Fiche de révision Amazon EC2

Amazon EC2

Généralités

- **Définition** : Infrastructure en tant que service, permet de louer des machines virtuelles.
- **Composants** :
  - **Instances EC2** : Machines virtuelles
  - **Volumes EBS** : Stockage persistant
  - **Elastic Load Balancer** : Répartition de charge
  - **Auto Scaling Group (ASG)** : Mise à l'échelle automatique
- **Choix des instances** :
  - **Système d'exploitation** : Linux, Windows, Mac OS
  - **Puissance de calcul** : Nombre de CPU
  - **Mémoire vive (RAM)** : Quantité de mémoire
  - **Stockage** :
    - **EBS** : Volumes de stockage réseau
    - **EFS** : Système de fichiers partagé
    - **Store d'instances** : Stockage connecté au matériel
  - **Réseau** :
    - **Cartes réseau** : Performances réseau
    - **IP publique** : Adresse IP accessible
  - **Sécurité** :
    - **Groupes de sécurité** : Règles de pare-feu
- **Bootstrapping (EC2 User Data)** :
  - Scripts exécutés au démarrage de l'instance pour automatiser des tâches (installation de logiciels, mises à jour, etc.)
- **Types d'instances EC2**
- **Exemple de types d'instances** :
  - **t2.micro** : 1 vCPU, 1 Go RAM, stockage EBS, performances réseau faibles à modérées.
  - **t2.xlarge** : 4 vCPU, 16 Go RAM, performances réseau modérées.
  - **c5d.4xlarge** : 16 vCPU, 32 Go RAM, stockage NVMe SSD, performances réseau élevées.
  - **r5.16xlarge** : 64 vCPU, 512 Go RAM, performances réseau très élevées.
  - **m5.8xlarge** : 32 vCPU, 128 Go RAM, performances réseau élevées.

- **Services Associés**
    - **EBS (Elastic Block Store)** : Stockage persistant pour les instances EC2.
    - **EFS (Elastic File System)** : Système de fichiers partagé.
    - **S3 (Simple Storage Service)** : Stockage d'objets évolutif.
- **Ressources**
    - **Documentation AWS** : Pour des détails approfondis sur chaque service.
    - **AWS Free Tier** : Pour expérimenter gratuitement avec certains services, comme les instances t2.micro (jusqu'à 750 heures par mois).

Types d'instances

- **Instances à Usage Général** :
    - **Exemples** : t2.micro, M5.large
    - **Cas d'utilisation** : Serveurs web, référentiels de code
    - **Caractéristiques** : Bon équilibre entre CPU, mémoire et mise en réseau
- **Instances Optimisées pour le Calcul** :
    - **Exemples** : C5, C6
    - **Cas d'utilisation** : Traitement de données par lots, transcodage de médias, HPC, apprentissage automatique, serveurs de jeux dédiés
    - **Caractéristiques** : Optimisées pour les tâches à forte intensité de calcul
- **Instances Optimisées pour la Mémoire** :
    - **Exemples** : R5, X1, Z1
    - **Cas d'utilisation** : Bases de données relationnelles et non relationnelles, caches distribués, applications de BI, traitement en temps réel de grandes données non structurées
    - **Caractéristiques** : Performances élevées pour les charges de travail nécessitant beaucoup de RAM
- **Instances Optimisées pour le Stockage** :
    - **Exemples** : I3, D2, H1
    - **Cas d'utilisation** : Systèmes OLTP, bases de données relationnelles et NoSQL, entreposage de données, systèmes de fichiers distribués
    - **Caractéristiques** : Accès à des ensembles de données volumineux sur le stockage local
- **Convention de Nommage des Instances**
    - **Format** : Classe.Génération.Taille
    - **Exemple** : m5.2xlarge

        - **Classe** : m (usage général)
        - **Génération** : 5 (cinquième génération)
        - **Taille** : 2xlarge (taille de l'instance)
- **Classes Courantes** :
    - **t** : Usage général (ex : t2.micro)
    - **m** : Usage général (ex : m5.large)
    - **c** : Optimisé pour le calcul (ex : c5.large)
    - **r** : Optimisé pour la mémoire (ex : r5.large)
    - **i** : Optimisé pour le stockage (ex : i3.large)
    - **x** : Optimisé pour la mémoire avec une très haute capacité (ex : x1e.xlarge)
    - **z** : Optimisé pour la mémoire (ex : z1d.large)
- **Caractéristiques des Instances**
    - **t2.micro** :
        - **vCPU** : 1
        - **Mémoire** : 1 Go RAM
        - **Stockage** : EBS
        - **Performances Réseau** : Faibles à modérées
        - **Niveau Gratuit AWS** : Jusqu'à 750 heures par mois
    - **r5.16xlarge** :
        - **vCPU** : 64
        - **Mémoire** : 512 Go RAM
        - **Stockage** : EBS
        - **Performances Réseau** : Très élevées
    - **c5d.4xlarge** :
        - **vCPU** : 16
        - **Mémoire** : 32 Go RAM
        - **Stockage** : NVMe SSD attaché à l'instance
        - **Performances Réseau** : Très élevées
- **Utilisation et Optimisation des Instances EC2**
    - **Utilisation** : Lancer des instances EC2 adaptées aux besoins spécifiques des applications
    - **Optimisation** : Choisir le type d'instance en fonction des exigences de performance (CPU, mémoire, stockage)

Tarification EC2

Instances réservées

- **Durée :** 1 ou 3 ans.
- **Utilisation :** Longues charges de travail, comme une base de données qui doit fonctionner longtemps.
- **Avantage :** Réductions significatives jusqu'à 72% par rapport aux instances à la demande.
- **Options :**
  - **Paiement à l'avance :** Complet, partiel ou aucun.
  - **Convertibles :** Permettent de modifier le type d'instance, la famille, le système d'exploitation, etc., avec des remises jusqu'à 66%.

Savings Plans

- **Durée :** 1 ou 3 ans.
- **Utilisation :** Longues charges de travail, avec flexibilité sur la taille des instances.
- **Avantage :** Engagement sur une utilisation en dollars (par exemple, 10 $ de l'heure).
- **Limitation :** Spécifique à une famille d'instances et à une région.

Spot Instances (Instances ponctuelles)

- **Utilisation :** Charges de travail très courtes et non critiques.
- **Avantage :** Remises agressives jusqu'à 90%.
- **Limitation :** Instances peuvent être perdues à tout moment si le prix au comptant dépasse votre maximum.

Hôtes dédiés

- **Utilisation :** Exigences de conformité, licences logicielles spécifiques.
- **Avantage :** Contrôle complet du serveur physique.
- **Coût :** Plus élevé, avec options à la demande ou réservées pour 1 ou 3 ans.

Instances dédiées

- **Utilisation :** Matériel dédié sans contrôle sur l'emplacement des instances.
- **Différence avec hôtes dédiés :** Partage du matériel avec d'autres instances du même compte.

Réservations de capacité

- **Utilisation :** Réservation de capacité dans une AZ spécifique pour n'importe quelle durée.
- **Avantage :** Accès garanti à la capacité réservée.
- **Coût :** Tarif à la demande, même si les instances ne sont pas utilisées.

Résumé

1. **À la demande :** Idéal pour des charges de travail courtes et imprévisibles, payant un prix élevé sans engagement à long terme.
2. **Réservées :** Pour des charges de travail prévisibles et longues, avec des réductions jusqu'à 72%.
3. **Plans d'épargne :** Engagement sur une utilisation spécifique en dollars, offrant flexibilité sur la taille des instances et réductions similaires aux instances réservées.
4. **Ponctuelles :** Pour des charges de travail courtes et non critiques, avec des remises très élevées mais moins fiables.
5. **Hôtes dédiés :** Pour des exigences de conformité et des licences logicielles spécifiques, offrant un contrôle complet du matériel.
6. **Instances dédiées :** Matériel dédié, sans contrôle sur l'emplacement des instances.
7. **Réservations de capacité :** Garantie de capacité dans une AZ spécifique, avec des coûts aux tarifs à la demande.

Comparaison de prix (exemple)

- **À la demande :** 0,10 $ par heure pour une instance m4.large dans us-east-1.
- **Ponctuelles :** Jusqu'à 61% de réduction par rapport à la demande.
- **Réservées :** Variations de prix en fonction de la durée et du paiement à l'avance.
- **Plans d'épargne :** Similaire aux réductions des instances réservées.
- **Convertibles réservées :** Réductions jusqu'à 66%.
- **Hôtes dédiés :** Tarifs à la demande, avec réservations offrant jusqu'à 70% de réduction.

Réservations de capacité : Tarifs à la demande, facturés même si les instances ne sont pas utilisées.

Amazon Machine Image (AMI)

1. **Définition d'AMI**
    - **AMI** : Amazon Machine Image, une image machine Amazon qui permet de lancer des instances EC2 avec une configuration spécifique.
    - **Personnalisation** : Les AMI peuvent être créées par AWS ou personnalisées par les utilisateurs selon leurs besoins.
2. **Contenu d'une AMI**
    - **Configuration Logicielle** : Comprend le système d'exploitation et tous les logiciels préinstallés.
    - **Préinstallation** : Permet un démarrage et une configuration plus rapides des instances EC2, car les logiciels nécessaires sont déjà inclus.
3. **Types d'AMI**
    - **AMI Publique** : Fournie par AWS, exemple populaire : Amazon Linux 2.
    - **AMI Personnalisée** : Créée et maintenue par l'utilisateur, offrant une flexibilité totale sur les configurations.
    - **AMI AWS Marketplace** : Créée et vendue par des tiers, permettant aux utilisateurs d'acheter des configurations préétablies.
4. **Création et Gestion des AMI**
    - **Processus de Création** :
        1. **Lancer une Instance EC2** : Déployer et personnaliser une instance EC2.
        2. **Arrêter l'Instance** : Assurer l'intégrité des données avant de créer une AMI.
        3. **Créer une AMI** : Générer une AMI à partir de l'instance personnalisée, ce qui inclut la création d'instantanés EBS.
5. **Copie Régionale** : Les AMI peuvent être copiées d'une région AWS à une autre pour profiter de l'infrastructure mondiale.

Évolutivité et Haute Disponibilité

Évolutivité et Haute Disponibilité

1. **Évolutivité** : Capacité d'une application à gérer une augmentation de la charge.
    - Évolutivité Verticale : Augmenter la taille d'une instance (ex. passer d'un t2.micro à un t2.large).
    - Évolutivité Horizontale (Élasticité) : Augmenter le nombre d'instances (ex. ajouter plusieurs instances EC2).
2. **Haute Disponibilité** : Exécution d'une application sur plusieurs zones de disponibilité (AZ) pour assurer la continuité du service en cas de défaillance d'une AZ.

Évolutivité

- **Évolutivité**
    - Capacité d'une application à gérer une augmentation de la charge.
- **Évolutivité Verticale**
    - Définition : Augmenter les ressources d'une instance individuelle.
    - Exemple : Passer d'un t2.micro à un t2.large.
    - Usage : Systèmes non distribués comme les bases de données.
    - Limites : Contrainte par les capacités matérielles maximales disponibles.
- **Évolutivité Horizontale (Élasticité)**
    - Définition : Ajouter ou retirer des instances pour gérer la charge.
    - Exemple : Ajouter plusieurs opérateurs dans un centre d'appels pour traiter plus d'appels.
    - Usage : Applications Web modernes et distribuées.
    - Implémentation AWS : Utilisation de groupes de mise à l'échelle automatique (Auto Scaling Groups) et des Load Balancers (ELB).

Haute Disponibilité

1. **Définition** :
    a. Assurer la disponibilité continue de l'application en déployant des instances sur plusieurs AZ.
2. **Implémentation AWS** :

a. Utilisation de plusieurs AZ pour déployer des instances via Auto Scaling Groups et Load Balancers.

Services AWS

1. **Auto Scaling Groups (ASG)**
   - Fonction : Ajuste automatiquement le nombre d'instances en fonction des besoins de l'application.
   - Composants :
     - Launch Configuration : Spécifie le type d'instance, l'AMI, les clés SSH, les groupes de sécurité.
     - Scaling Policies : Règles pour ajuster le nombre d'instances.
2. **Elastic Load Balancing (ELB)**
   - Fonction : Répartit le trafic entrant sur plusieurs instances pour assurer une distribution de la charge.
   - Types :
     - Application Load Balancer (ALB) : Pour les applications HTTP/HTTPS, routage basé sur le contenu.
     - Network Load Balancer (NLB) : Pour les applications nécessitant une faible latence, routage basé sur IP.
     - Classic Load Balancer (CLB) : Pour les applications existantes nécessitant un équilibrage de charge simple.

Résumé

- Évolutivité Verticale : Augmentation de la taille d'une instance.
- Évolutivité Horizontale : Augmentation du nombre d'instances via Auto Scaling.
- Haute Disponibilité : Déploiement sur plusieurs AZ pour tolérance aux pannes.
- Elastic Load Balancing : Répartition de la charge pour équilibrer le trafic.
- Auto Scaling : Ajustement automatique des instances pour répondre à la demande.
- Elasticité : Ajustement dynamique des ressources, optimisé pour les coûts.
- Agilité : Déploiement rapide des ressources pour une réponse rapide aux besoins.

Elastic Load Balancing (ELB)

- Amazon Elastic Load Balancing (ELB) permet de distribuer le trafic entrant vers plusieurs instances en aval pour assurer la haute disponibilité et l'élasticité de vos applications.
- **Concepts Clés**
    - **Fonction** : Répartir le trafic internet vers plusieurs serveurs backend (instances EC2).
    - **Bénéfices** :
        - Répartition de la charge sur plusieurs instances.
        - Un seul point d'accès (DNS) pour les utilisateurs.
        - Gestion des défaillances des instances en aval.
        - Terminaison SSL (HTTPS).
        - Utilisation sur plusieurs zones de disponibilité pour la haute disponibilité.
    - **Types d'ELB**
        - Application Load Balancer (ALB) : Couches 7, pour HTTP/HTTPS.
        - Network Load Balancer (NLB) : Couches 4, pour TCP/UDP, très performant.
        - Gateway Load Balancer (GWLB) : Couches 3, pour router le trafic vers des appliances de sécurité.
        - Classic Load Balancer (CLB) : Couches 4 et 7, obsolète en 2023.
    - **Fonctionnement**
        - Répartition de la Charge
            1. Les utilisateurs accèdent à l'ELB via un DNS.
            2. L'ELB dirige le trafic vers les instances EC2 en fonction de la charge et des contrôles de santé.
            3. Si une instance est défaillante, l'ELB ne lui enverra pas de trafic.
        - Haute Disponibilité
            1. Utilisation d'instances sur plusieurs AZ pour assurer la continuité du service.
            2. Masquer les défaillances d'instances grâce à l'ELB.
- **Types d'ELB et Cas d'Usage**
- **Application Load Balancer (ALB)**
    - Couches 7 : HTTP, HTTPS, gRPC.
    - Usage : Applications Web, nécessitant du routage HTTP avancé.

    - Caractéristiques :
        - DNS statique.
        - Routage basé sur le contenu.
        - Intégration avec des services AWS (ex. ECS, Lambda).
- **Network Load Balancer (NLB)**
    - Couches 4 : TCP, UDP.
    - Usage : Applications nécessitant des performances élevées.
    - Caractéristiques :
        - IP statique via IP élastiques.
        - Capacité de gérer des millions de requêtes par seconde.
- **Gateway Load Balancer (GWLB)**
    - Couches 3 : GENEVE sur paquets IP.
    - Usage : Routage vers des appliances de sécurité pour inspection du trafic.
    - Caractéristiques :
        - Permet l'analyse du trafic avec des appliances virtuelles de sécurité (pare-feu, détection d'intrusion).
- **Différences et Sélection d'ELB**
    - ALB vs. NLB vs. GWLB
        - ALB : Pour les applications Web nécessitant du routage HTTP/HTTPS.
        - NLB : Pour des applications nécessitant des performances élevées et des protocoles TCP/UDP.
        - GWLB : Pour l'inspection du trafic et des opérations de sécurité sur les paquets IP.
- **Architecture et Implémentation**
    - Architecture de Base
        - Les utilisateurs accèdent à l'ELB via DNS.
        - L'ELB répartit le trafic vers les instances EC2 cibles.
        - Les réponses des instances EC2 sont renvoyées aux utilisateurs via l'ELB.
    - Configuration et Gestion
        - AWS gère l'infrastructure de l'ELB (maintenance, mises à jour, haute disponibilité).
        - Configuration via la console AWS, CLI, ou des API.
- **Cas d'Utilisation Spécifiques**
    - Routage et Sécurité
        - ALB : Routage basé sur les URL, headers HTTP, cookies.

        - NLB : Utilisation d'IP statiques pour des connexions stables et performantes.
        - GWLB : Inspection approfondie des paquets IP pour des applications nécessitant une sécurité renforcée.
    - Exemples Concrets
        - ALB : E-commerce, plateformes de contenu.
        - NLB : Applications financières, jeux en ligne.
        - GWLB : Solutions de sécurité, pare-feu, systèmes de détection d'intrusion.
- **Résumé**
    - ELB : Service géré par AWS pour répartir le trafic.
    - Types d'ELB : ALB (HTTP/HTTPS), NLB (TCP/UDP), GWLB (sécurité).
    - Avantages : Haute disponibilité, élasticité, simplification de la gestion.
    - Utilisation : Adapté aux besoins spécifiques de l'application (Web, performance, sécurité).

Auto Scaling Groups (ASG)

- Amazon Auto Scaling Groups (ASG) permet d'ajuster automatiquement le nombre d'instances EC2 en fonction de la demande de votre application, assurant ainsi une utilisation optimale des ressources et une haute disponibilité.
- **Concepts Clés**
    - Auto Scaling Groups (ASG)
        - Fonction : Ajouter ou supprimer automatiquement des instances EC2 en fonction de la charge.
        - Bénéfices :
            - Évolutivité automatique selon la demande.
            - Haute disponibilité des applications.
            - Optimisation des coûts grâce à une capacité optimale.
            - Surveillance et remplacement des instances défectueuses.
    - Évolutivité et Elasticité
        - Évolutivité (Scaling)
            - Verticale : Augmenter la taille des instances.
            - Horizontale : Ajouter ou supprimer des instances.

        - Elasticité : Capacité du système à adapter dynamiquement les ressources pour répondre à la demande.
- **Fonctionnement des ASG**
    - Paramètres de l'ASG
        - Taille minimale : Nombre minimum d'instances EC2 en cours d'exécution.
        - Capacité souhaitée : Nombre d'instances EC2 que vous souhaitez généralement avoir en cours d'exécution.
        - Taille maximale : Nombre maximum d'instances EC2 en cours d'exécution.
    - Processus d'Auto Scaling
        - Scaling Out : Ajout d'instances EC2 lorsque la charge augmente.
        - Scaling In : Suppression d'instances EC2 lorsque la charge diminue.
        - Surveillance : Détection des instances défectueuses et leur remplacement.
    - Intégration avec ELB
        - Enregistrement des Instances : Les nouvelles instances EC2 ajoutées par l'ASG sont enregistrées auprès de l'ELB.
        - Routage du Trafic : L'ELB distribue le trafic vers les instances EC2 en fonction de la charge.
- **Cas d'Usage**
    - Applications Web Variables
        - Exemple : Site de commerce en ligne avec des pics de trafic pendant la journée et une baisse la nuit.
        - Avantages : Capacité à s'adapter aux variations de la demande en ajoutant ou supprimant des instances EC2.
    - Applications Haute Disponibilité
        - Surveillance et Remplacement : Remplacement automatique des instances défectueuses pour maintenir la disponibilité.
        - Multi-AZ : Déploiement des instances EC2 sur plusieurs zones de disponibilité pour une meilleure résilience.
- **Configuration des ASG**
    - Création d'un ASG
        - Définir les paramètres : Taille minimale, capacité souhaitée, taille maximale.
        - Spécifier un modèle de lancement : Configuration des instances EC2 à utiliser (type d'instance, AMI, etc.).

        - Configurer les politiques de scaling : Basées sur des métriques comme l'utilisation du CPU, la mémoire, etc.
    - Surveillance et Alarmes
        - CloudWatch : Utilisation d'Amazon CloudWatch pour surveiller les performances des instances et déclencher des actions de scaling.
        - Alarmes : Configuration d'alarmes pour informer des changements de charge et des événements critiques.
- **Résumé**
    - ASG : Service géré par AWS pour ajuster automatiquement le nombre d'instances EC2.
    - Avantages : Haute disponibilité, optimisation des coûts, remplacement automatique des instances défectueuses.
    - Utilisation : Adapté aux applications avec des charges variables et nécessitant une haute disponibilité.
- **Stratégies de Mise à l'Échelle des Groupes Auto Scaling (ASG)**
    - Les groupes Auto Scaling (ASG) d'AWS permettent de s'adapter dynamiquement à la demande de votre application en ajustant le nombre d'instances EC2. Voici les différentes stratégies de mise à l'échelle que vous pouvez utiliser pour optimiser la gestion de vos ressources.
- **Stratégies de Mise à l'Échelle**
    - Mise à l'Échelle Manuelle
        - Description : Ajustement manuel de la taille d'un ASG.
        - Exemple : Modifier manuellement la capacité de 1 à 2 instances ou de 2 à 1 instance.
    - Mise à l'Échelle Dynamique
        - Description : Ajustement automatique de la taille de l'ASG en réponse à des demandes changeantes.
        - Types de Politiques :
            - Simple Scaling :
                - Déclencheur : Une alarme CloudWatch.
                - Action : Ajouter ou supprimer un nombre fixe d'instances EC2.
                - Exemple : Ajouter deux instances si l'utilisation moyenne du CPU dépasse 70% pendant 5 minutes.
            - Step Scaling :
                - Déclencheur : Une alarme CloudWatch.

        - Action : Ajouter ou supprimer des instances par étapes, selon des seuils prédéfinis.
        - Exemple : Ajouter ou retirer des unités de capacité en fonction de différents niveaux d'utilisation du CPU.
    - Target Tracking Scaling :
        - Description : Maintien d'une métrique cible pour les instances EC2.
        - Exemple : Maintenir l'utilisation moyenne du CPU à 40% en ajustant automatiquement le nombre d'instances EC2.
    - Mise à l'Échelle Programmée
        - Description : Ajustement de la taille de l'ASG à des moments spécifiques, basés sur des prévisions de la demande.
        - Exemple : Augmenter la capacité minimale à 10 instances EC2 chaque vendredi à 17h00 avant un événement sportif.
    - Mise à l'Échelle Prédictive
        - Description : Utilisation de l'apprentissage automatique pour prédire et ajuster la taille de l'ASG en fonction de la demande future.
        - Fonctionnement : Analyse des modèles de trafic passés pour prévoir les charges futures.
        - Exemple : Provisionner le bon nombre d'instances EC2 avant une période de charge élevée prévue.
- **Surveillance et Optimisation**
    - CloudWatch : Utilisez CloudWatch pour surveiller les performances et ajuster les stratégies en conséquence.
    - Alarmes et Notifications : Configurez des alarmes pour être informé des changements et des événements critiques.

Suivez-nous

- **Site web** : https://lecloudfacile.com
- **Youtube** : https://www.youtube.com/@lecloudfacile
- **Linkedin** : https://www.linkedin.com/company/lecloudfacile/
- **Udemy** : https://www.udemy.com/course/nouveau-aws-cloud-practitioner-clf-c02/?referralCode=8CE99E6C2100F1998BDF
- **Communauté WhatsApp** : https://chat.whatsapp.com/HIeIlLVOJ9xCJKX8VhbLSr