



# Elasticité, Haute disponibilité & Surveillance

Par Lahda Biassou Alphonsine



# Lahda Biassou Alphonsine

## Ingénieure cloud et formatrice





# Plan

- **Besoin architecturale**
- **Evolution de vos solutions de calculs**
- **Evolution de vos bases de données**
- **Conception d'un environnement hautement disponible**
- **Surveillance**





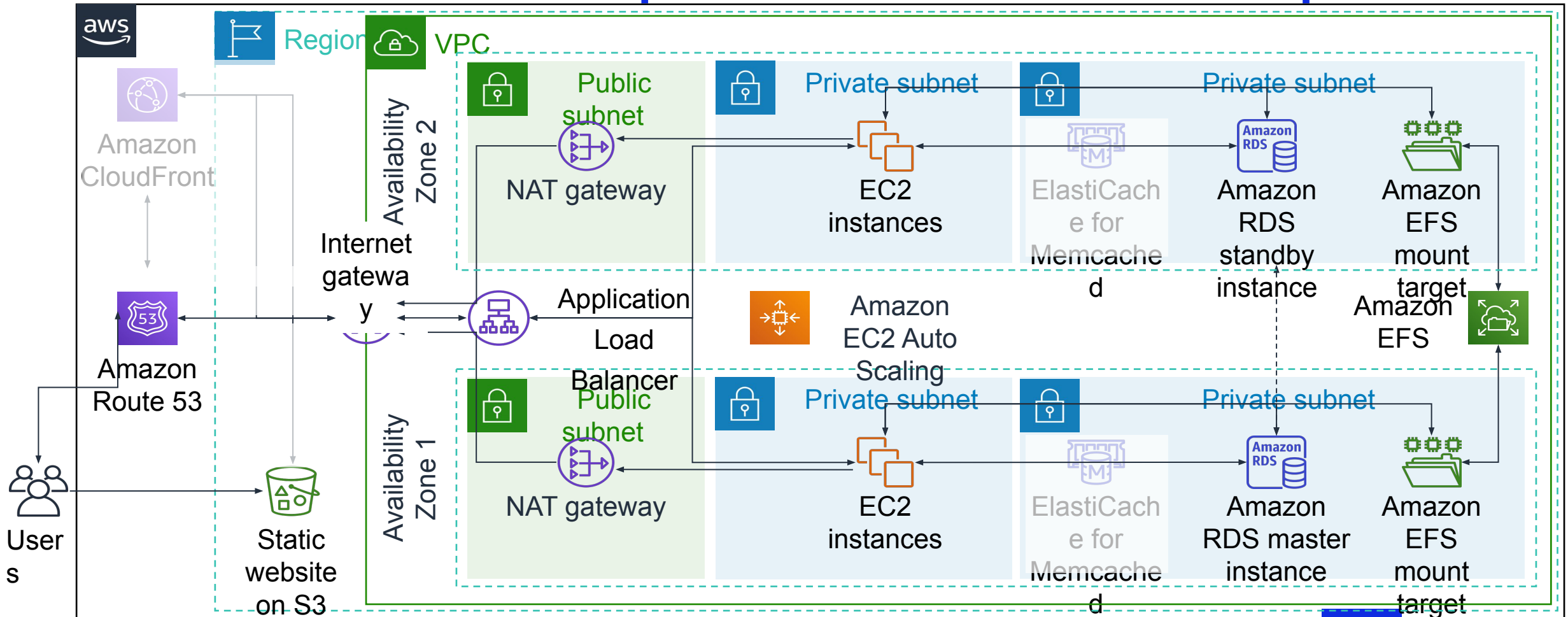
# Objectifs

A la fin de ce module, vous devriez être capable de :

- ➡ Utiliser Amazon EC2 Auto Scaling dans une architecture pour promouvoir l'élasticité
- ➡ Expliquer comment mettre à l'échelle les ressources de votre base de données
- ➡ Déployer un équilibreur de charge d'application pour créer un environnement hautement disponible
- ➡ Utiliser Amazon Route 53 pour le basculement du système de noms de domaine (DNS)
- ➡ Créer un environnement hautement disponible
- ➡ Concevoir des architectures qui utilisent Amazon CloudWatch pour surveiller les ressources et réagir en conséquence



# Besoin architecturale - implémentation de la haute disponibilité





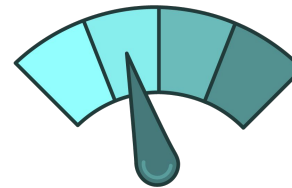
# besoin architectural - architecture réactive



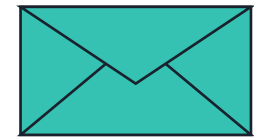
Élastique et  
évolutive



Résiliente



Responsive



Axe sur message



# Plan

- **Besoin architecturale**
- **Evolution de vos solutions de calculs**
- **Evolution de vos bases de données**
- **Conception d'un environnement hautement disponible**
- **Surveillance**





# C'est quoi l'élasticité?

Une infrastructure élastique peut s'étendre et se **contracter** en fonction de l'évolution des besoins en capacité.

## Exemples :

- Augmenter le nombre de serveurs web en cas de pic de trafic
- Réduire la capacité d'écriture de votre base de données en cas de baisse du trafic
- Gérer les fluctuations quotidiennes de la demande dans l'ensemble de votre architecture



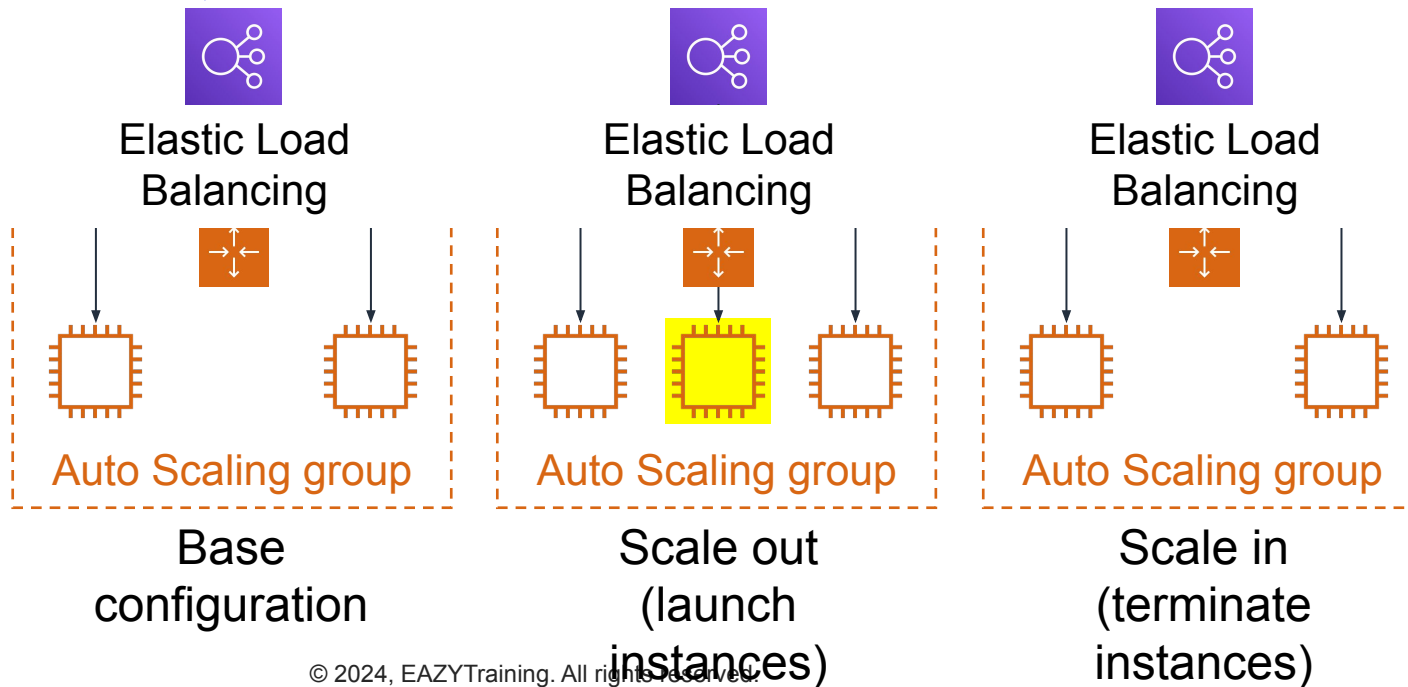


# C'est quoi l'élasticité?

Une technique utilisée pour obtenir de l'élasticité

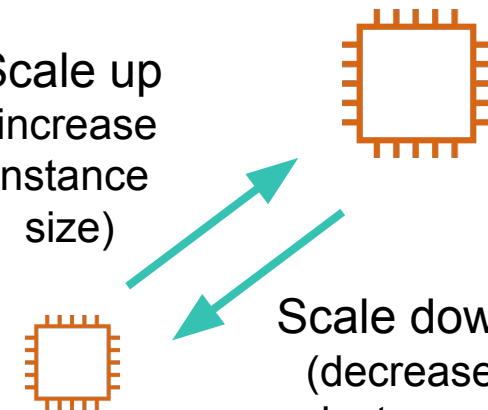
Mise à l'échelle **horizontale** ou **scaling horizontale**

Scaling **verticale**



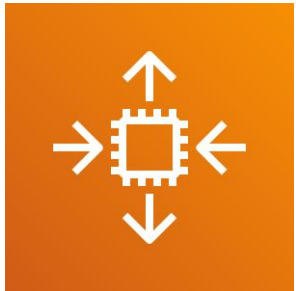
Scale up  
(increase  
instance  
size)

Scale down  
(decrease  
instance  
size)





## Evolution de vos ressources de calcul -Amazon EC2 Auto Scaling



Amazon EC2  
Auto Scaling

- **Lancement ou arrêt des instances** en fonction des conditions spécifiées
- **Enregistrement automatique des nouvelles instances** auprès des répartiteurs de charge lorsque cela est spécifié
- Peut être lancé à travers des **zones de disponibilité**



# Amazon EC2 Auto Scaling -options de scaling

## Scheduled

Good for predictable workloads



Scale based on date and time

**Use case:** Éteindre les instances de développement et de test de développement et de test

## Dynamic

Bon pour les conditions changeantes



Supports target tracking

**Use case:** Mise à l'échelle en fonction de l'utilisation de l'unité centrale

## Prédictive

Bon pour la prévision de la demande

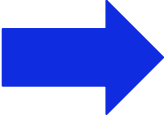


Scale based on machine learning

**Use case:** Gestion d'une augmentation de la charge de travail pour un site de commerce électronique lors d'un événement commercial majeur



# Dynamic scaling policy types

 **Simple scaling** - Ajustement unique de la mise à l'échelle

**Exemples de cas d'utilisation** : Nouvelles charges de travail, charges de travail en dents de scie

 **Mise à l'échelle progressive ou step scaling** - L'ajustement dépend de la taille de la violation de l'alarme

**Exemple de cas d'utilisation** : Charges de travail prévisibles

 **Mise à l'échelle du suivi de l'objectif ou Target tracking scaling** - Valeur cible pour une mesure spécifique

**Exemple de cas d'utilisation** : Applications horizontalement évolutives, telles que les applications à équilibrage de charge et les applications de traitement de données par lots.

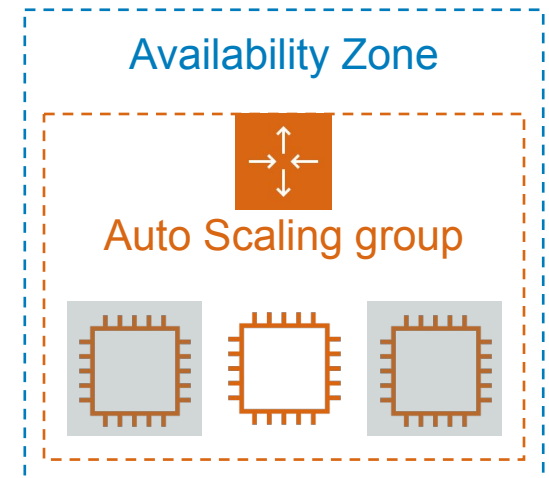


# Amazon Ec2 Auto Scaling -Auto scaling groups

Un groupe de mise à l'échelle automatique définit

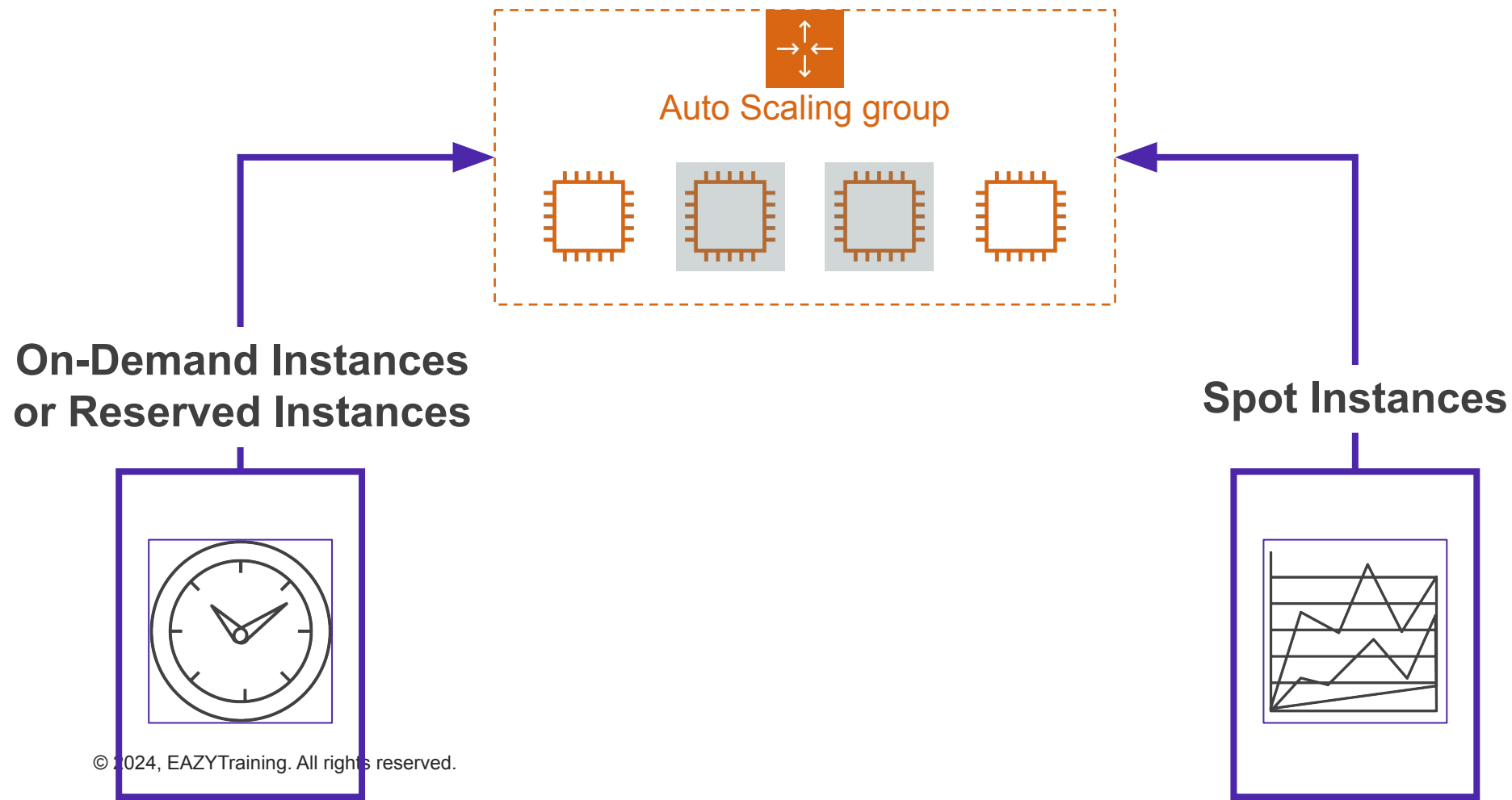
- la capacité minimale
- la capacité maximale
- Capacité souhaitée\*

Capacité?





# Amazon EC2 Auto Scaling -Purchasing options





# Considération de Scaling automatique

- Plusieurs types de mise à l'échelle automatique
- Mise à l'échelle simple, par étapes ou par suivi d'objectifs
- Mesures multiples (pas seulement l'unité centrale)
- Quand réduire ou augmenter l'échelle
- Utilisation de crochets de cycle de vie



# Points clés



- Une infrastructure élastique peut s'étendre et se contracter en fonction de l'évolution des besoins en capacité.
- Amazon EC2 Auto Scaling ajoute ou supprime automatiquement des instances EC2 en fonction des politiques que vous définissez, des planifications et des contrôles de santé.
- Amazon EC2 Auto Scaling propose plusieurs options de mise à l'échelle pour répondre au mieux aux besoins de vos applications.
- Lorsque vous configurez un groupe de mise à l'échelle automatique, vous pouvez spécifier les types d'instances EC2 et la combinaison de modèles de tarification qu'il utilise.





# Plan

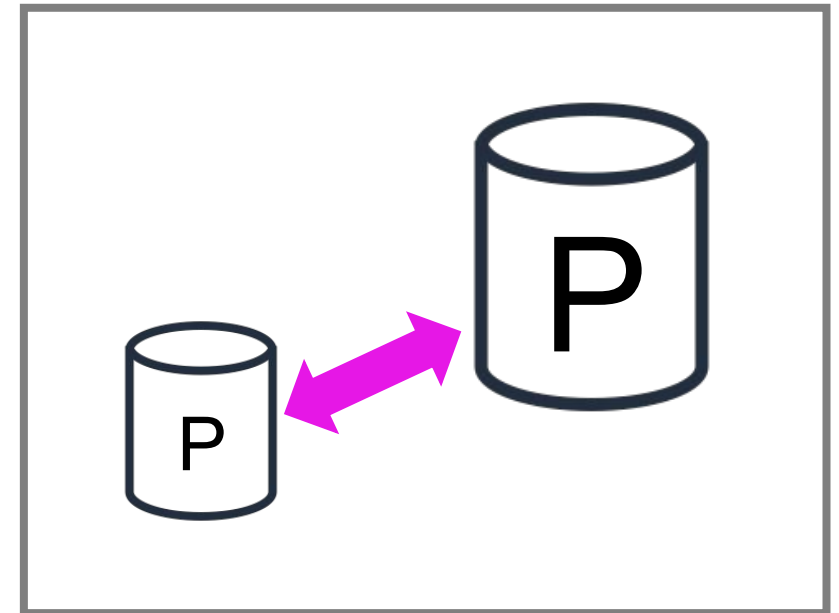
- **Besoin architecturale**
- **Evolution de vos solutions de calculs**
- **Evolution de vos bases de données**
- **Conception d'un environnement hautement disponible**
- **Surveillance**





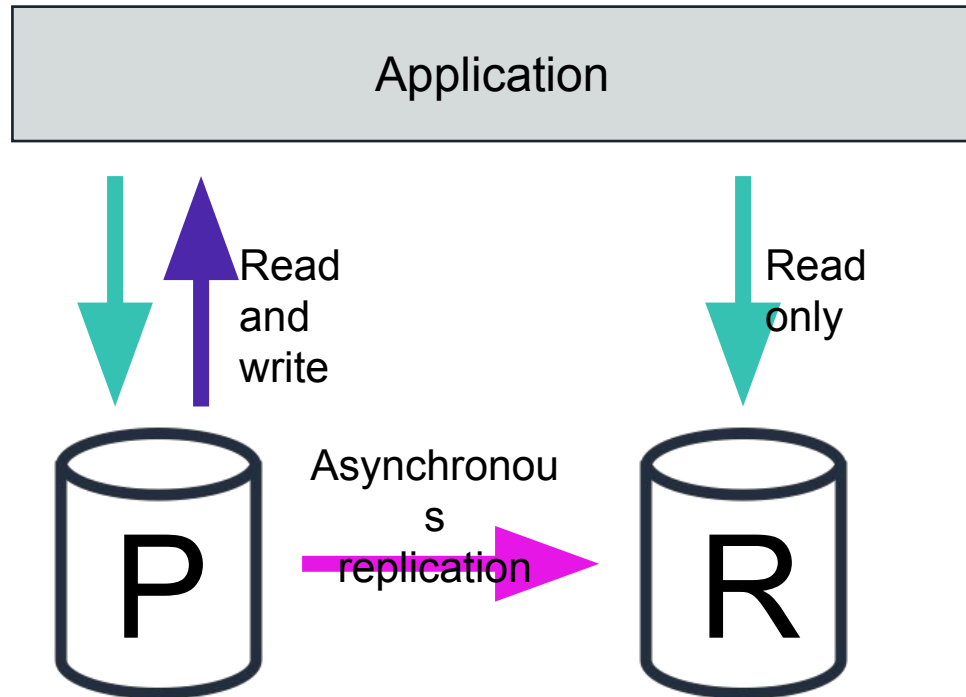
# Vertical scaling avec Amazon RDS -Push button scaling

- Augmentation ou diminution de la taille des instances de la base de données
- De **micro à 24xlarge** et tout ce qui se trouve entre les deux
- Évolution verticale avec **un temps d'arrêt minimal**





# Horizontal scaling avec Amazon RDS -Read replicas



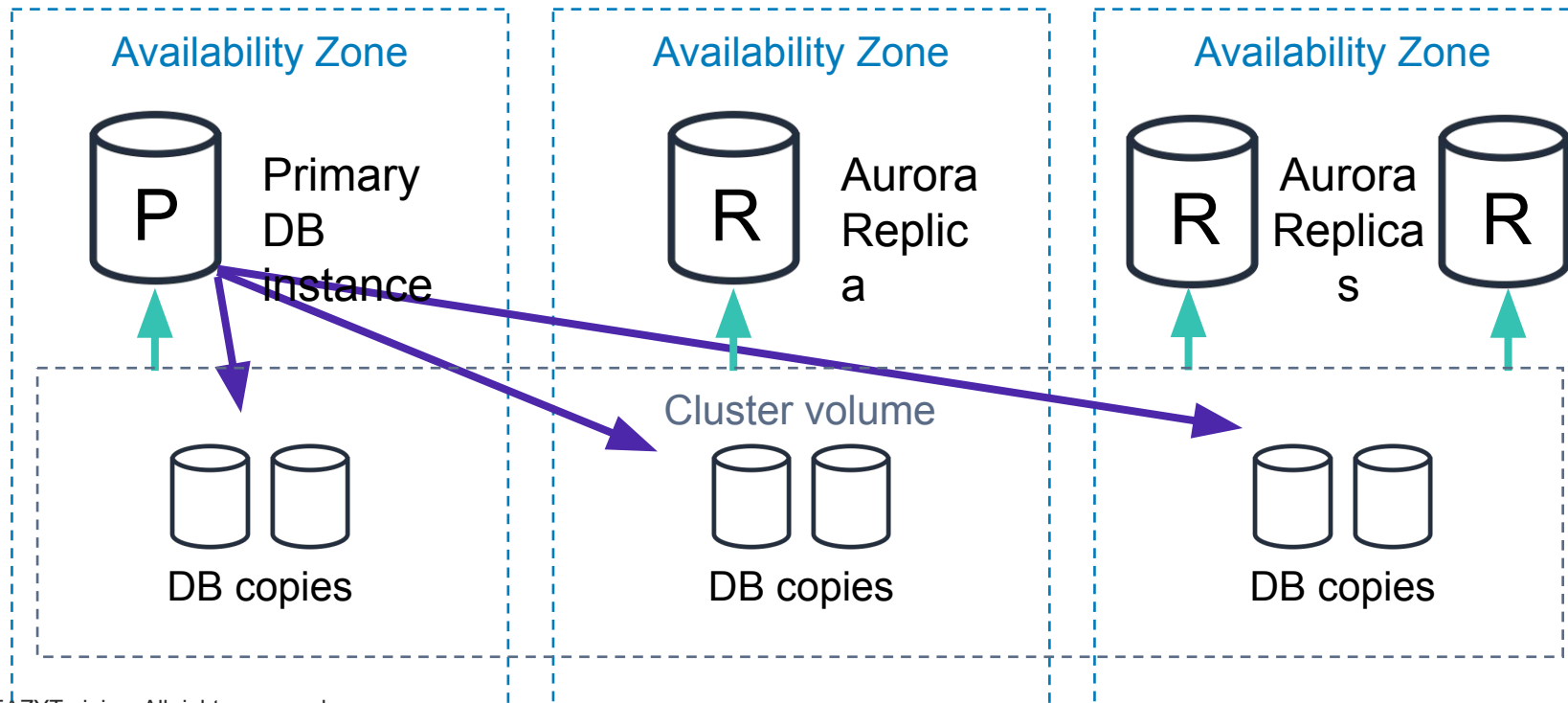
P = Primary (source) DB instance; R = Read Replica

- Évolution horizontale pour les **charges de travail lourdes en lecture**
- Jusqu'à **cinq réplicas** de lecture et **jusqu'à 15 réplicas Aurora**
- Réplication asynchrone
- Disponible pour Amazon RDS pour MySQL, MariaDB, PostgreSQL et Oracle



# Scaling avec Amazon Aurora

Chaque cluster Aurora DB peut avoir jusqu'à 15 répliques Aurora.





## Amazon Aurora Serverless

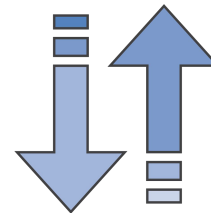


Répond automatiquement à  
votre demande :

- Scales les capacités
- Démarre
- Arrête



Payer pour le nombre  
d'unités de capacité  
unités de capacité Aurora  
(ACU) utilisées



Bon pour des situations  
intermittentes  
et charges de travail  
imprévisibles

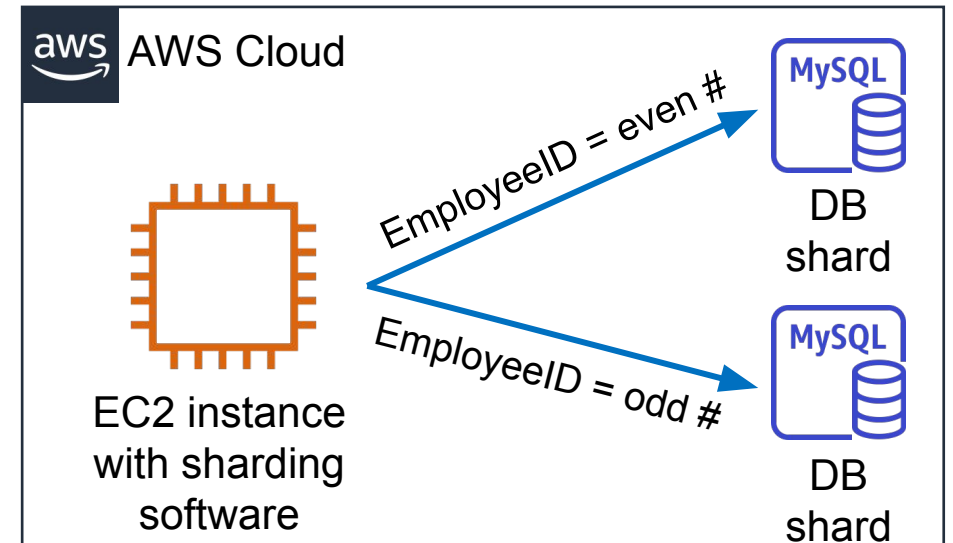


# Horizontal scaling -partage de bases de données ou database sharding

En l'absence de schémas, toutes les données résident dans une seule partition.

- Exemple : Identifiants des employés dans une seule base de données  
Avec le sharding, les données sont divisées en grands gros morceaux (shards).
- Exemple : Les numéros d'identification des employés pairs dans une base de données et les numéros d'identification des employés impairs dans une autre base de données.

Dans de nombreuses circonstances, la répartition améliore les performances d'écriture.





# Scaling avec Amazon DynamoDB -On-demand

## On-Demand

---

Pay per request



No more  
provisioning

**Use case:** Charges de travail irrégulières et imprévisibles.  
S'adapte rapidement aux besoins.



# Scaling avec Amazon DynamoDB -Auto scaling

## On-Demand

Pay per request

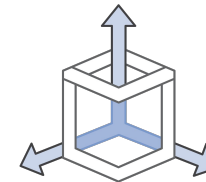


No more  
provisioning

**Use case:** Spiky, unpredictable workloads.  
Rapidly accommodates to need.

## Auto scaling

Default for all new tables



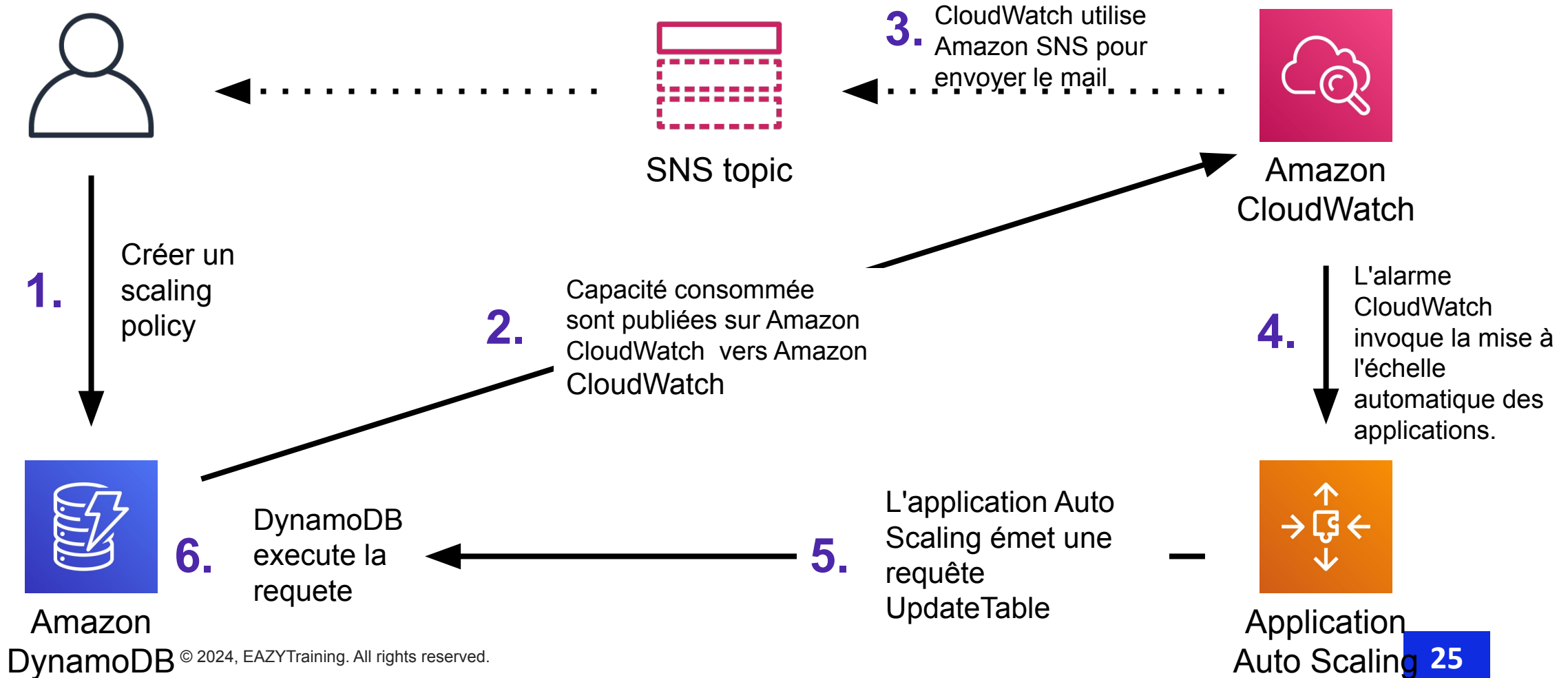
Specify upper and  
lower bounds

**Use case:** General scaling, good  
solution for most applications.





# Implementation d'auto scaling avec DynamoDB





# Augmentation de la capacité de débit - Capacité d'adaptation de DynamoDB

- Permet de lire et d'écrire sur des partitions chaudes **sans limitation de débit**
- **Augmente automatiquement la capacité de débit** pour les partitions qui reçoivent plus de trafic\*.
- Est **activé automatiquement pour chaque table DynamoDB**

\*Le trafic ne peut pas dépasser la capacité totale provisionnée de la table ou la capacité maximale de la partition.



Augmentation de la capacité d'adaptation du débit

Hot partition



## Exemple 1/3-Capacité d'adaptation

Exemple de tableau avec **capacité d'adaptation**

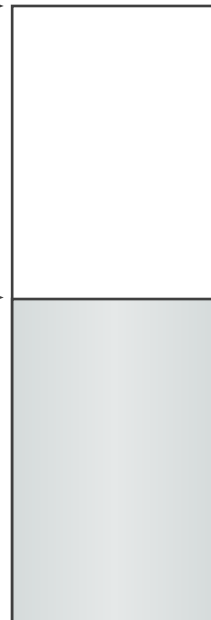
Capacité totale fournie = 400 UCE

Capacité totale consommée = 200 WCU

Provisioned: 100 WCU



Consumed: 50  
WCU



Partition 1



Partition 2



Partition 3



Partition 4



## Exemple 2/3-Capacité d'adaptation

Exemple de tableau avec **capacité d'adaptation**

Capacité totale fournie = 400 UCE

Capacité totale consommée = 250 WCU

Provisioned: 100 WCU



Consumed: 50 WCU



Partition 1



Partition 2



Partition 3



Partition 4



## Exemple 3/3-Capacité d'adaptation

Exemple de tableau avec **capacité d'adaptation**

Capacité totale fournie = 400 UCE

Capacité totale consommée = 300 WCU

Provisioned: 100 WCU

Consumed: 50 WCU

Consumed: 150 WCU

Augmentation de  
la capacité  
d'adaptation du  
débit

Partition 1

Partition 2

Partition 3

Partition 4



# La capacité d'adaptation ne corrige pas les hot keys et les hot partitions

Partition key value	Uniformité
ID de l'utilisateur, lorsque l'application a plusieurs utilisateurs	Bien
Code d'état, lorsqu'il n'y a que quelques codes d'état possibles	Mauvais
Date de création de l'élément, arrondie à la période la plus proche (par exemple, jour, heure ou minute)	Mauvais
ID de l'appareil, où chaque appareil accède aux données à des intervalles relativement similaires	bien
L'identification des appareils, où même si de nombreux appareils sont suivis, l'un d'entre eux est beaucoup plus populaire que tous les autres.	Mauvais



- Vous pouvez utiliser la mise à l'échelle par bouton-poussoir pour augmenter verticalement la capacité de calcul de votre instance RDS DB.
- Vous pouvez utiliser des répliques de lecture ou des shards pour faire évoluer horizontalement votre instance de base de données RDS.
- Avec Amazon Aurora, vous pouvez choisir la taille de la classe de l'instance de base de données et le nombre de répliques Aurora (jusqu'à 15).
- Aurora Serverless met automatiquement à l'échelle les ressources en fonction des spécifications de capacité minimale et maximale.
- Amazon DynamoDB On-Demand propose un modèle de tarification à la demande.
- La mise à l'échelle automatique de DynamoDB utilise Amazon Application Auto Scaling pour ajuster dynamiquement la capacité de débit provisionnée.
- DynamoDB adaptive capacity fonctionne en augmentant automatiquement la capacité de débit pour les partitions qui reçoivent plus de trafic.

## Points clés





# Plan

- **Besoin architecturale**
- **Evolution de vos solutions de calculs**
- **Evolution de vos bases de données**
- **Conception d'un environnement hautement disponible**
- **Surveillance**







## Systèmes Hautement Disponible

- Peuvent supporter une certaine dégradation tout en restant disponibles
- Avoir des temps d'arrêt réduits au minimum
- Nécessiter une intervention humaine minimale
- Reprendre après une panne ou passer à une source secondaire dans un délai acceptable de dégradation des performances.

Percentage of Uptime	Maximum Downtime Per Year	Equivalent Downtime Per Day
90%	36.5 days	2.4 hours
99%	3.65 days	14 minutes
99.9%	8.76 hours	86 seconds
99.99%	52.6 minutes	8.6 seconds
99.999%	5.25 minutes	0.86 seconds



## HA- Elastic Load Balancing



Elastic Load  
Balancing

Un **service d'équilibrage de charge géré** qui distribue le trafic applicatif entrant sur plusieurs instances EC2, conteneurs, adresses IP et fonctions Lambda.

- Peut être orienté vers **l'extérieur ou vers l'intérieur**
- Chaque équilibreur de charge reçoit **un nom DNS**
- Reconnaît les instances en **mauvaise santé et y répond.**



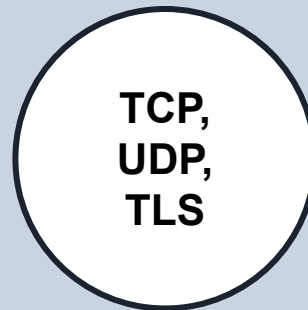
# Types de Load Balancers

## Application Load Balancer



- Gestion flexible des applications
- Équilibrage de charge avancé du trafic HTTP et HTTPS
- Fonctionne au niveau des requêtes(couche 7)

## Network Load Balancer



- Très haute performance et adresse IP statique pour votre application
- Répartition de la charge du trafic TCP, UDP et TLS
- Fonctionne au niveau de la connexion (couche 4)

## Gateway Load Balancer

PREVIOUS GENERATION  
for HTTP, HTTPS, TCP, and SSL

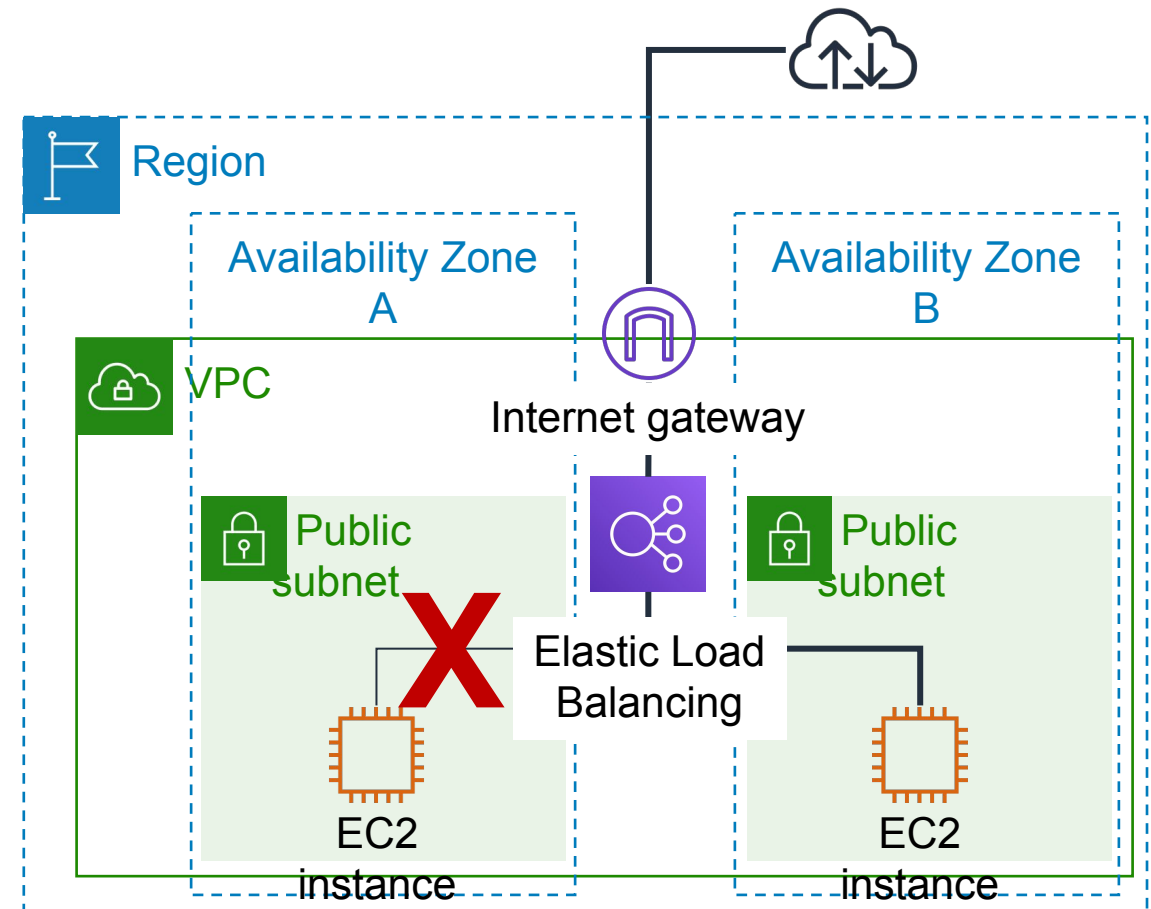
- Simplifie le déploiement, la mise à l'échelle et la gestion des appliances virtuelles de réseau tierces
- Operates at both the request level and connection level



# Implémentation de la haute disponibilité

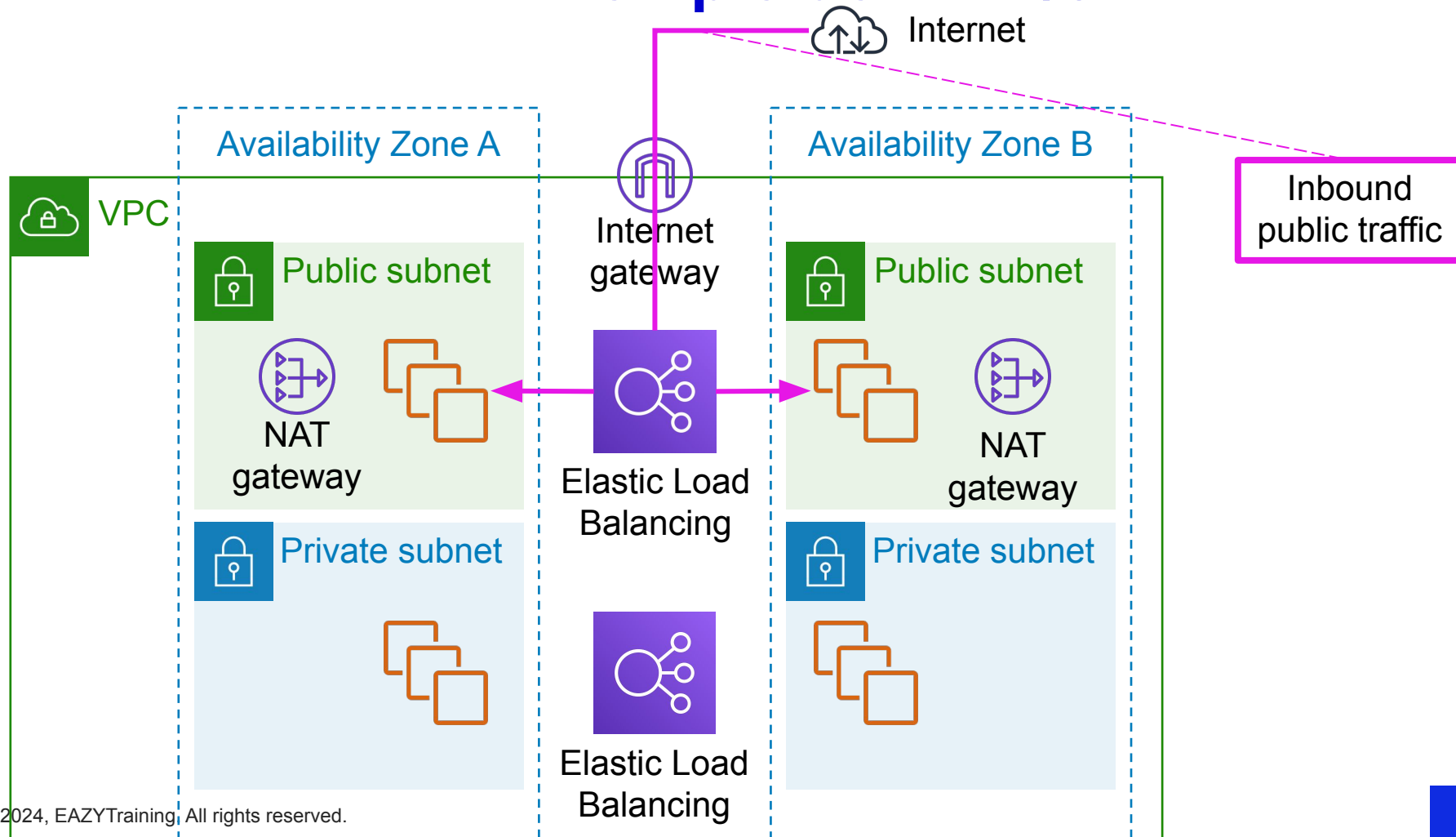
Commencez par deux zones de disponibilité par région AWS.

Si les ressources d'une zone de disponibilité sont inaccessibles, votre application ne devrait pas tomber en panne.



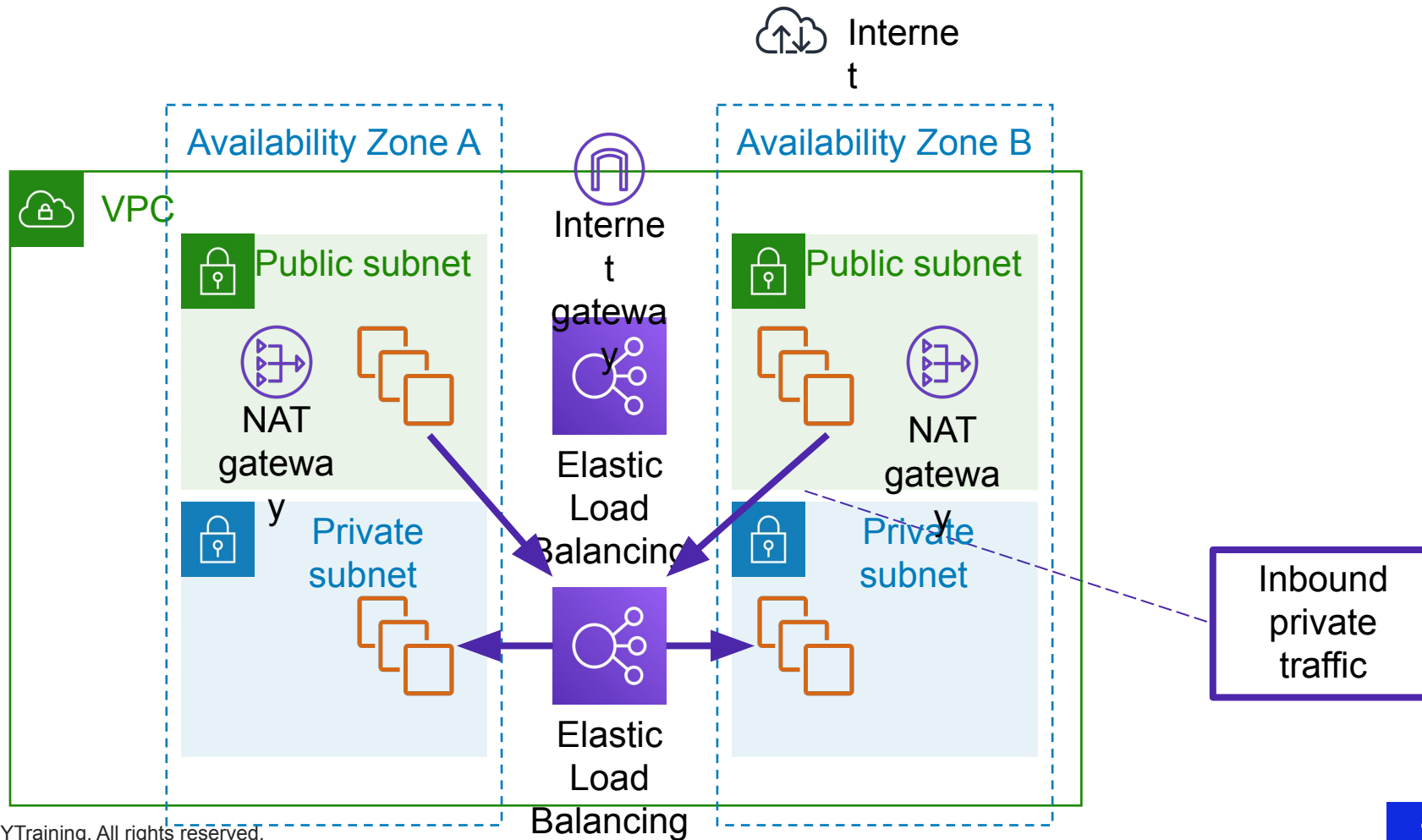


## Exemple de HA 1/3



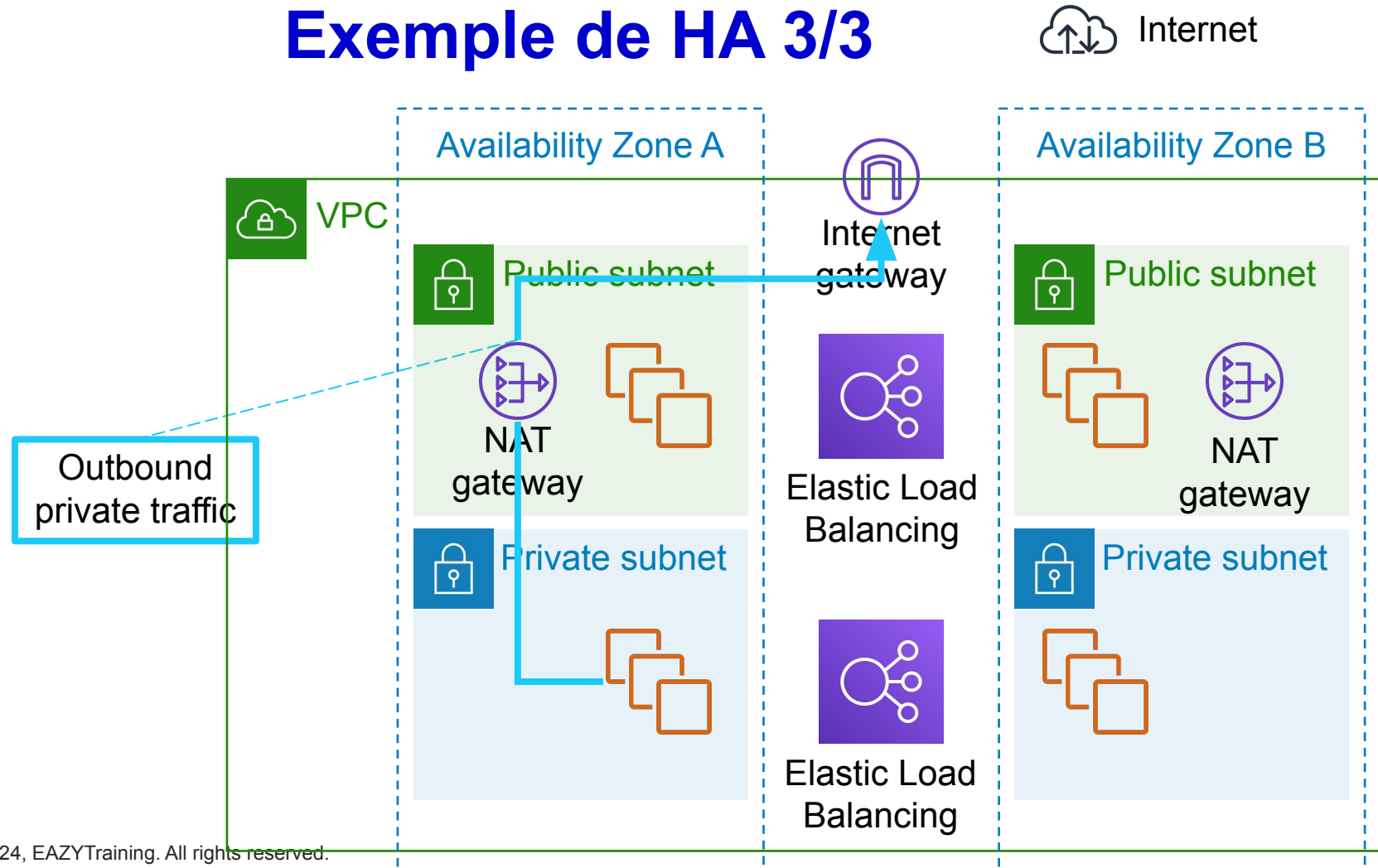


## Exemple HA 2/3



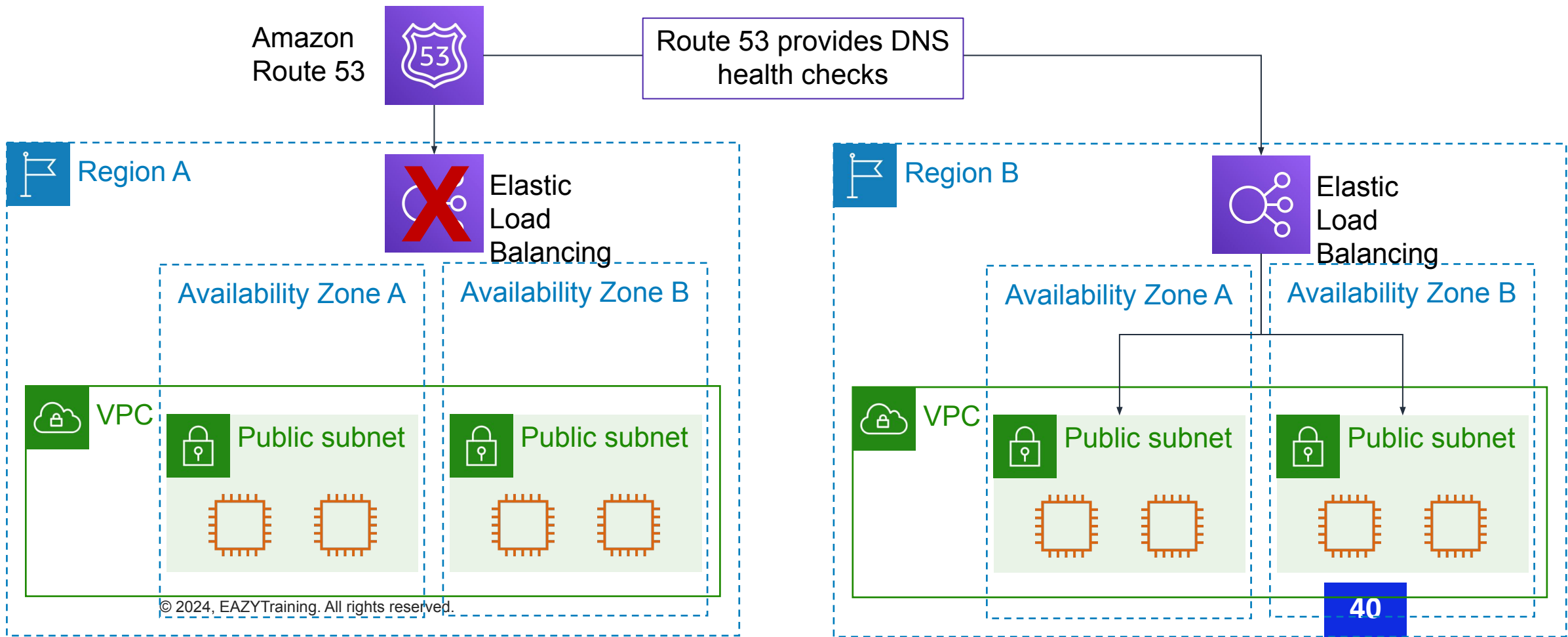


## Exemple de HA 3/3





# Haute disponibilité et DNS







# Plan

- **Besoin architecturale**
- **Evolution de vos solutions de calculs**
- **Evolution de vos bases de données**
- **Conception d'un environnement hautement disponible**
- **Surveillance**

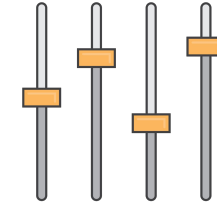




# Surveillance -Contrôle de l'utilisation, des opérations et des performances



Operational Health



Utilisation des ressources



Performance de l'application



Audit de securite



# Surveillance des coûts

Pour créer une architecture plus souple et plus élastique, vous devez savoir où vous dépensez de l'argent.

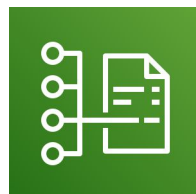
AWS Cost Explorer



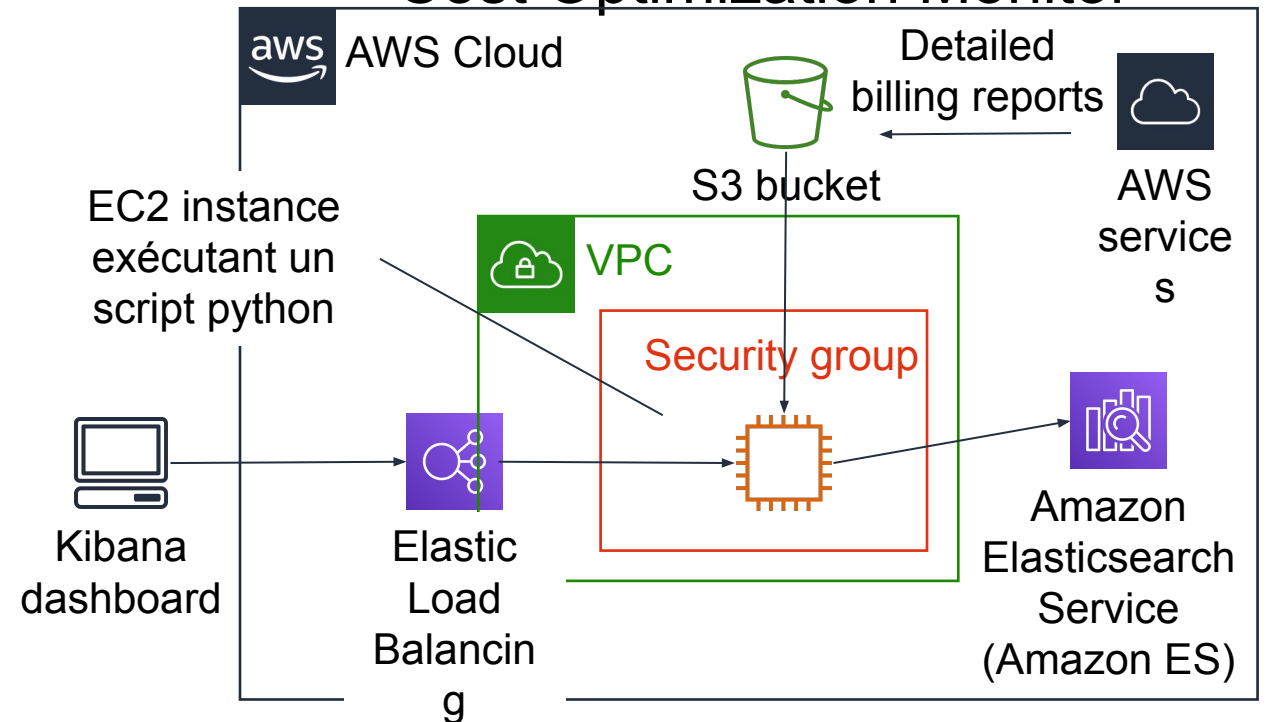
AWS Budgets



AWS Cost and Usage Report



## Cost Optimization Monitor





## Amazon CloudWatch



Amazon  
CloudWatch

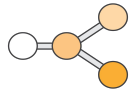
- Collecte et suit les métriques de vos ressources et applications
- Permet de corrélérer, de visualiser et d'analyser les métriques et les journaux.
- Permet de créer des alarmes et de détecter les comportements anormaux
- Permet d'envoyer des notifications ou d'apporter des modifications aux ressources que vous surveillez.



# Amazon CloudWatch



Metrics



Logs



Alarms



Events



Rules



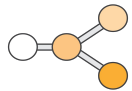
Targets



# Amazon CloudWatch -metrics



Metrics



Logs



Alarms



Events



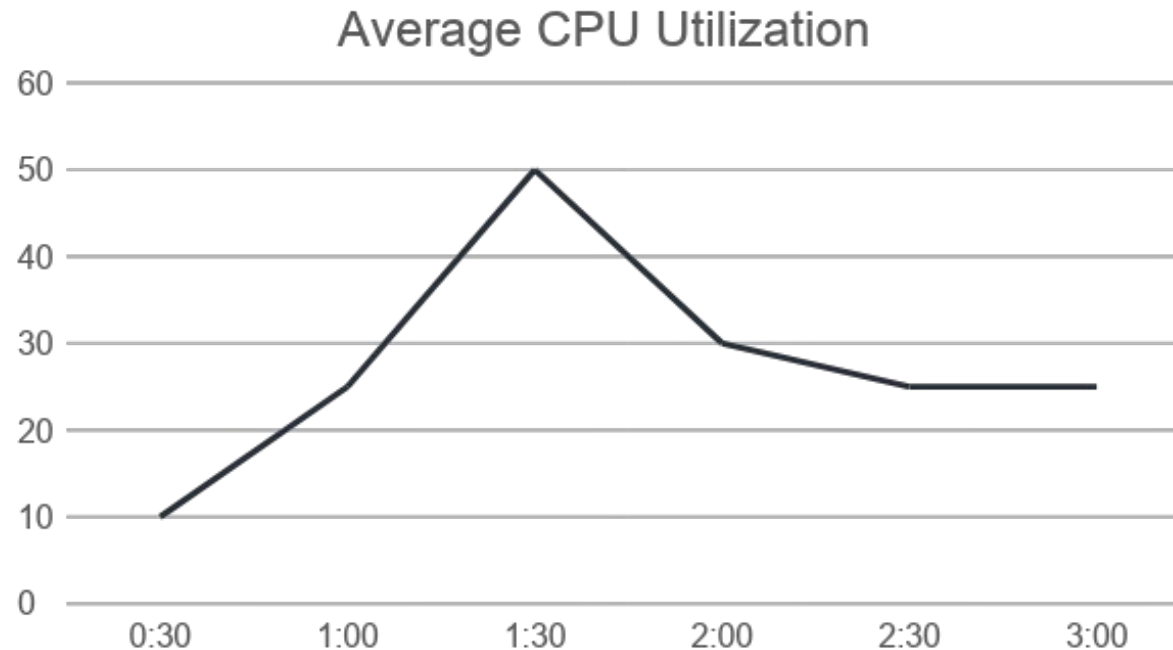
Rules

s



Targets

© 2024, EAZYTraining. All rights reserved.



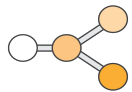
Metric data is kept for 15 months



# Amazon CloudWatch -Logs



Metrics



Logs



Alarms



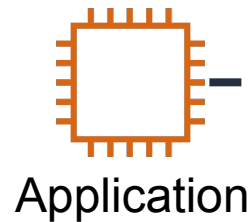
Events



Rules



Targets



Application

Log\_File.txt

Errors: 3

Warnings: 12

Connections: 20

Print out...



Amazon  
CloudWatch



Amazon S3

Source examples

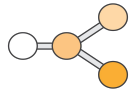
- VPC Flow Logs
- Amazon Route 53
- Elastic Load Balancing access logs



# Amazon CloudWatch -alarms



Metrics



Logs



Alarms



Events

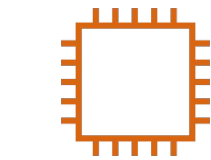


Rules



Targets

© 2024, EAZYTraining. All rights reserved.



Application

CPUUtilization metric

80% 60% 45% 25% 10% 10% 10% 10% 5%

Alarm

If CPUUtilization metric is > 50% for 5 minutes

Trigger an action like:

- Envoyer une notification à l'équipe
- Créer une autre instance pour contenir la charge

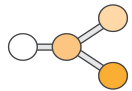




# Amazon EventBridges -alarms



Metrics



Logs



Alarms



Events

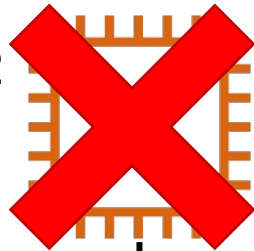


Rules



Targets

Event: EC2  
instance  
termination



## Exemple d' Event

- Modification d'une ressource AWS, telle que -
  - Connexion à la console
  - Changement d'état d'une instance EC2
  - Changement d'état de l'EC2 Auto Scaling
  - Création d'un volume EBS
- Appel à l'API AWS
- Événements des partenaires SaaS
- Événements provenant de vos propres applications

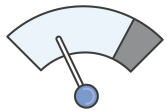


Amazon  
EventBridge

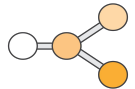


# Amazon EventBridges-rules

Rule example



Metrics



Logs



Alarms



Events



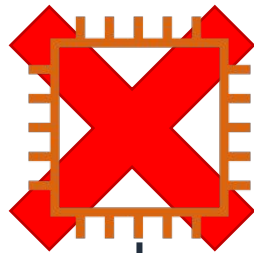
Rules



Targets

© 2024, EAZYTraining. All rights reserved.

Event



```
{
  "source": [
    "aws.ec2"
  ],
  "detail-type": [
    "EC2 Instance State-change Notification"
  ],
  "detail": {
    "state": [
      "terminated" ]
  }
}
```



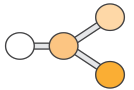
Amazon  
EventBridge



# Amazon EventBridges -Targets



Metrics



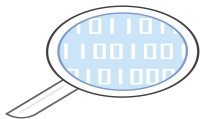
Logs



Alarms



Events



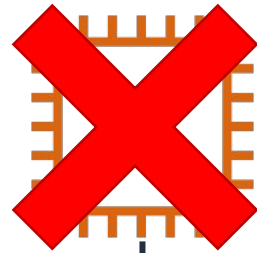
Rules



Targets

© 2024, EAZYTraining. All rights reserved.

Event

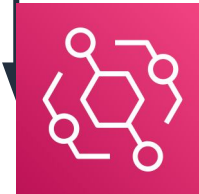


Exemple de Rule

```
{
  "source": [ "aws.ec2" ],
  "detail-type": [ "EC2
Instance State-change
Notification" ],
  "detail": {
    "state": [ "terminated" ]
  }
}
```

Exemple de Target

- EC2 instances
- AWS Lambda
- Kinesis streams
- Amazon ECS
- Step Functions
- Amazon SNS
- Amazon SQS

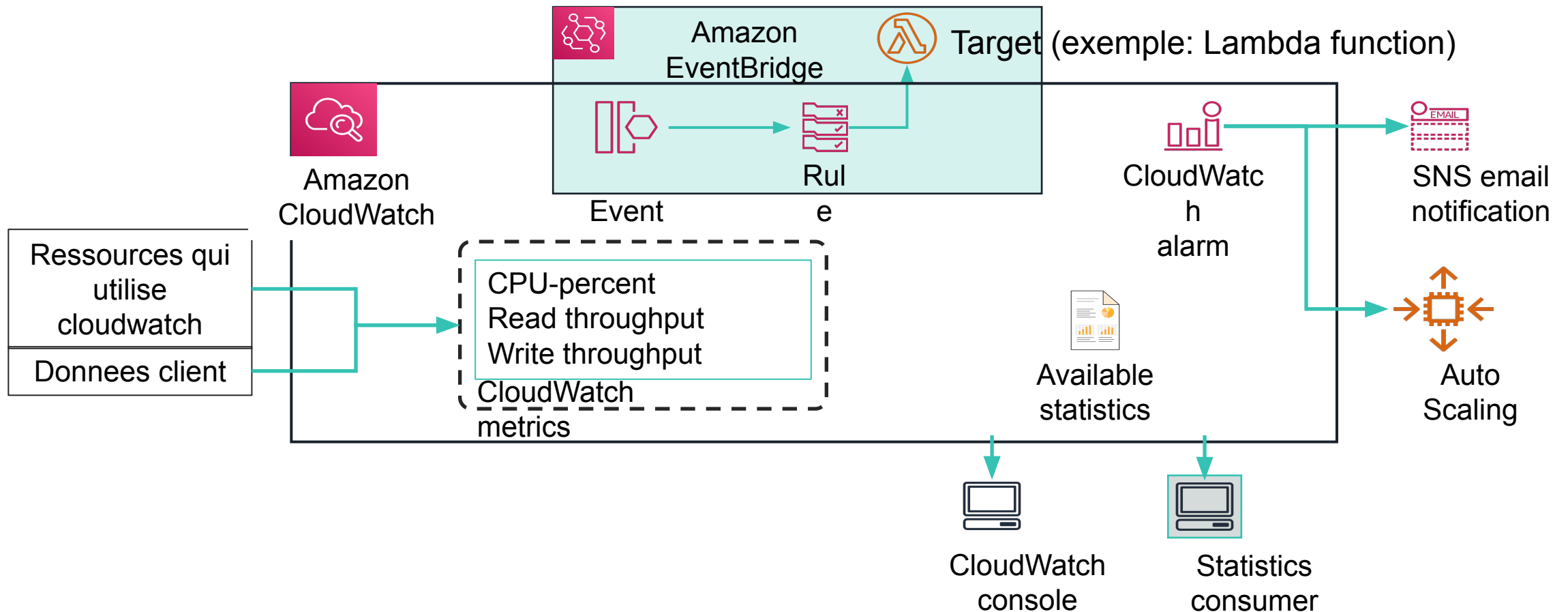


Amazon  
EventBridge





# Fonctionnement de Cloudwatch & EventBridge





# Points clés



- AWS Cost Explorer, AWS Budgets, AWS Cost and Usage Report et Cost Optimization Monitor peuvent vous aider à comprendre et à gérer le coût de votre infrastructure AWS.
- CloudWatch collecte des données de surveillance et d'exploitation sous forme de journaux, de mesures et d'événements. Il visualise les données à l'aide de tableaux de bord automatisés afin que vous puissiez obtenir une vue unifiée de vos ressources, applications et services AWS qui s'exécutent sur AWS et sur site.
- EventBridge est un service de bus d'événements sans serveur qui connecte vos applications avec des données provenant de diverses sources. EventBridge fournit un flux de données en temps réel provenant de vos propres applications, d'applications SaaS et de services AWS. Il achemine ensuite ces données vers des cibles.



# Résumé

En résumé, dans ce module, vous avez appris à :

- Utiliser Amazon EC2 Auto Scaling au sein d'une architecture pour promouvoir l'élasticité
- Expliquer comment mettre à l'échelle les ressources de votre base de données
- Déployer un équilibreur de charge d'application pour créer un environnement hautement disponible
- Utiliser Amazon Route 53 pour le basculement DNS
- Créer un environnement hautement disponible
- Concevoir des architectures qui utilisent Amazon CloudWatch pour surveiller les ressources et réagir en conséquence

MERCI POUR VOTRE AIMABLE  
ATTENTION!



**Lahda Biassou Alphonsine**

Ingénieure cloud et Formatrice