guide warehouse workers to complex rendering workloads based on GPUs in media and entertainment and everything in between. Many customers tell us they need solutions that include a simple and cost-effective mechanism for delivering productivity applications like Microsoft 365 to hybrid and remote users, while improving security posture for corporate data and intellectual property.

Bibliography

I. Official
II. Unofficial
III. Critical
IV. General

I. Official

[Kumar 2023]

Empowering your workforce with Amazon WorkSpaces services Microsoft 365. Available at: <https://aws.amazon.com/blogs/desktop-and-application-streaming/empowering-your-workforce-with-amazon-workspaces-services-and-microsoft-365/?ck_subscriber_id=1560524742>

AWS Certifications

April 30, 2013

Amazon Web Services Announces Launch of Certification Program for AWS Cloud Computing Professionals

New program addresses growing demand for IT pros with demonstrated knowledge of AWS best practices worldwide

SEATTLE--(BUSINESS WIRE)--Apr. 30, 2013-- Amazon Web Services, Inc. (AWS), an Amazon.com company (NASDAQ: AMZN), today announced the launch of the new AWS Certification Program with the first of several exams that will made available in 2013. With the accelerating adoption of cloud computing and the AWS Cloud around the world, organizations are increasingly seeking mechanisms to identify candidates and consultants with demonstrated knowledge of AWS best practices. The new AWS Certification Program helps to fill this need to recognize IT professionals that possess the skills and technical knowledge necessary for building and maintaining applications and services on the AWS Cloud. To learn more about the AWS Certification Program, visit http://aws.amazon.com/certification.

AWS Certifications help to recognize the skills, knowledge and expertise of IT professionals in designing, deploying and managing applications on the AWS platform. To earn an AWS Certification, individuals must demonstrate their proficiency in a particular area by passing an AWS Certification Exam. Individuals looking to prepare for an exam can attend courses through AWS Training to help gain proficiency with AWS services. Individuals that pass an AWS Certification Exam can display the applicable AWS Certified logo on business cards and resumes to gain visibility for their AWS expertise while fostering credibility with employers and peers.

The first available AWS Certification Exam is for the “AWS Certified Solutions Architect – Associate Level” certification, which tests skills for technical professionals and solutions architects involved in the design and development of applications on AWS. Additional role-based certifications, including certifications for Systems Operations (SysOps) Administrators and Developers, will follow later this year.

AWS Certification Exams are administered through testing centers in more than 100 countries and 750 testing locations worldwide.

Additional benefits of the AWS Certification Program include:

Helps organizations to identify engineering and/or IT staff with the skills and technical knowledge necessary for building and maintaining solutions on the AWS Cloud;

Tests an individual's IT skills and technical knowledge are in alignment with AWS's architectural best practices for building highly secure and reliable cloud applications;

Increases differentiation for AWS Partner Network (APN) members that have AWS Certified individuals on staff;

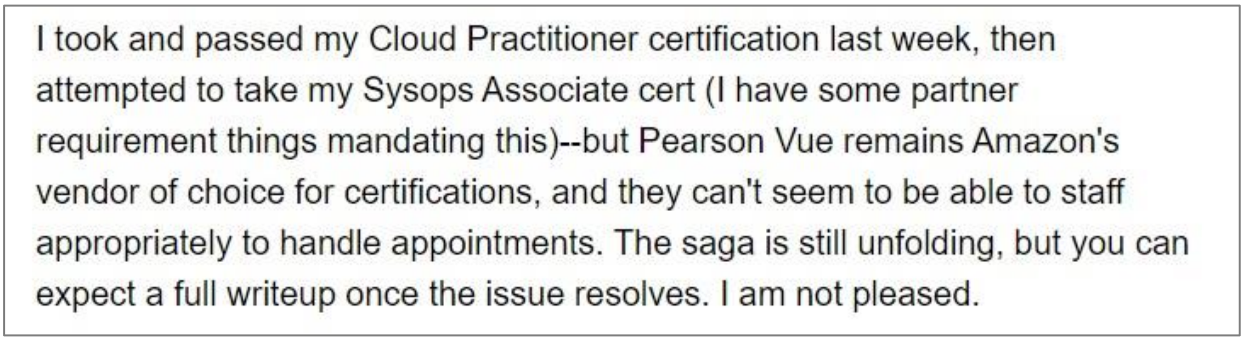
Allows technical professionals to develop, certify and advertise their expertise with cloud computing on AWS.

"With cloud computing being quickly adopted by organizations of all sizes around the world, in-depth training programs as well as certifications for individuals who have demonstrated competence with AWS are increasingly important," said Adam Selipsky, Vice President, Amazon Web Services. "The AWS Certification Program helps organizations identify that the employees, partners and consultants they depend on for their AWS solutions are well-versed in the best practices of building cloud applications on AWS and have the skills to help them be successful."

"We have many mission critical programs running on the AWS Cloud across our various business groups at Samsung, so training with AWS has been a priority for the whole company for some time," said Seok Kyun Choi, Head of Training Division at Samsung SDS. "The new AWS Certification program will allow us to recognize our employees who have expanded their skills as well as identify that the partners we are working with have solid knowledge in building AWS-based applications."

"Assisting clients with their AWS implementations is increasingly strategic to the future of our business," said Joseph Coyle, North America Chief Technology Officer, Capgemini. "As we've been expanding our team of AWS-trained professionals to keep up with the growing demand from our enterprise clients, AWS Certifications will allow us to further differentiate our cloud computing practice in the market and provide an added level of assurance to our clients."

"With a rapidly growing demand for cloud services, the industry needs trained and certified resources to help businesses unlock the optimal value of cloud computing in driving transformation, innovation and competitiveness," said Kaushik Bhaumik, Senior Vice President for Technology, Industry and Alliances at Cognizant. "As a member of the AWS Partner Network (APN), we have been leveraging AWS training programs to bring the right talent to customer engagements. The AWS Certification Program will help us further enhance and differentiate our cloud capabilities in helping customers run better and run different."



I took and passed my Cloud Practitioner certification last week, then attempted to take my Sysops Associate cert (I have some partner requirement things mandating this)--but Pearson Vue remains Amazon's vendor of choice for certifications, and they can't seem to be able to staff appropriately to handle appointments. The saga is still unfolding, but you can expect a full writeup once the issue resolves. I am not pleased.

Corey Quinn writing on July 1st 2024

Bibliography

I. Official

II. Unofficial

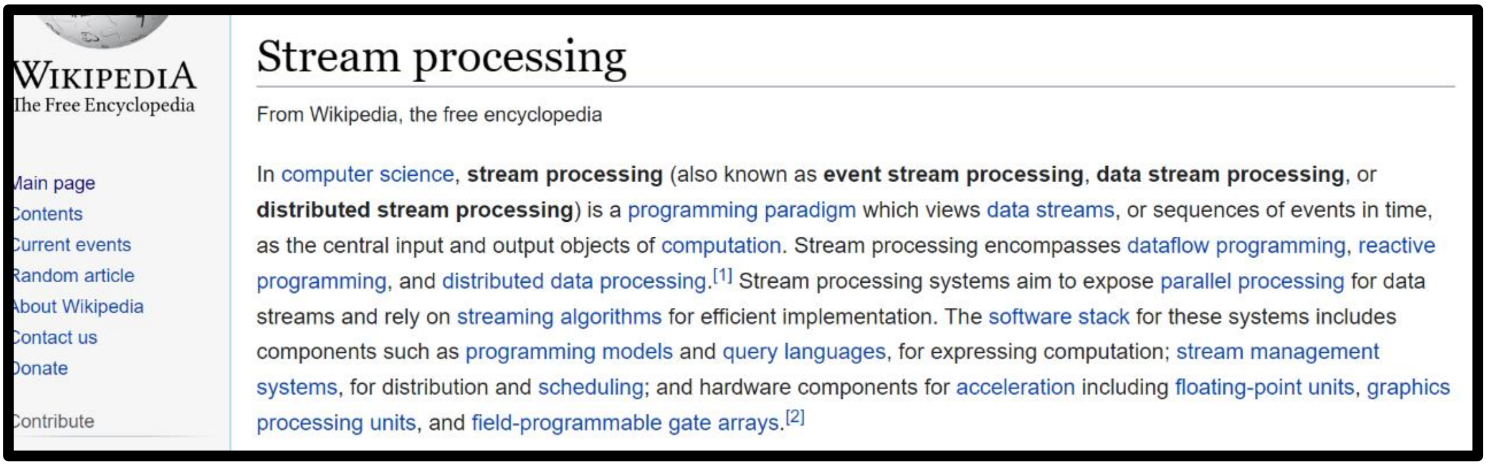
[Quinn 2024]

Quinn, Corey (2024). *Last Week in AWS* [Newsletter]. July 1st 2024.

III. Critical

IV. General

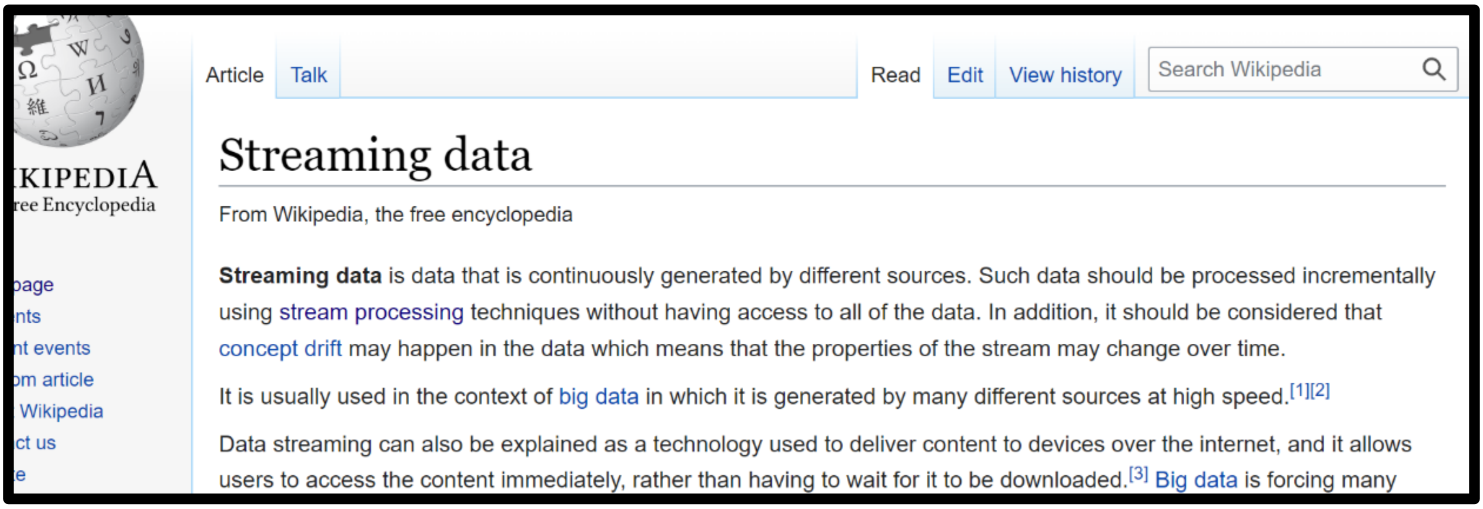
Kinesis



Stream processing

From Wikipedia, the free encyclopedia

In computer science, **stream processing** (also known as **event stream processing**, **data stream processing**, or **distributed stream processing**) is a programming paradigm which views data streams, or sequences of events in time, as the central input and output objects of computation. Stream processing encompasses dataflow programming, reactive programming, and distributed data processing.[1] Stream processing systems aim to expose parallel processing for data streams and rely on streaming algorithms for efficient implementation. The software stack for these systems includes components such as programming models and query languages, for expressing computation; stream management systems, for distribution and scheduling; and hardware components for acceleration including floating-point units, graphics processing units, and field-programmable gate arrays.[2]



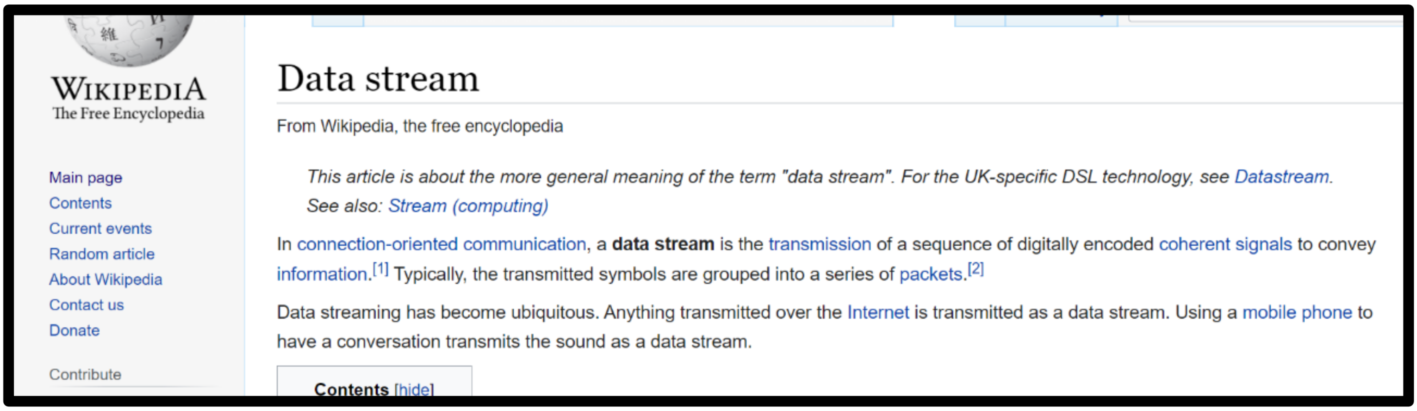
Streaming data

From Wikipedia, the free encyclopedia

Streaming data is data that is continuously generated by different sources. Such data should be processed incrementally using stream processing techniques without having access to all of the data. In addition, it should be considered that concept drift may happen in the data which means that the properties of the stream may change over time.

It is usually used in the context of big data in which it is generated by many different sources at high speed.[1][2]

Data streaming can also be explained as a technology used to deliver content to devices over the internet, and it allows users to access the content immediately, rather than having to wait for it to be downloaded.[3] Big data is forcing many



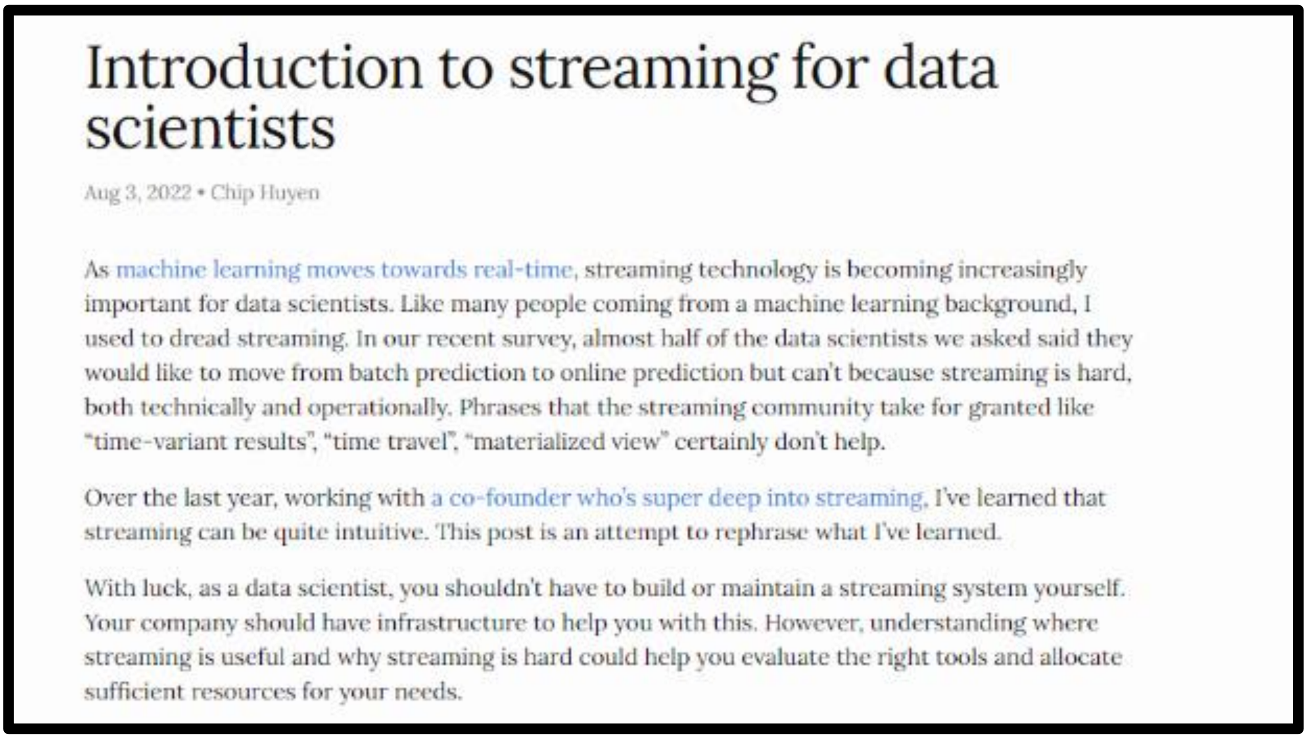
Data stream

From Wikipedia, the free encyclopedia

This article is about the more general meaning of the term "data stream". For the UK-specific DSL technology, see Datastream. See also: Stream (computing)

In connection-oriented communication, a **data stream** is the transmission of a sequence of digitally encoded coherent signals to convey information.[1] Typically, the transmitted symbols are grouped into a series of packets.[2]

Data streaming has become ubiquitous. Anything transmitted over the Internet is transmitted as a data stream. Using a mobile phone to have a conversation transmits the sound as a data stream.



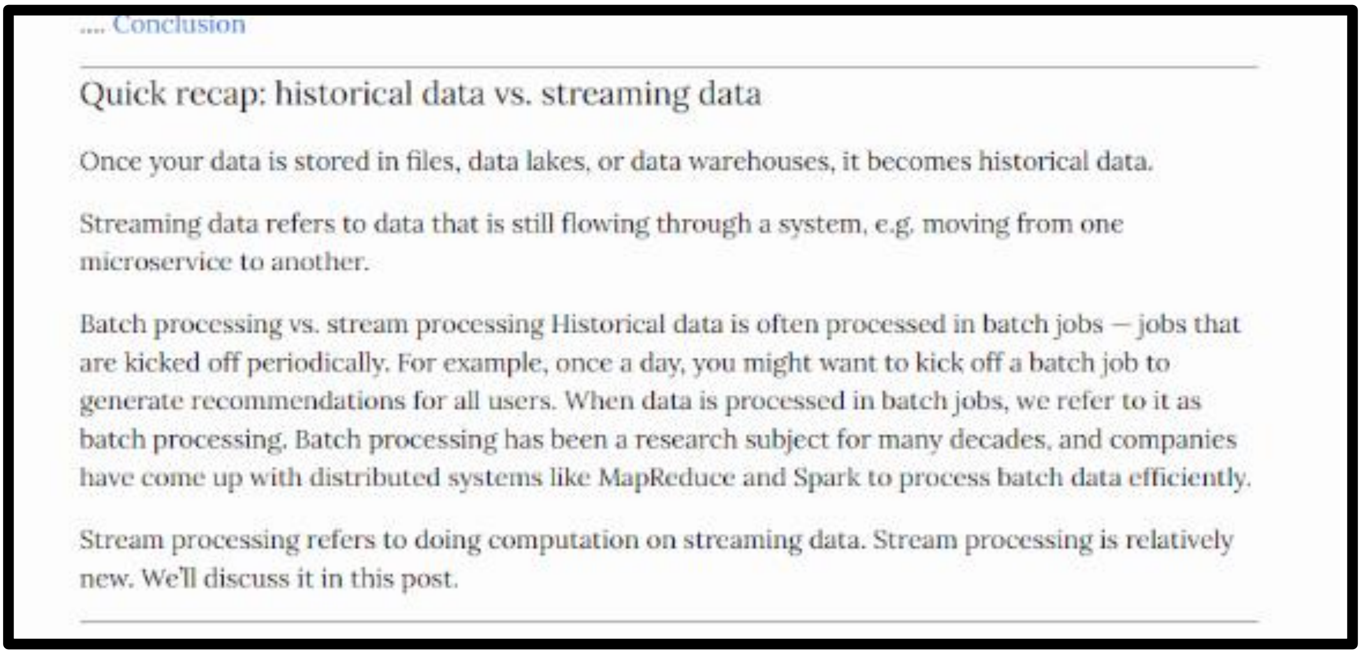
Introduction to streaming for data scientists

Aug 3, 2022 • Chip Huyen

As machine learning moves towards real-time, streaming technology is becoming increasingly important for data scientists. Like many people coming from a machine learning background, I used to dread streaming. In our recent survey, almost half of the data scientists we asked said they would like to move from batch prediction to online prediction but can't because streaming is hard, both technically and operationally. Phrases that the streaming community take for granted like "time-variant results", "time travel", "materialized view" certainly don't help.

Over the last year, working with a co-founder who's super deep into streaming, I've learned that streaming can be quite intuitive. This post is an attempt to rephrase what I've learned.

With luck, as a data scientist, you shouldn't have to build or maintain a streaming system yourself. Your company should have infrastructure to help you with this. However, understanding where streaming is useful and why streaming is hard could help you evaluate the right tools and allocate sufficient resources for your needs.



.... Conclusion

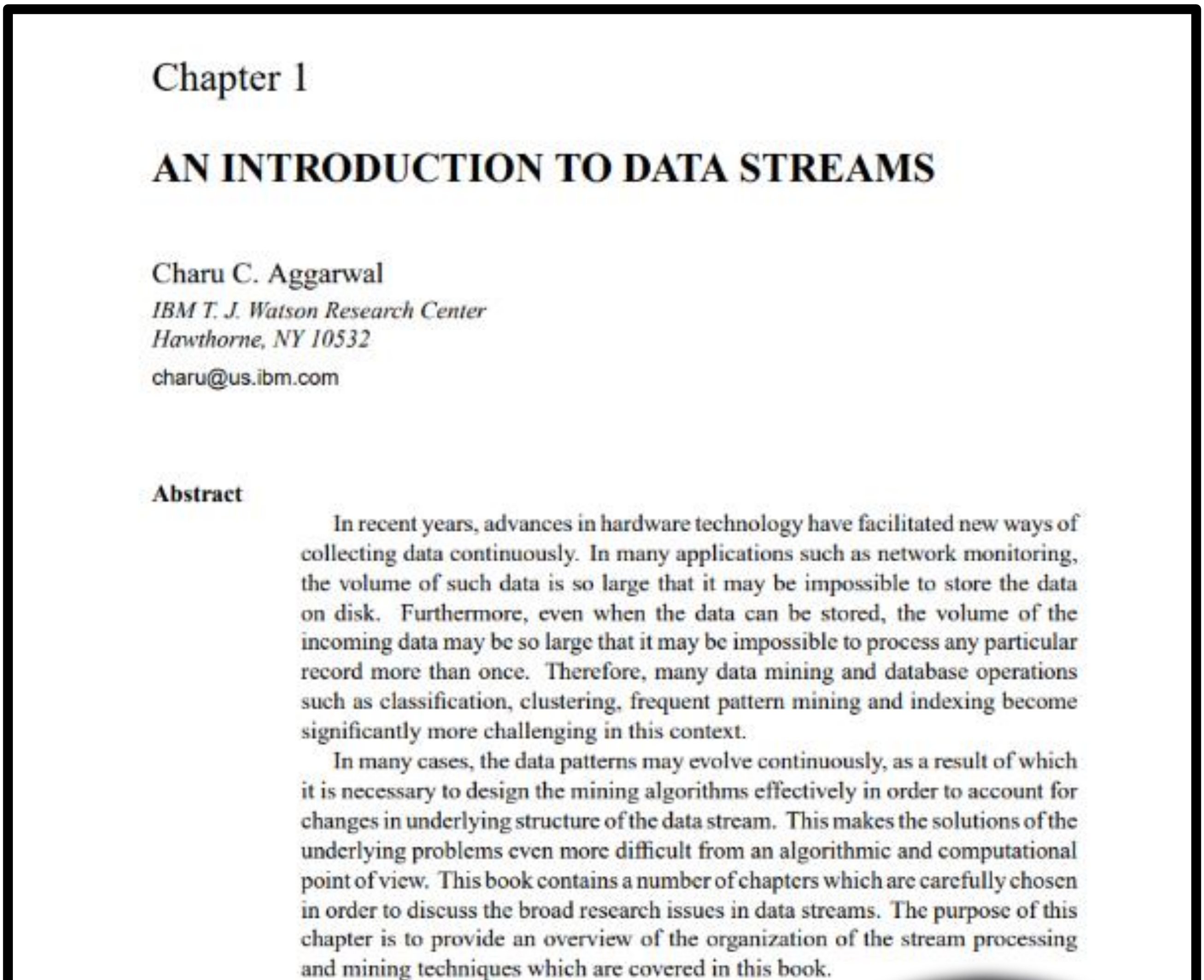
Quick recap: historical data vs. streaming data

Once your data is stored in files, data lakes, or data warehouses, it becomes historical data.

Streaming data refers to data that is still flowing through a system, e.g. moving from one microservice to another.

Batch processing vs. stream processing Historical data is often processed in batch jobs — jobs that are kicked off periodically. For example, once a day, you might want to kick off a batch job to generate recommendations for all users. When data is processed in batch jobs, we refer to it as batch processing. Batch processing has been a research subject for many decades, and companies have come up with distributed systems like MapReduce and Spark to process batch data efficiently.

Stream processing refers to doing computation on streaming data. Stream processing is relatively new. We'll discuss it in this post.



Chapter 1

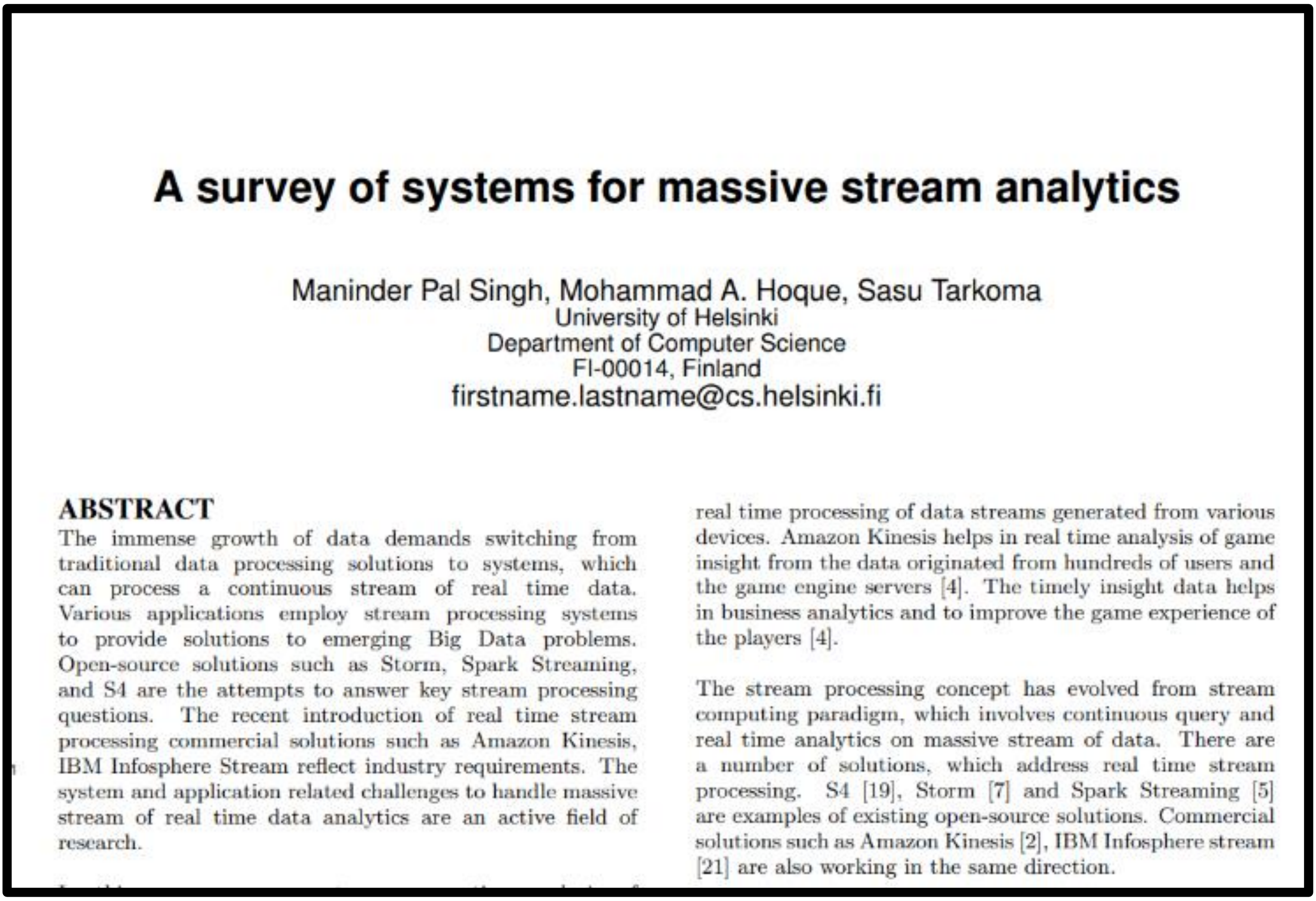
AN INTRODUCTION TO DATA STREAMS

Charu C. Aggarwal
IBM T. J. Watson Research Center
Hawthorne, NY 10532
charu@us.ibm.com

Abstract

In recent years, advances in hardware technology have facilitated new ways of collecting data continuously. In many applications such as network monitoring, the volume of such data is so large that it may be impossible to store the data on disk. Furthermore, even when the data can be stored, the volume of the incoming data may be so large that it may be impossible to process any particular record more than once. Therefore, many data mining and database operations such as classification, clustering, frequent pattern mining and indexing become significantly more challenging in this context.

In many cases, the data patterns may evolve continuously, as a result of which it is necessary to design the mining algorithms effectively in order to account for changes in underlying structure of the data stream. This makes the solutions of the underlying problems even more difficult from an algorithmic and computational point of view. This book contains a number of chapters which are carefully chosen in order to discuss the broad research issues in data streams. The purpose of this chapter is to provide an overview of the organization of the stream processing and mining techniques which are covered in this book.



A survey of systems for massive stream analytics

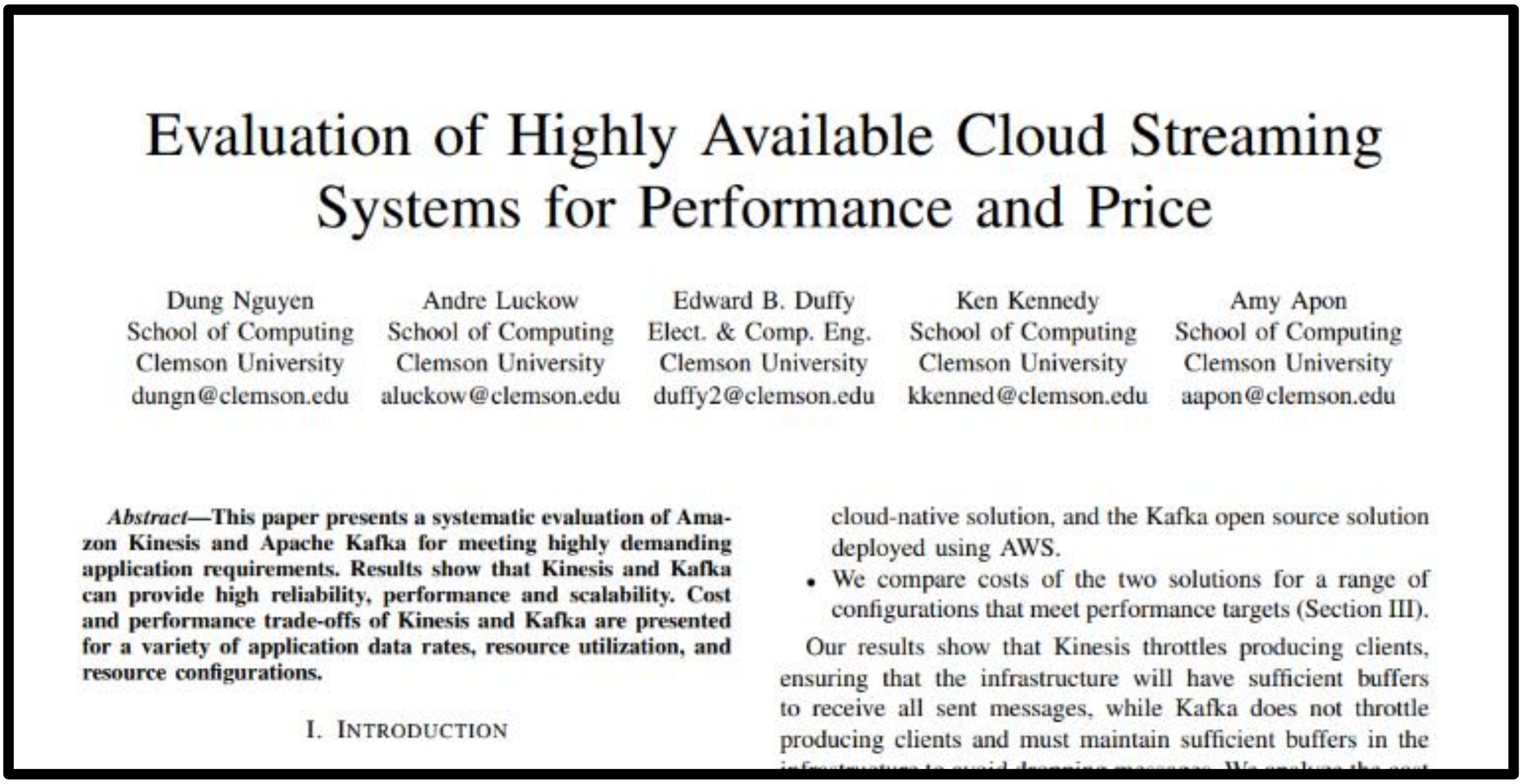
Maninder Pal Singh, Mohammad A. Hoque, Sasu Tarkoma
University of Helsinki
Department of Computer Science
FI-00014, Finland
firstname.lastname@cs.helsinki.fi

ABSTRACT

The immense growth of data demands switching from traditional data processing solutions to systems, which can process a continuous stream of real time data. Various applications employ stream processing systems to provide solutions to emerging Big Data problems. Open-source solutions such as Storm, Spark Streaming, and S4 are the attempts to answer key stream processing questions. The recent introduction of real time stream processing commercial solutions such as Amazon Kinesis, IBM Infosphere Stream reflect industry requirements. The system and application related challenges to handle massive stream of real time data analytics are an active field of research.

real time processing of data streams generated from various devices. Amazon Kinesis helps in real time analysis of game insight from the data originated from hundreds of users and the game engine servers [4]. The timely insight data helps in business analytics and to improve the game experience of the players [4].

The stream processing concept has evolved from stream computing paradigm, which involves continuous query and real time analytics on massive stream of data. There are a number of solutions, which address real time stream processing. S4 [19], Storm [7] and Spark Streaming [5] are examples of existing open-source solutions. Commercial solutions such as Amazon Kinesis [2], IBM Infosphere stream [21] are also working in the same direction.



Evaluation of Highly Available Cloud Streaming Systems for Performance and Price

Dung Nguyen
School of Computing
Clemson University
dungn@clemson.edu

Andre Luckow
School of Computing
Clemson University
aluckow@clemson.edu

Edward B. Duffy
Elect. & Comp. Eng.
Clemson University
duffy2@clemson.edu

Ken Kennedy
School of Computing
Clemson University
kkenned@clemson.edu

Amy Apon
School of Computing
Clemson University
aapon@clemson.edu

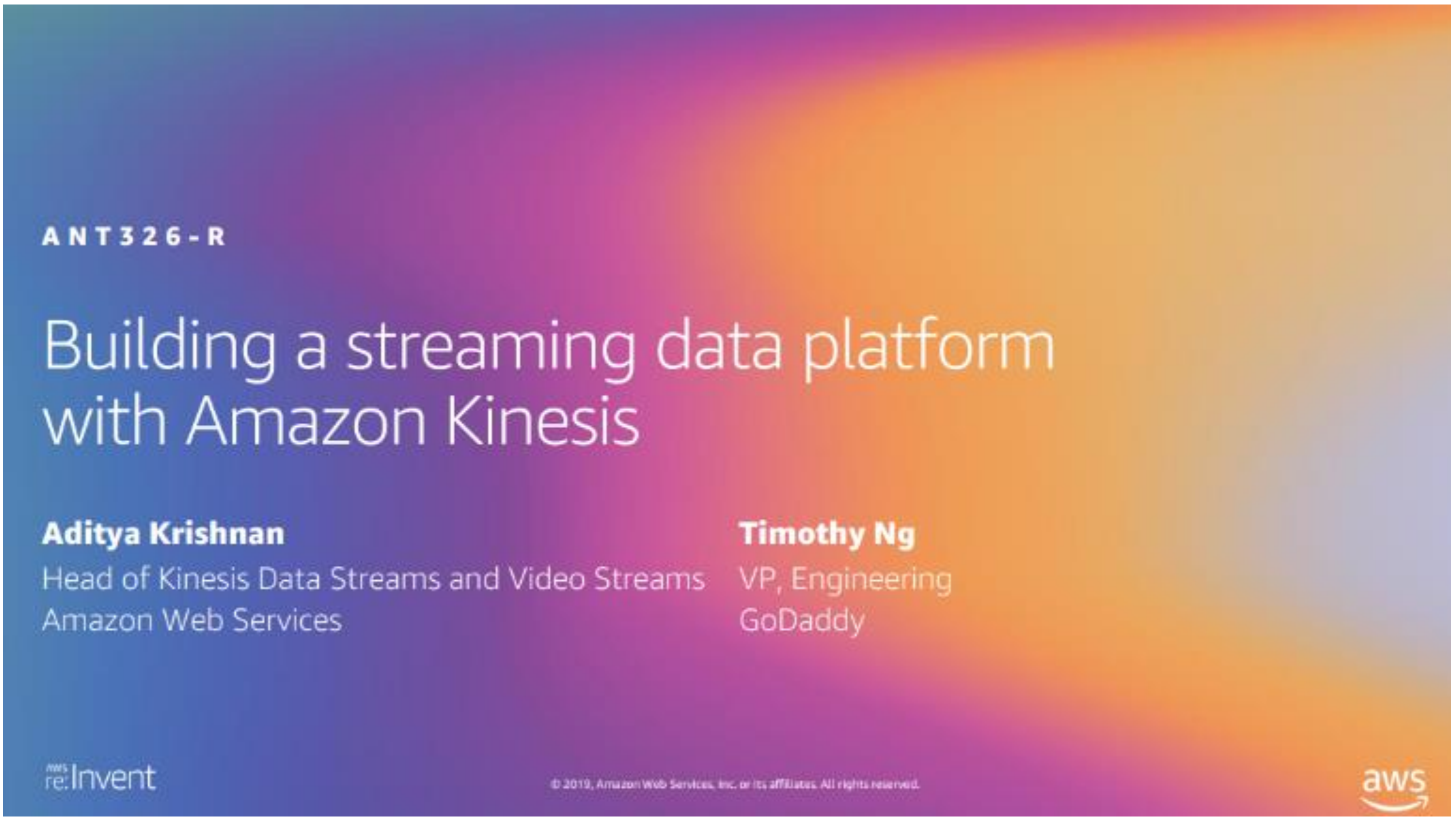
***Abstract*—This paper presents a systematic evaluation of Amazon Kinesis and Apache Kafka for meeting highly demanding application requirements. Results show that Kinesis and Kafka can provide high reliability, performance and scalability. Cost and performance trade-offs of Kinesis and Kafka are presented for a variety of application data rates, resource utilization, and resource configurations.**

I. Introduction

cloud-native solution, and the Kafka open source solution deployed using AWS.

- We compare costs of the two solutions for a range of configurations that meet performance targets (Section III).

Our results show that Kinesis throttles producing clients, ensuring that the infrastructure will have sufficient buffers to receive all sent messages, while Kafka does not throttle producing clients and must maintain sufficient buffers in the infrastructure to avoid dropping messages. We analyze the cost



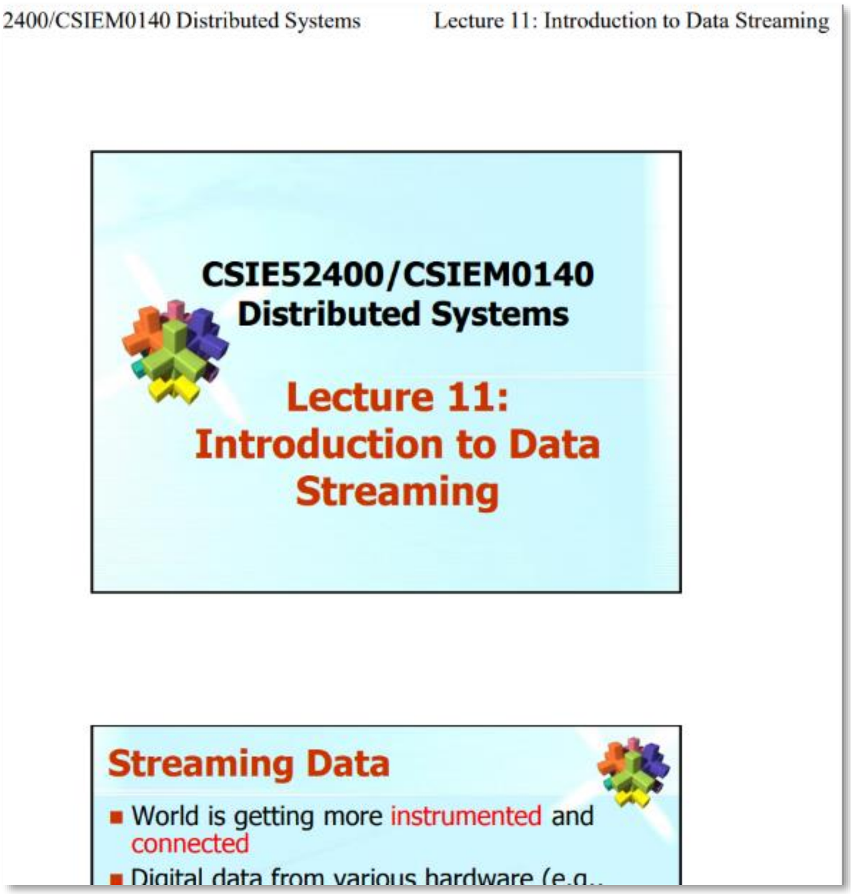
ANT326-R

Building a streaming data platform with Amazon Kinesis

Aditya Krishnan
Head of Kinesis Data Streams and Video Streams
Amazon Web Services

Timothy Ng
VP, Engineering
GoDaddy

re:Invent



2400/CSIEM0140 Distributed Systems Lecture 11: Introduction to Data Streaming

CSIE52400/CSIEM0140
Distributed Systems

Lecture 11:
Introduction to Data
Streaming

Streaming Data

- World is getting more instrumented and connected
- Digital data from various hardware (e.g.,

[Showyang 2022]

What is the difference between SQS and Kinesis data streams?

This confusion has lived a long time. It stems from the fact that XXXX But day we lay it to rest.

I'd like to say a few words if I may.

Aristotle distinguishes between two different senses of the term *dunamis*. In the strictest sense, a *dunamis* is the *power* that a thing has to produce a change. A thing has a *dunamis* in this sense when it has within it "a starting-point of change in another thing or in itself insofar as it is other" (Θ.1, 1046^{a}12; cf. Δ.12). The exercise of such a power is a *kinêsis*—a movement or process. So, for example, the housebuilder's craft is a power whose exercise is the process of housebuilding. But there is a second sense of *dunamis*—and it is the one in which Aristotle is mainly interested—that might be better translated as 'potentiality'. For, as Aristotle tells us, in this sense *dunamis* is related not to movement (*kinêsis*) but to activity (*energeia*)(Θ.6, 1048^{a}25). A *dunamis* in this sense is not a thing's power to produce a change but rather its capacity to be in a different and more completed state. Aristotle thinks that potentiality so understood is indefinable (1048^{a}37), claiming that the general idea can be grasped from a consideration of cases. Activity is to potentiality, Aristotle tells us, as "what is awake is in relation to what is asleep, and what is seeing is in relation to what has its eyes closed but has sight, and what has been shaped out of the matter is in relation to the matter" (1048^{b}1–3).

https://plato.stanford.edu/entries/aristotle-metaphysics/#ActuPote

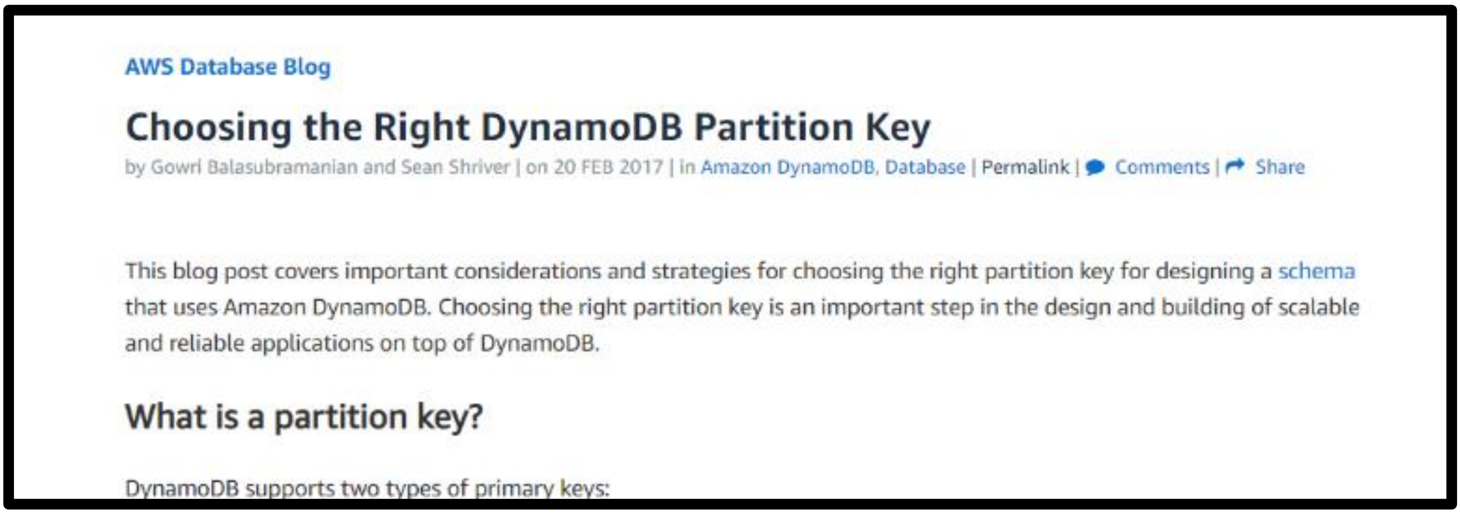
"REAL TIME"

SHARD

PARTITION KEY

What on earth is a "partition key"?

Recall that we discussed partition keys in the context of DynamoDB:



AWS Database Blog

Choosing the Right DynamoDB Partition Key

by Gowri Balasubramanian and Sean Shriver | on 20 FEB 2017 | in Amazon DynamoDB, Database | Permalink | Comments | Share

This blog post covers important considerations and strategies for choosing the right partition key for designing a schema that uses Amazon DynamoDB. Choosing the right partition key is an important step in the design and building of scalable and reliable applications on top of DynamoDB.

What is a partition key?

DynamoDB supports two types of primary keys:

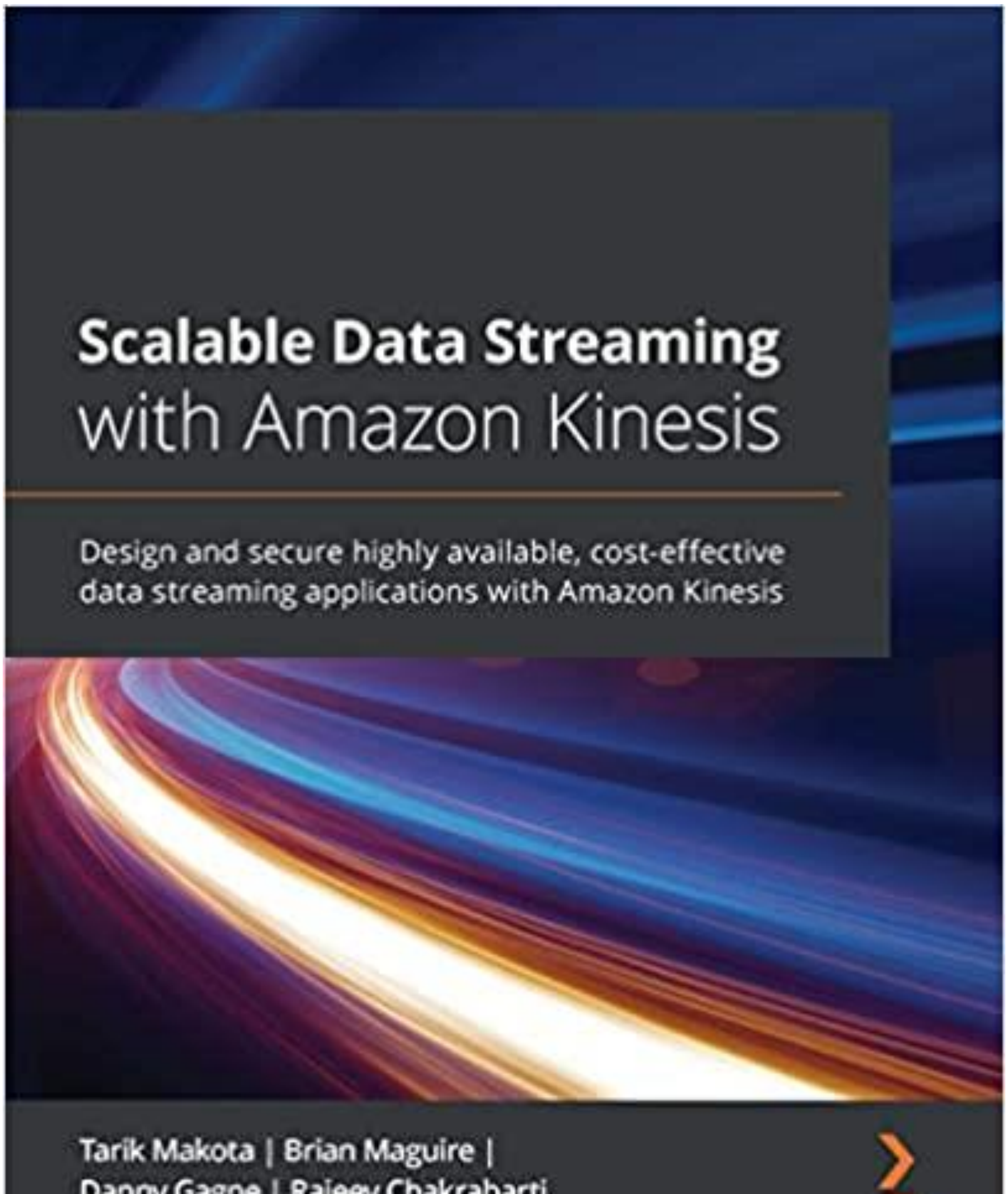
Streaming Data Solutions on AWS

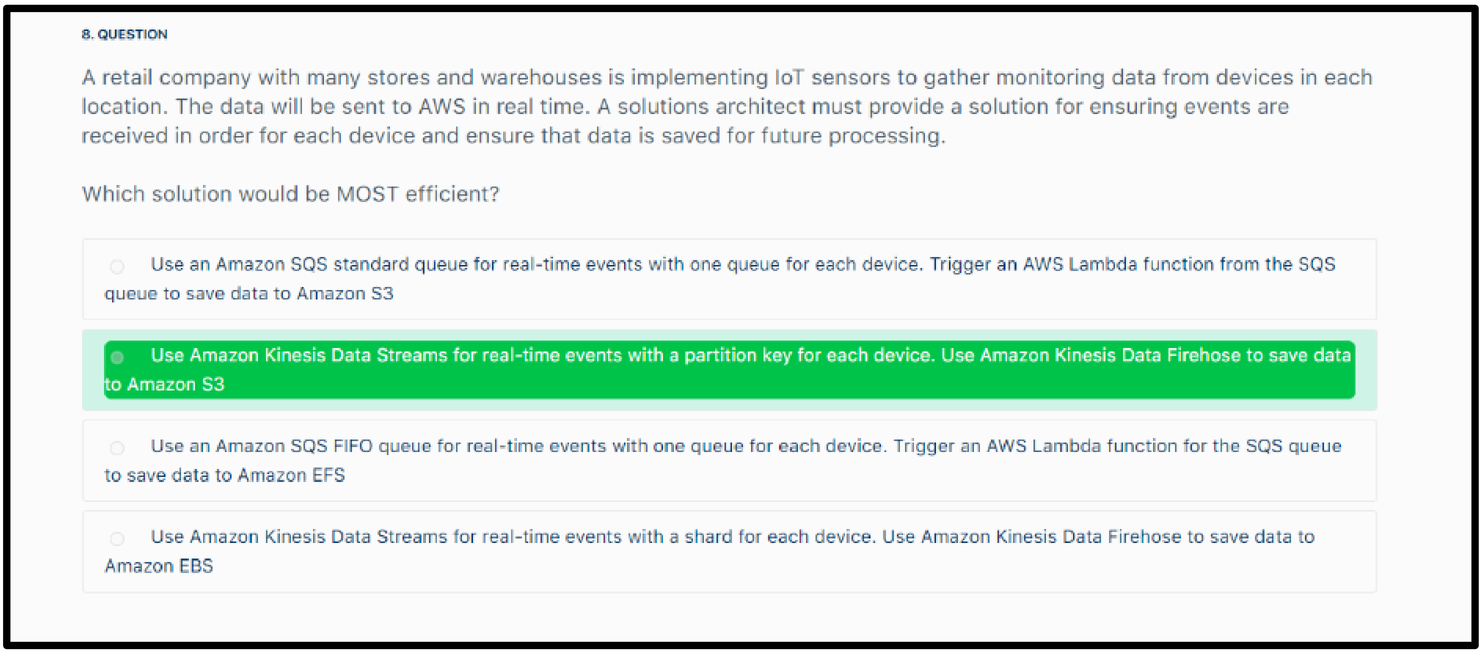
AWS Whitepaper

Contributors

- Amalia Rabinovitch, Sr. Solutions Architect, AWS
- Priyanka Chaudhary, Data Lake, Data Architect, AWS
- Zohair Nasimi, Solutions Architect, AWS
- Rob Kuhr, Solutions Architect, AWS
- Ejaz Sayyed, Sr. Partner Solutions Architect, AWS
- Allan MacInnis, Solutions Architect, AWS
- Chander Matrubhutam, Product Marketing Manager, AWS

How can I learn more about Amazon Kinesis?


Scalable Data Streaming
with Amazon Kinesis
Design and secure highly available, cost-effective data streaming applications with Amazon Kinesis
Tarik Makota | Brian Maguire |
Danny Gagne | Rajeev Chakrabarti



8. QUESTION

A retail company with many stores and warehouses is implementing IoT sensors to gather monitoring data from devices in each location. The data will be sent to AWS in real time. A solutions architect must provide a solution for ensuring events are received in order for each device and ensure that data is saved for future processing.

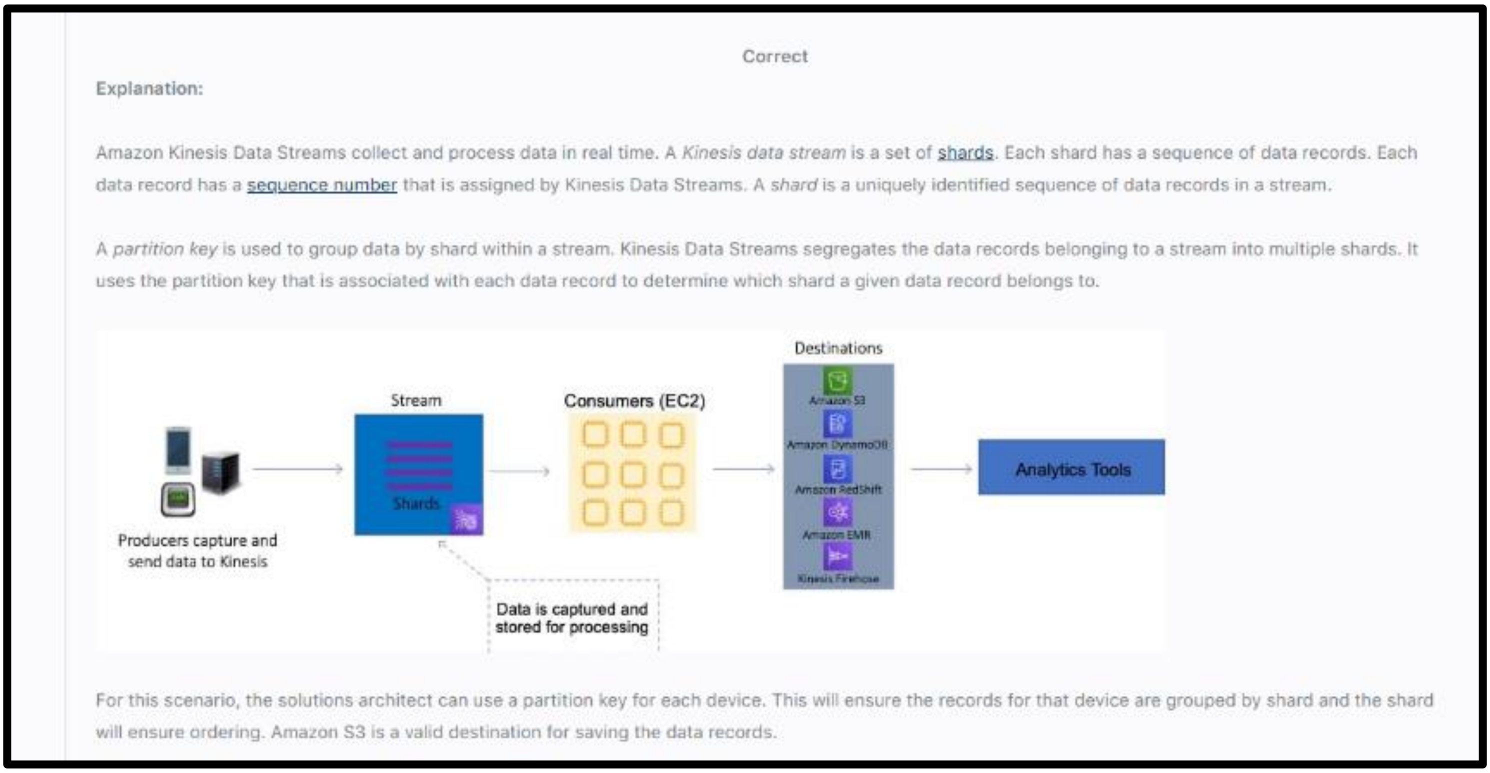
Which solution would be MOST efficient?

○ Use an Amazon SQS standard queue for real-time events with one queue for each device. Trigger an AWS Lambda function from the SQS queue to save data to Amazon S3

● Use Amazon Kinesis Data Streams for real-time events with a partition key for each device. Use Amazon Kinesis Data Firehose to save data to Amazon S3

○ Use an Amazon SQS FIFO queue for real-time events with one queue for each device. Trigger an AWS Lambda function for the SQS queue to save data to Amazon EFS

○ Use Amazon Kinesis Data Streams for real-time events with a shard for each device. Use Amazon Kinesis Data Firehose to save data to Amazon EBS



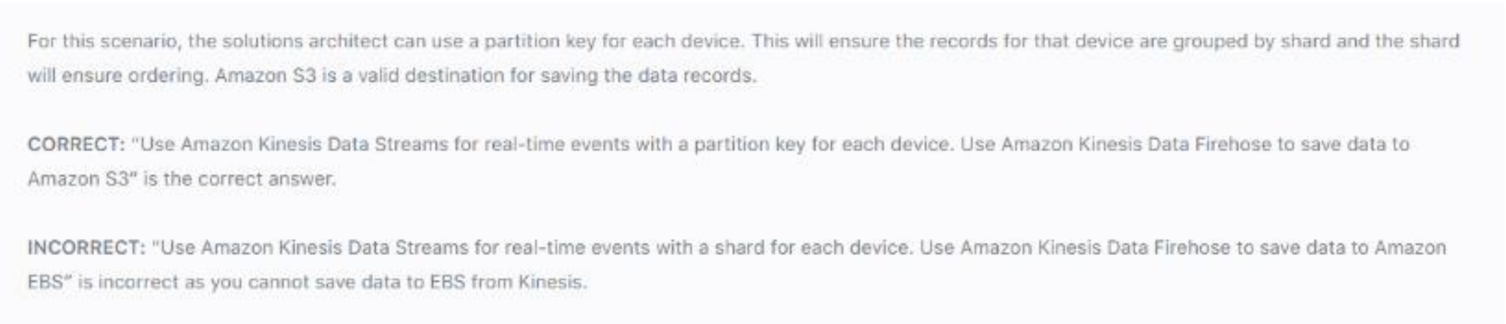
Correct

Explanation:

Amazon Kinesis Data Streams collect and process data in real time. A *Kinesis data stream* is a set of shards. Each shard has a sequence of data records. Each data record has a sequence number that is assigned by Kinesis Data Streams. A *shard* is a uniquely identified sequence of data records in a stream.

A *partition key* is used to group data by shard within a stream. Kinesis Data Streams segregates the data records belonging to a stream into multiple shards. It uses the partition key that is associated with each data record to determine which shard a given data record belongs to.

For this scenario, the solutions architect can use a partition key for each device. This will ensure the records for that device are grouped by shard and the shard will ensure ordering. Amazon S3 is a valid destination for saving the data records.



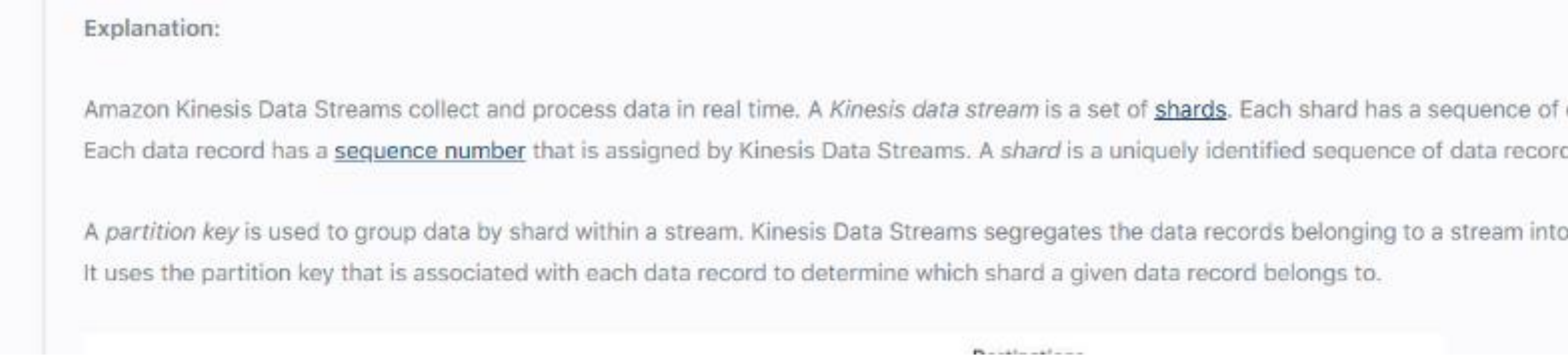
For this scenario, the solutions architect can use a partition key for each device. This will ensure the records for that device are grouped by shard and the shard will ensure ordering. Amazon S3 is a valid destination for saving the data records.

CORRECT: "Use Amazon Kinesis Data Streams for real-time events with a partition key for each device. Use Amazon Kinesis Data Firehose to save data to Amazon S3" is the correct answer.

INCORRECT: "Use Amazon Kinesis Data Streams for real-time events with a shard for each device. Use Amazon Kinesis Data Firehose to save data to Amazon EBS" is incorrect as you cannot save data to EBS from Kinesis.

"A **shard** is a uniquely identified *sequence of* data *records* in a stream."

- Neal Davis



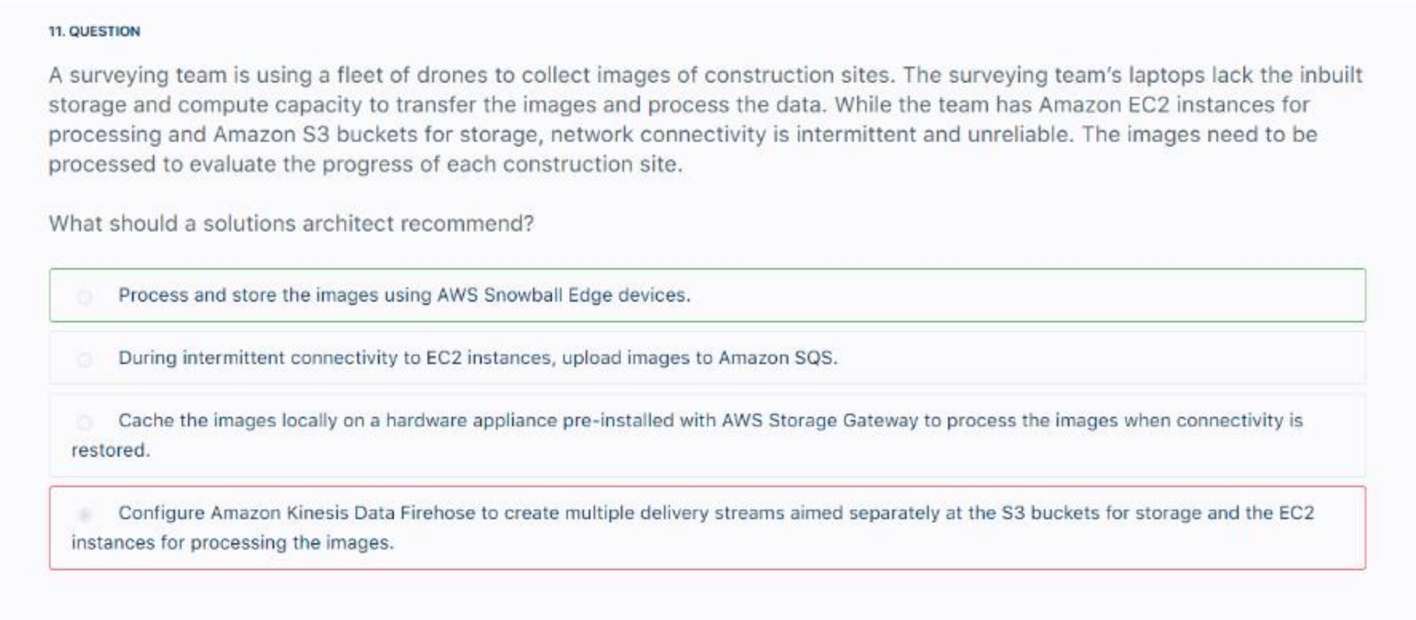
Explanation:

Amazon Kinesis Data Streams collect and process data in real time. A *Kinesis data stream* is a set of shards. Each shard has a sequence of
Each data record has a sequence number that is assigned by Kinesis Data Streams. A *shard* is a uniquely identified sequence of data recor

A *partition key* is used to group data by shard within a stream. Kinesis Data Streams segregates the data records belonging to a stream into
It uses the partition key that is associated with each data record to determine which shard a given data record belongs to.

> “A Kinesis data stream **is**
>
> a *Set of Shards*”
>
> - Neal Davis

KINESIS DATA FIREHOSE



11. QUESTION

A surveying team is using a fleet of drones to collect images of construction sites. The surveying team's laptops lack the inbuilt storage and compute capacity to transfer the images and process the data. While the team has Amazon EC2 instances for processing and Amazon S3 buckets for storage, network connectivity is intermittent and unreliable. The images need to be processed to evaluate the progress of each construction site.

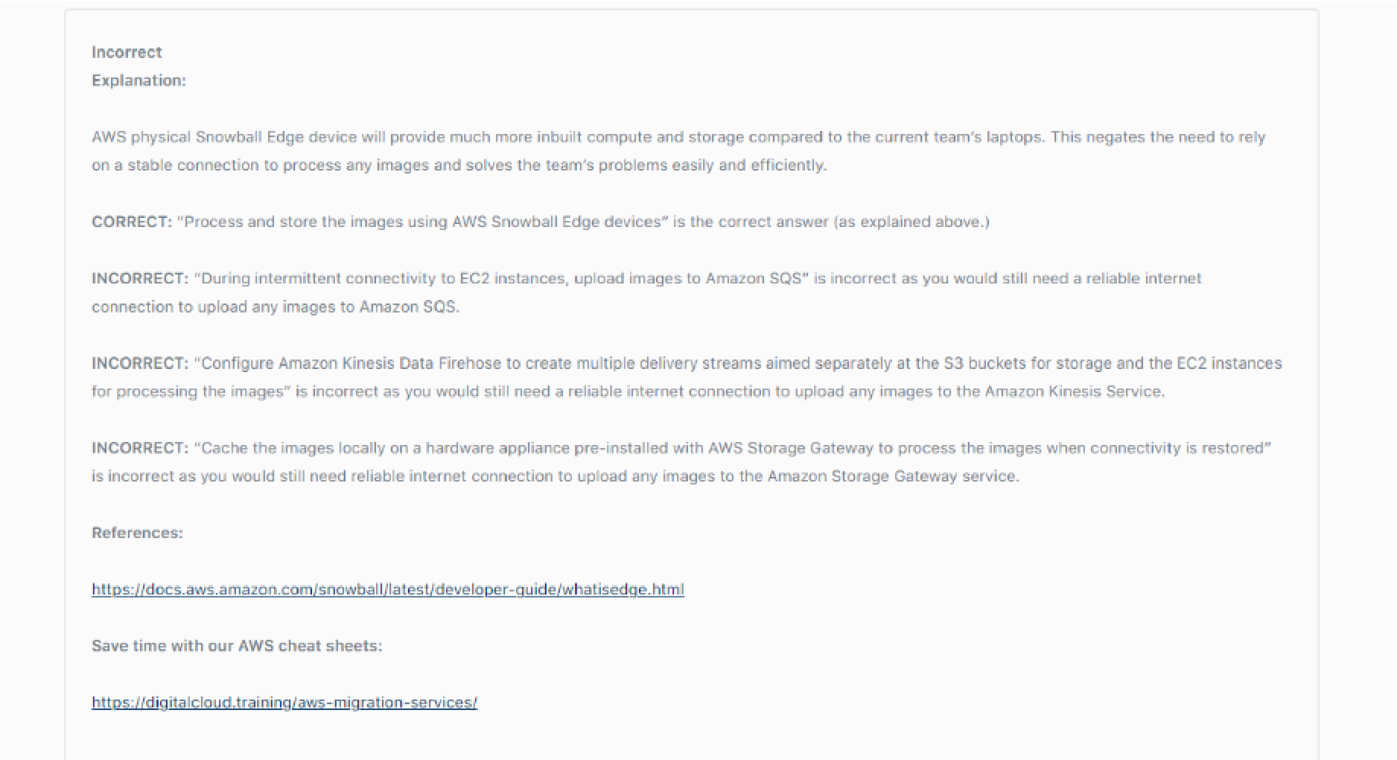
What should a solutions architect recommend?

○ Process and store the images using AWS Snowball Edge devices.

○ During intermittent connectivity to EC2 instances, upload images to Amazon SQS.

○ Cache the images locally on a hardware appliance pre-installed with AWS Storage Gateway to process the images when connectivity is restored.

● Configure Amazon Kinesis Data Firehose to create multiple delivery streams aimed separately at the S3 buckets for storage and the EC2 instances for processing the images.



Incorrect
Explanation:

AWS physical Snowball Edge device will provide much more inbuilt compute and storage compared to the current team's laptops. This negates the need to rely on a stable connection to process any images and solves the team's problems easily and efficiently.

CORRECT: "Process and store the images using AWS Snowball Edge devices" is the correct answer (as explained above.)

INCORRECT: "During intermittent connectivity to EC2 instances, upload images to Amazon SQS" is incorrect as you would still need a reliable internet connection to upload any images to Amazon SQS.

INCORRECT: "Configure Amazon Kinesis Data Firehose to create multiple delivery streams aimed separately at the S3 buckets for storage and the EC2 instances for processing the images" is incorrect as you would still need a reliable internet connection to upload any images to the Amazon Kinesis Service.

INCORRECT: "Cache the images locally on a hardware appliance pre-installed with AWS Storage Gateway to process the images when connectivity is restored" is incorrect as you would still need reliable internet connection to upload any images to the Amazon Storage Gateway service.

References:

https://docs.aws.amazon.com/snowball/latest/developer-guide/whatisedge.html

Save time with our AWS cheat sheets:

https://digitalcloud.training/aws-migration-services/

Amazon Kinesis Data Firehose adds support for data stream delivery to Amazon Redshift Serverless

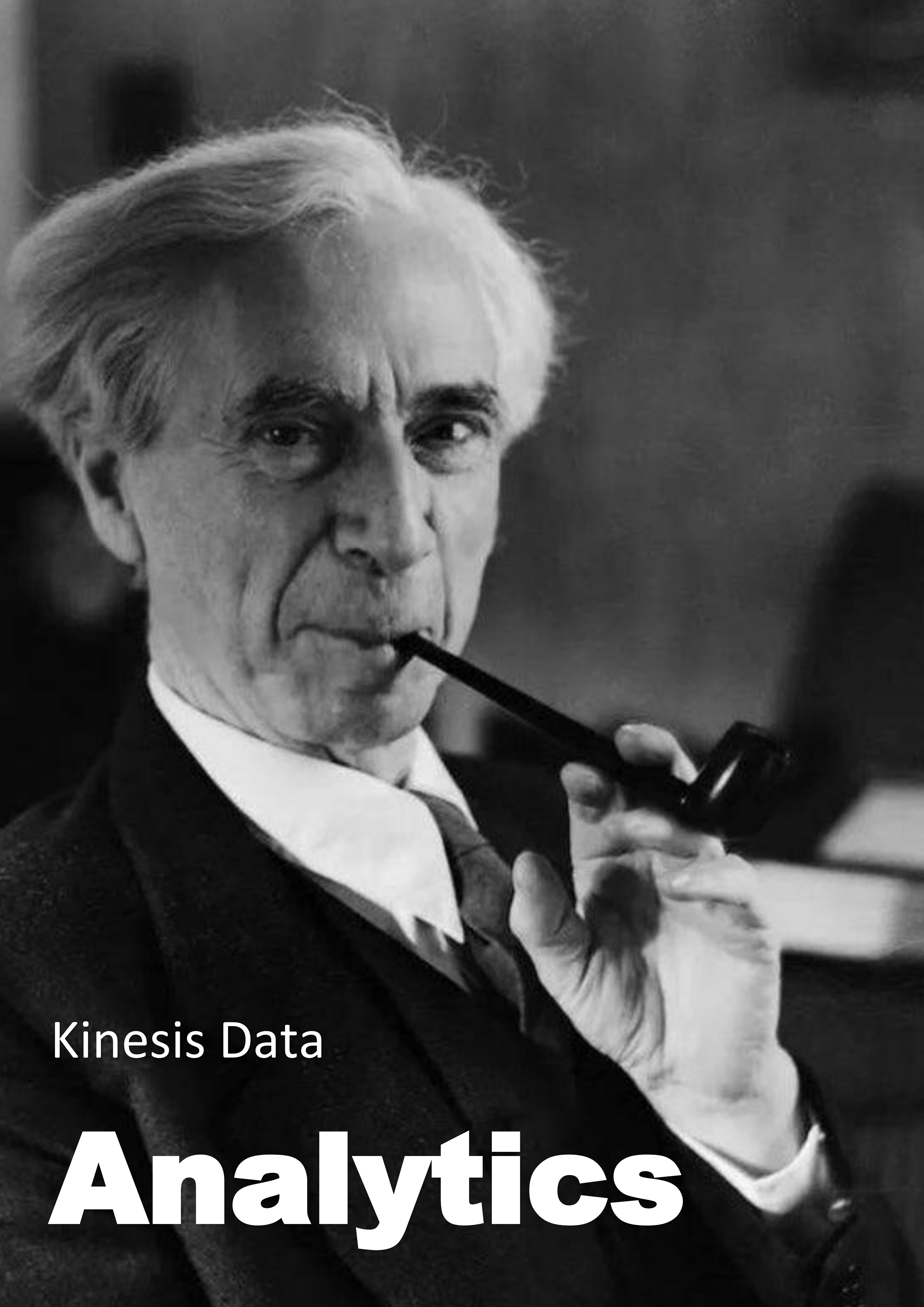
Posted On: Jun 19, 2023

Amazon Kinesis Data Firehose can now deliver streaming data to Amazon Redshift Serverless. With few clicks, you can more easily ingest, transform, and reliably deliver streaming data into Amazon Redshift Serverless without building and managing your own data ingestion and delivery infrastructure. Kinesis Data Firehose is a fully managed service that automatically scales to match the throughput of your data and without ongoing administration.

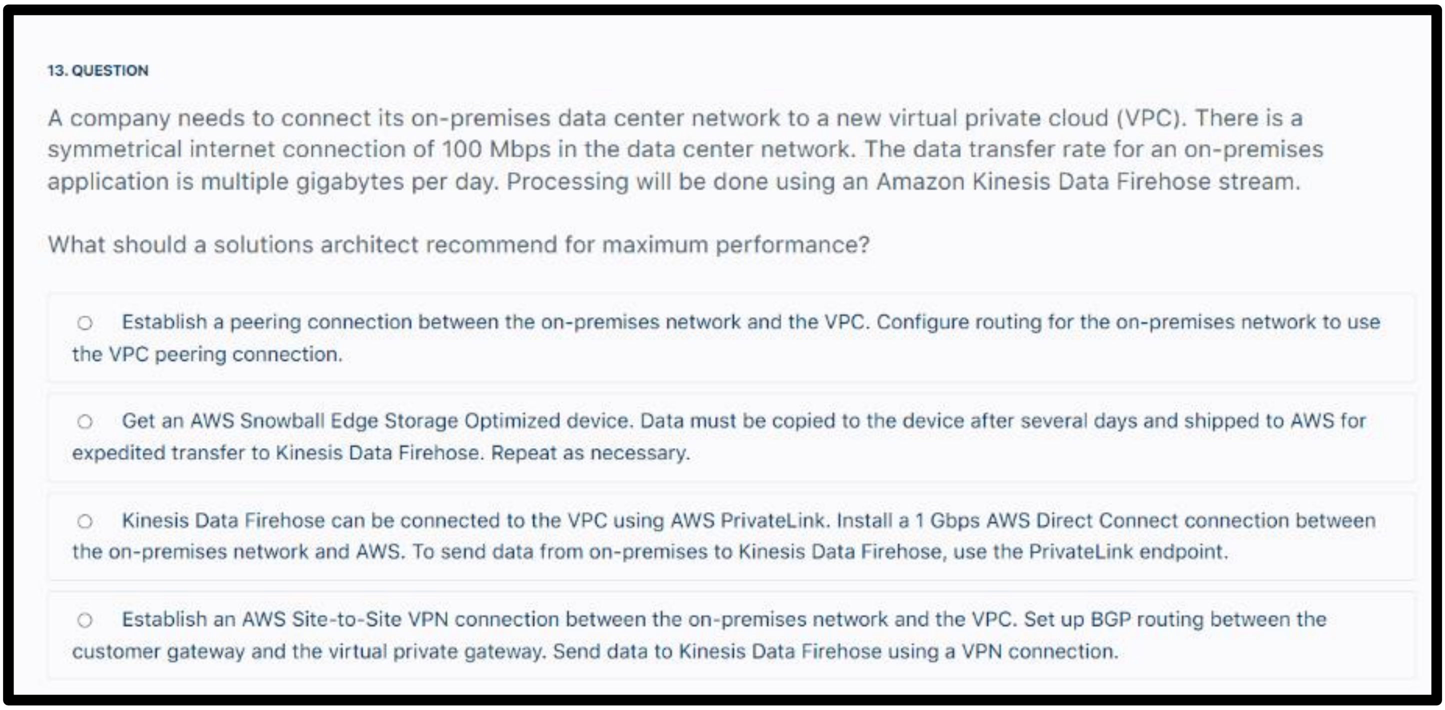
Amazon Redshift Serverless allows you to run and scale analytics without having to provision and manage data warehouse clusters. With Amazon Redshift Serverless, all users including data analysts, developers, and data scientists, can use Amazon Redshift to get insights from data in seconds. You only pay for the compute used for the duration of the workloads on a per-second basis. You can benefit from this simplicity without making any changes to your existing analytics and business intelligence applications. You can configure the Amazon Redshift Serverless instance to be publicly accessible and start using Amazon Kinesis Data Firehose with it to reliably load real-time streams.

Amazon Kinesis Data Firehose with Amazon Redshift Serverless is generally available in the regions here under Redshift Serverless API section.

To get started, you need an AWS account. Once you have an account, you can create a delivery stream in the Amazon Kinesis Console. To learn more, explore the Amazon Kinesis Data Firehose developer guide.



Kinesis Data Analytics



13. QUESTION

A company needs to connect its on-premises data center network to a new virtual private cloud (VPC). There is a symmetrical internet connection of 100 Mbps in the data center network. The data transfer rate for an on-premises application is multiple gigabytes per day. Processing will be done using an Amazon Kinesis Data Firehose stream.

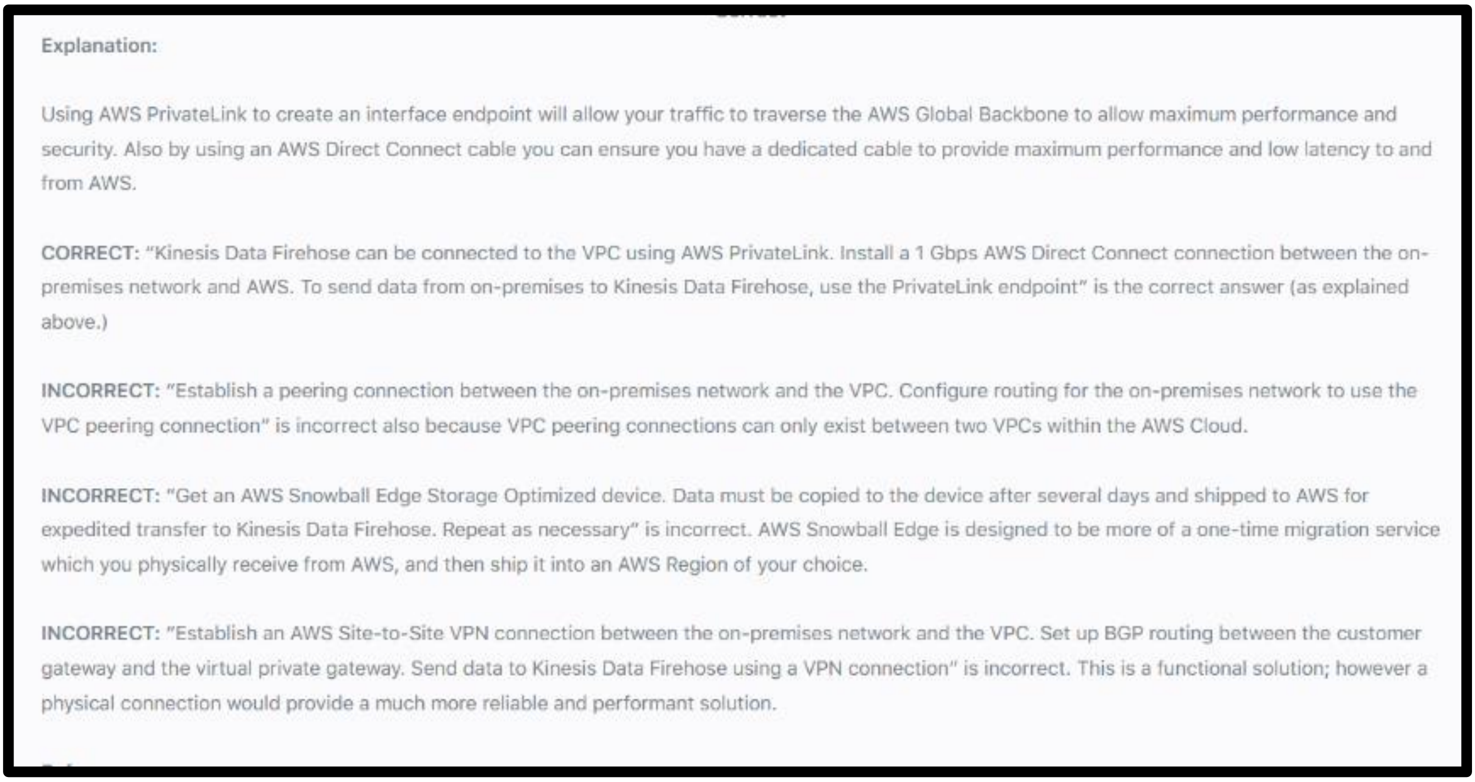
What should a solutions architect recommend for maximum performance?

○ Establish a peering connection between the on-premises network and the VPC. Configure routing for the on-premises network to use the VPC peering connection.

○ Get an AWS Snowball Edge Storage Optimized device. Data must be copied to the device after several days and shipped to AWS for expedited transfer to Kinesis Data Firehose. Repeat as necessary.

○ Kinesis Data Firehose can be connected to the VPC using AWS PrivateLink. Install a 1 Gbps AWS Direct Connect connection between the on-premises network and AWS. To send data from on-premises to Kinesis Data Firehose, use the PrivateLink endpoint.

○ Establish an AWS Site-to-Site VPN connection between the on-premises network and the VPC. Set up BGP routing between the customer gateway and the virtual private gateway. Send data to Kinesis Data Firehose using a VPN connection.



Explanation:

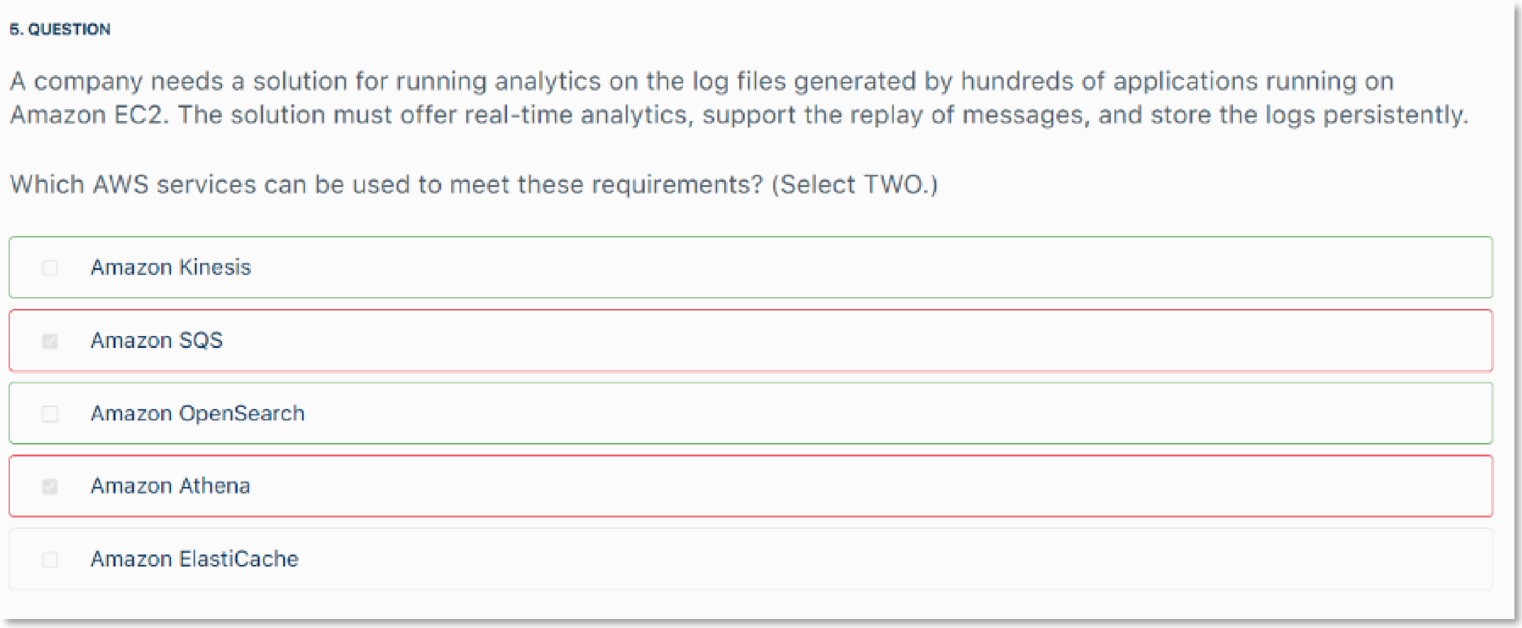
Using AWS PrivateLink to create an interface endpoint will allow your traffic to traverse the AWS Global Backbone to allow maximum performance and security. Also by using an AWS Direct Connect cable you can ensure you have a dedicated cable to provide maximum performance and low latency to and from AWS.

CORRECT: "Kinesis Data Firehose can be connected to the VPC using AWS PrivateLink. Install a 1 Gbps AWS Direct Connect connection between the on-premises network and AWS. To send data from on-premises to Kinesis Data Firehose, use the PrivateLink endpoint" is the correct answer (as explained above.)

INCORRECT: "Establish a peering connection between the on-premises network and the VPC. Configure routing for the on-premises network to use the VPC peering connection" is incorrect also because VPC peering connections can only exist between two VPCs within the AWS Cloud.

INCORRECT: "Get an AWS Snowball Edge Storage Optimized device. Data must be copied to the device after several days and shipped to AWS for expedited transfer to Kinesis Data Firehose. Repeat as necessary" is incorrect. AWS Snowball Edge is designed to be more of a one-time migration service which you physically receive from AWS, and then ship it into an AWS Region of your choice.

INCORRECT: "Establish an AWS Site-to-Site VPN connection between the on-premises network and the VPC. Set up BGP routing between the customer gateway and the virtual private gateway. Send data to Kinesis Data Firehose using a VPN connection" is incorrect. This is a functional solution; however a physical connection would provide a much more reliable and performant solution.



5. QUESTION

A company needs a solution for running analytics on the log files generated by hundreds of applications running on Amazon EC2. The solution must offer real-time analytics, support the replay of messages, and store the logs persistently.

Which AWS services can be used to meet these requirements? (Select TWO.)

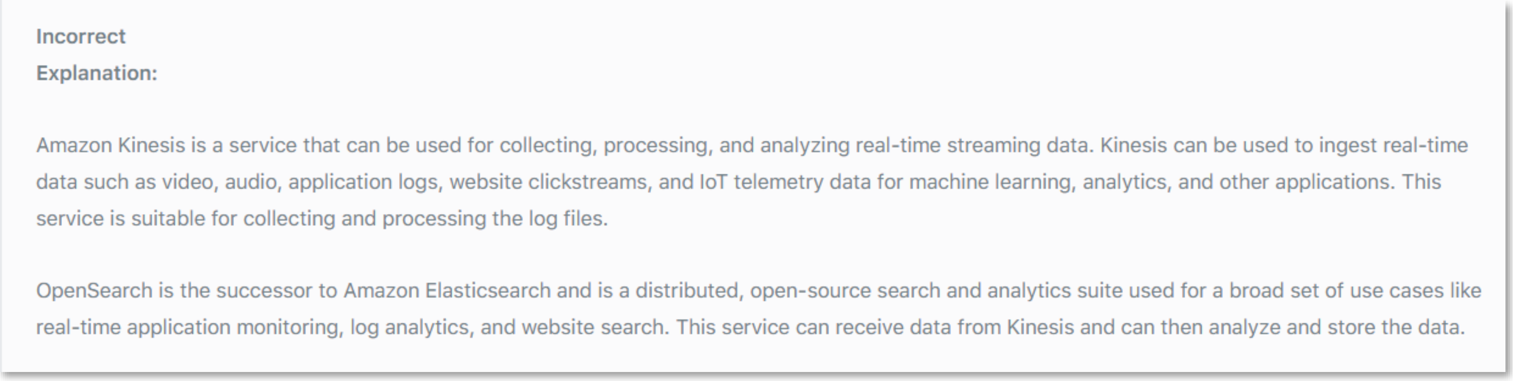
☐ Amazon Kinesis

☑ Amazon SQS

☐ Amazon OpenSearch

☑ Amazon Athena

☐ Amazon ElastiCache

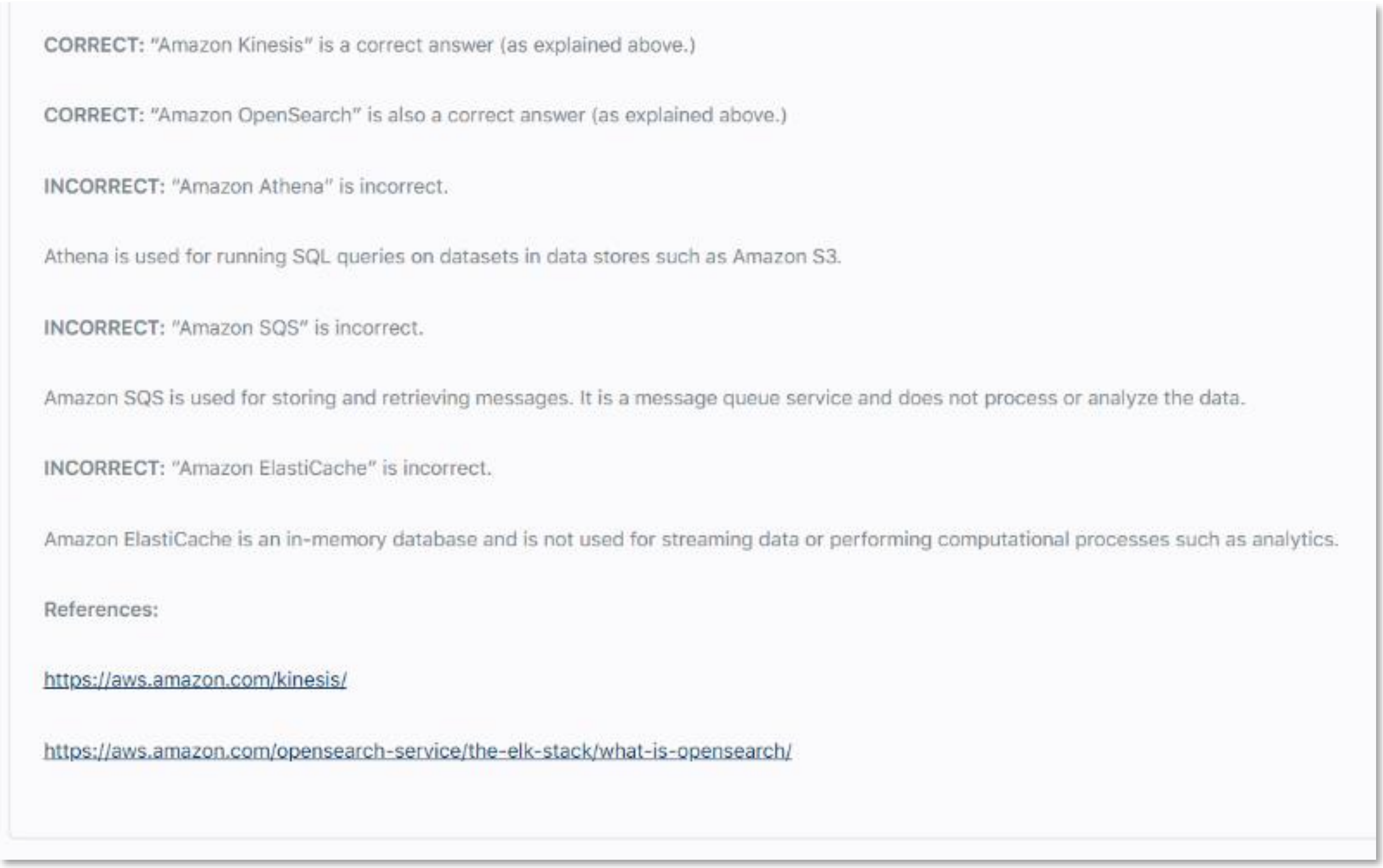


Incorrect

Explanation:

Amazon Kinesis is a service that can be used for collecting, processing, and analyzing real-time streaming data. Kinesis can be used to ingest real-time data such as video, audio, application logs, website clickstreams, and IoT telemetry data for machine learning, analytics, and other applications. This service is suitable for collecting and processing the log files.

OpenSearch is the successor to Amazon Elasticsearch and is a distributed, open-source search and analytics suite used for a broad set of use cases like real-time application monitoring, log analytics, and website search. This service can receive data from Kinesis and can then analyze and store the data.



CORRECT: "Amazon Kinesis" is a correct answer (as explained above.)

CORRECT: "Amazon OpenSearch" is also a correct answer (as explained above.)

INCORRECT: "Amazon Athena" is incorrect.

Athena is used for running SQL queries on datasets in data stores such as Amazon S3.

INCORRECT: "Amazon SQS" is incorrect.

Amazon SQS is used for storing and retrieving messages. It is a message queue service and does not process or analyze the data.

INCORRECT: "Amazon ElastiCache" is incorrect.

Amazon ElastiCache is an in-memory database and is not used for streaming data or performing computational processes such as analytics.

References:

https://aws.amazon.com/kinesis/

https://aws.amazon.com/opensearch-service/the-elk-stack/what-is-opensearch/

Bibliography

I. Official
II. Unofficial
III. Critical

IV. General

I. Official

[AWS 2024]

Netflix on AWS. Available at: <https://aws.amazon.com/solutions/case-studies/netflix-kinesis-data-streams/>

II. Unofficial

[Bhaduri 2023]

Bhaduri, Korak (2021). "Message Queue vs Streaming. Feb 19th 2023. *The Iron Blog*. Available at: < https://blog.iron.io/message-queue-vs-streaming/>.

[Name 2022]

Amazon Kinesis: the core of real time streaming data analysis on AWS. Available at: <https://www.sungardas.com/en-us/cto-labs-blog/amazon-kinesis-the-core-of-real-time-streaming-data-analysis-on-aws/>

[Wikipedia 2022]

Wikipedia. "Streaming Media". Available at: < https://en.wikipedia.org/wiki/Streaming_media>.

III. Critical

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher. Available at:

<URL here>.

[Surname1]

Smith, David (year). Title of Work Here. 1st Jan 2022. City: Publisher.
Available at:
<URL here>.

IV. General

[Showyang 2022]

Lecture 11 – Introduction to Data Streaming. *Distributed Systems.*
Available at:
<https://web.csie.ndhu.edu.tw/showyang/DistrSys2020s/11DataStreaming.pdf>