aws

Guide de l'utilisateur

Amazon EC2 Auto Scaling

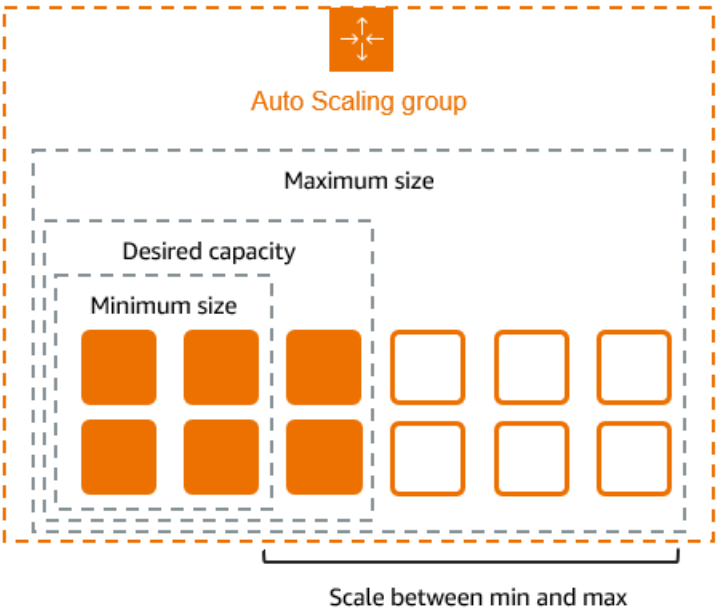
Amazon EC2 Auto Scaling: Guide de l'utilisateur

Table of Contents

Qu'est-ce qu'Amazon EC2 Auto Scaling ?

Amazon EC2 Auto Scaling permet de vous assurer que vous disposez du bon nombre d'instances Amazon EC2 disponibles pour gérer la charge de l'application. Vous créez des ensembles d'instances EC2, appelés groupes Auto Scaling. Vous pouvez spécifier le nombre minimum d'instances dans chaque groupe Auto Scaling, et Amazon EC2 Auto Scaling veille à ce que le groupe ne descende jamais en-dessous de cette taille. Vous pouvez spécifier le nombre maximum d'instances dans chaque groupe Auto Scaling et Amazon EC2 Auto Scaling veille à ce que le groupe ne dépasse jamais cette taille. Si vous spécifiez la capacité souhaitée, lorsque vous créez le groupe ou à tout moment par la suite, et Amazon EC2 Auto Scaling veille à ce que le groupe possède autant d'instances. Si vous spécifiez des politiques de mise à l'échelle, Amazon EC2 Auto Scaling peut lancer ou résilier des instances à mesure que la demande sur l'application augmente ou diminue.

Par exemple, le groupe Auto Scaling suivant possède une taille minimale de quatre instances, une capacité souhaitée de six instances et une taille maximale de douze instances. Les politiques de mise à l'échelle que vous définissez ajustent le nombre d'instances, entre le nombre minimum et maximum d'instances, en fonction des critères que vous spécifiez.



Caractéristiques d'Amazon EC2 Auto Scaling

Avec Amazon EC2 Auto Scaling, vos instances EC2 sont organisées en groupes Auto Scaling afin qu'elles puissent être traitées comme une unité logique à des fins de dimensionnement et de gestion.

Les groupes Auto Scaling utilisent des modèles de lancement (ou des configurations de lancement) comme modèles de configuration pour leurs instances EC2.

Voici les principales fonctionnalités d'Amazon EC2 Auto Scaling :

Surveillance de l'état des instances en cours d'exécution

Amazon EC2 Auto Scaling surveille automatiquement l'état et la disponibilité de vos instances à l'aide des bilans de santé EC2 et remplace les instances résiliées ou défectueuses afin de maintenir la capacité souhaitée.

Surveillances d'état personnalisées

Outre les contrôles de santé intégrés, vous pouvez définir des contrôles de santé personnalisés spécifiques à votre application afin de vérifier qu'elle répond comme prévu. Si une instance échoue à votre bilan de santé personnalisé, elle est automatiquement remplacée pour conserver la capacité souhaitée.

Équilibrer les capacités entre les zones de disponibilité

Vous pouvez spécifier plusieurs zones de disponibilité pour votre groupe Auto Scaling, et Amazon EC2 Auto Scaling équilibre vos instances de manière égale entre les zones de disponibilité au fur et à mesure que le groupe évolue. Cela garantit une disponibilité et une résilience élevées en protégeant vos applications contre les défaillances en un seul endroit.

Types d’instances et options d’achat multiples

Au sein d'un même groupe Auto Scaling, vous pouvez lancer plusieurs types d'instances et options d'achat (instances ponctuelles et à la demande), ce qui vous permet d'optimiser les coûts grâce à l'utilisation d'instances ponctuelles. Vous pouvez également profiter des remises sur les instances réservées et le Savings Plan en les utilisant conjointement avec les instances à la demande du groupe.

Remplacement automatique des instances Spot

Si votre groupe inclut des instances Spot, Amazon EC2 Auto Scaling peut automatiquement demander une capacité Spot de remplacement si vos instances Spot sont interrompues. Grâce au rééquilibrage des capacités, Amazon EC2 Auto Scaling peut également surveiller et remplacer de manière proactive vos instances Spot présentant un risque élevé d'interruption.

Equilibrage de charge

Vous pouvez utiliser l'équilibrage de charge et les contrôles de santé d'Elastic Load Balancing pour garantir une répartition uniforme du trafic applicatif vers vos instances saines. Chaque

fois que des instances sont lancées ou résiliées, Amazon EC2 Auto Scaling enregistre et désenregistre automatiquement les instances de l'équilibreur de charge.

Evolutivité

Amazon EC2 Auto Scaling propose également plusieurs méthodes pour redimensionner vos groupes Auto Scaling. L'utilisation de la mise à l'échelle automatique vous permet de maintenir la disponibilité des applications et de réduire les coûts en augmentant la capacité pour faire face aux pics de charge et en supprimant de la capacité lorsque la demande est plus faible. Vous pouvez également ajuster manuellement la taille de votre groupe Auto Scaling selon vos besoins.

Actualisation d'instance

La fonctionnalité d'actualisation des instances fournit un mécanisme permettant de mettre à jour les instances de manière continue lorsque vous mettez à jour votre AMI ou votre modèle de lancement. Vous pouvez également utiliser une approche progressive, connue sous le nom de déploiement Canary, pour tester une nouvelle AMI ou un nouveau modèle de lancement sur un petit nombre d'instances avant de le déployer dans l'ensemble du groupe.

Hooks de cycle de vie

Les hooks du cycle de vie sont utiles pour définir des actions personnalisées qui sont invoquées lors du lancement de nouvelles instances ou avant leur fermeture. Cette fonctionnalité est particulièrement utile pour créer des architectures axées sur les événements, mais elle vous aide également à gérer les instances tout au long de leur cycle de vie.

Support pour les charges de travail dynamiques

Les Lifecycle Hooks offrent également un mécanisme permettant de conserver l'état à l'arrêt. Pour garantir la continuité des applications dynamiques, vous pouvez également utiliser une protection évolutive ou des politiques de résiliation personnalisées pour empêcher les instances dont les processus sont longs de s'arrêter prématurément.

Pour plus d'informations sur les avantages d'Amazon EC2 Auto Scaling, consultez Avantages d'Auto Scaling pour l'architecture des applications.

Tarification d'Amazon EC2 Auto Scaling

Amazon EC2 Auto Scaling est gratuit. Il est donc facile de l'essayer et de découvrir les avantages qu'il peut apporter à votre AWS architecture. Vous ne payez que pour les AWS ressources (par exemple, les instances EC2, les volumes EBS et les CloudWatch alarmes) que vous utilisez.

Mise en route

Pour commencer, suivez le didacticiel Create your first Auto Scaling group pour créer un groupe Auto Scaling et voir comment il réagit lorsqu'une instance de ce groupe se termine.

Utiliser des groupes Auto Scaling

Vous pouvez créer vos groupes Auto Scaling, y accéder et les gérer à l'aide des interfaces suivantes :

- AWS Management Console – offre une interface Web que vous pouvez utiliser pour accéder à vos groupes Auto Scaling. Si vous êtes inscrit à un Compte AWS, vous pouvez accéder à vos groupes Auto Scaling en vous connectant au AWS Management Console, en utilisant le champ de recherche de la barre de navigation pour rechercher des groupes Auto Scaling, puis en choisissant Auto Scaling groups.
- AWS Command Line Interface (AWS CLI) — Fournit des commandes pour un large éventail de Services AWS, et est compatible avec Windows, macOS et Linux. Consultez Préparez-vous à utiliser AWS CLI pour démarrer. Pour plus d'informations, consultez update-auto-scaling-group dans le guide de référence des commandes AWS CLI .
- AWS Tools for Windows PowerShell— Fournit des commandes pour un large éventail de AWS produits pour ceux qui écrivent des scripts dans l' PowerShell environnement. Consultez le Guide de l'utilisateur AWS Tools for Windows PowerShell pour démarrer. Pour plus d'informations, consultez le Guide de référence des cmdlets AWS Tools for PowerShell.
- AWS SDK — Fournit des opérations d'API spécifiques au langage et prend en charge de nombreux détails de connexion, tels que le calcul des signatures, la gestion des nouvelles tentatives de demande et la gestion des erreurs. Pour plus d'informations, consultez Kits SDK AWS.
- API de requête : Fournit des actions d'API de bas niveau appelées à l'aide de demandes HTTPS. L'utilisation de l'API de requête est le moyen le plus direct d'accéder à un Services AWS. Toutefois, il faut alors que votre application gère les détails de bas niveau, notamment la génération du hachage pour signer la demande et la gestion des erreurs. Pour de plus amples informations, veuillez consulter la Référence d'API Amazon EC2 Auto Scaling.
- AWS CloudFormation— Permet de créer des groupes Auto Scaling à l'aide CloudFormation de modèles. Pour plus d’informations, consultez Créer un groupe Auto Scaling avec AWS CloudFormation.

Pour vous connecter par programmation à un Service AWS, vous utilisez un point de terminaison. .

Avantages d'Auto Scaling pour l'architecture des applications

L'ajout d'Amazon EC2 Auto Scaling à l'architecture de votre application est un moyen de maximiser les avantages du AWS cloud. Lorsque vous utilisez Amazon EC2 Auto Scaling, vos applications profitent des avantages suivants :

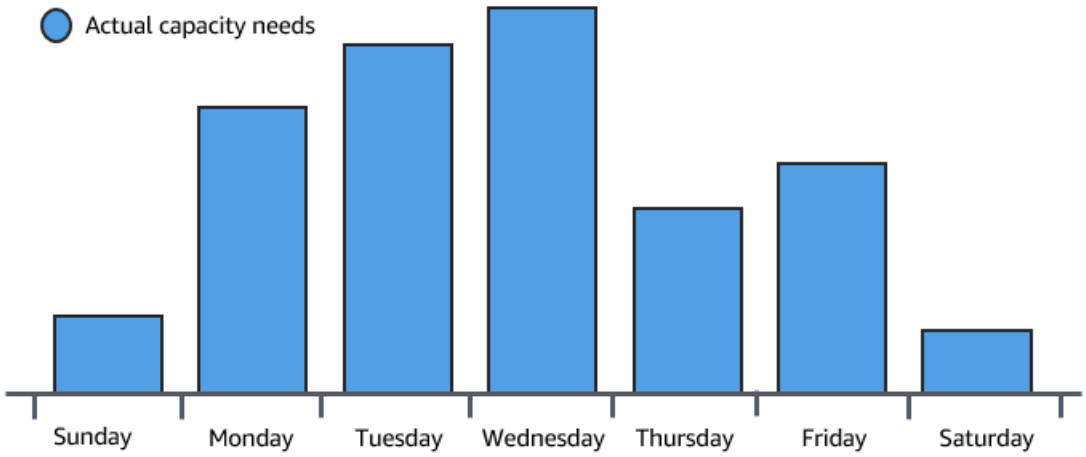
- Meilleure tolérance aux pannes. Amazon EC2 Auto Scaling peut détecter lorsqu'une instance est défaillante, la résilier et en lancer une nouvelle pour la remplacer. Vous pouvez également configurer Amazon EC2 Auto Scaling pour utiliser plusieurs zones de disponibilité. Si une zone de disponibilité devient indisponible, Amazon EC2 Auto Scaling peut lancer des instances dans une autre pour compenser.
- Meilleure disponibilité. Amazon EC2 Auto Scaling garantit que l'application dispose toujours de la bonne capacité pour gérer la demande de trafic actuelle.
- Meilleure gestion des coûts. Amazon EC2 Auto Scaling peut augmenter et diminuer dynamiquement la capacité selon les besoins. Parce que vous payez pour les instances EC2 que vous utilisez, vous économisez de l'argent en lançant des instances lorsqu'elles sont nécessaires et en les résiliant lorsqu'elles ne le sont plus.

Table des matières

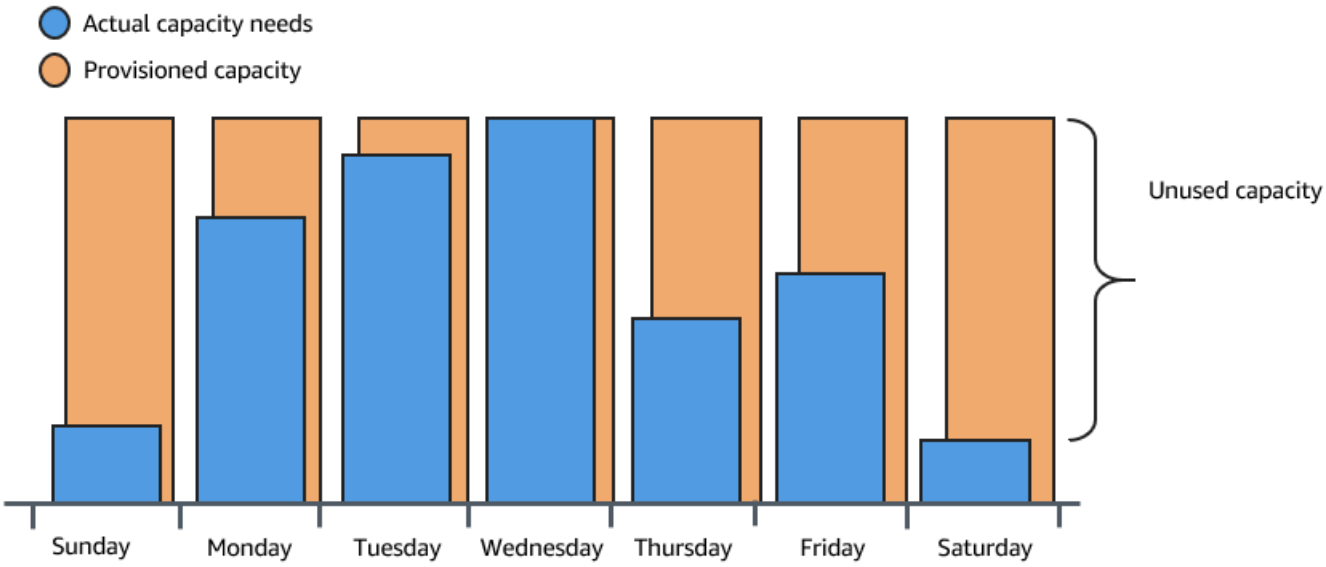
Exemple : couvrir la demande variable

Pour montrer quelques-uns des avantages d'Amazon EC2 Auto Scaling, prenez une application Web de base sur AWS. Cette application permet aux employés de chercher des salles de conférence qu'ils peuvent utiliser pour des réunions. Au début et à la fin de la semaine, l'utilisation de cette application est minime. Au milieu de la semaine, davantage d'employés planifient des réunions, la demande sur l'application augment donc considérablement.

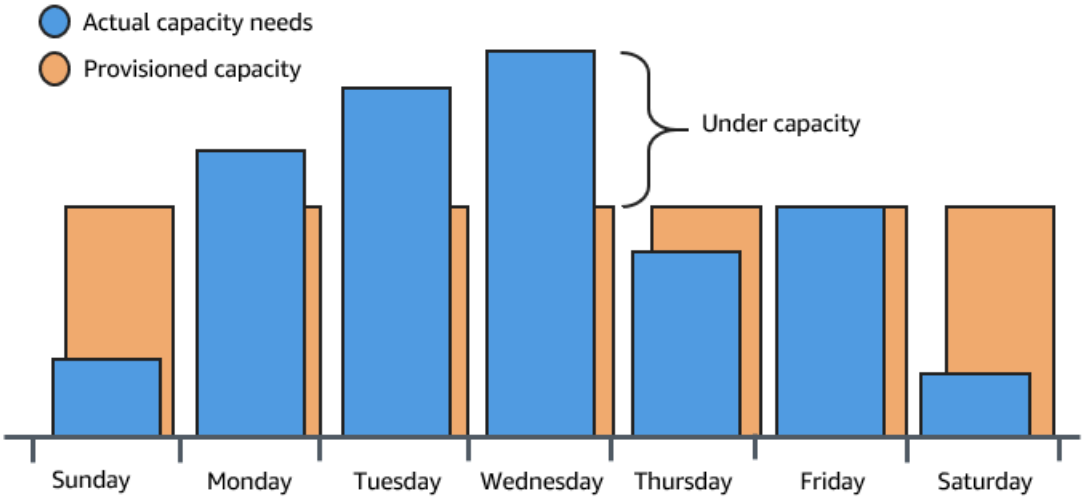
Le graphique suivant montre l'utilisation de la capacité de l'application sur une semaine.



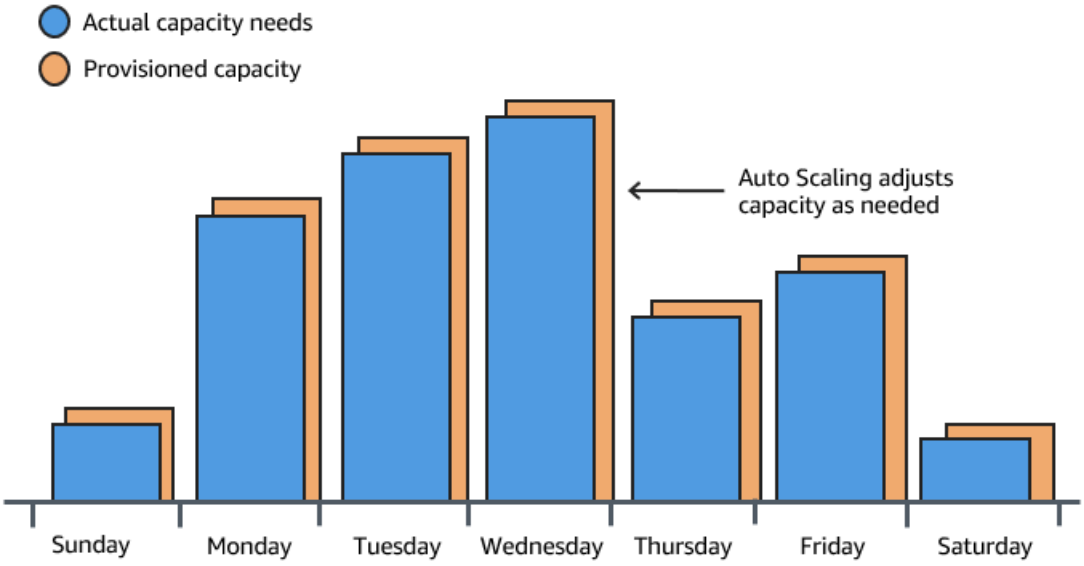
Généralement, il existe deux manières d'anticiper ces changements de capacité. La première option consiste à ajouter suffisamment de serveurs pour que l'application ait toujours assez de capacité pour répondre à la demande. L'inconvénient de cette option, cependant, est que l'application n'a pas besoin d'autant de capacité tous les jours. La capacité supplémentaire reste inutilisée et, en substance, augmente le coût de fonctionnement de l'application.



La deuxième option consiste à disposer de suffisamment de capacité pour gérer la demande moyenne sur l'application. Cette option est moins onéreuse car vous n'achetez pas d'équipement que vous utiliserez uniquement à l'occasion. Cependant, vous risquez de créer une mauvaise expérience client si la demande sur l'application dépasse sa capacité.



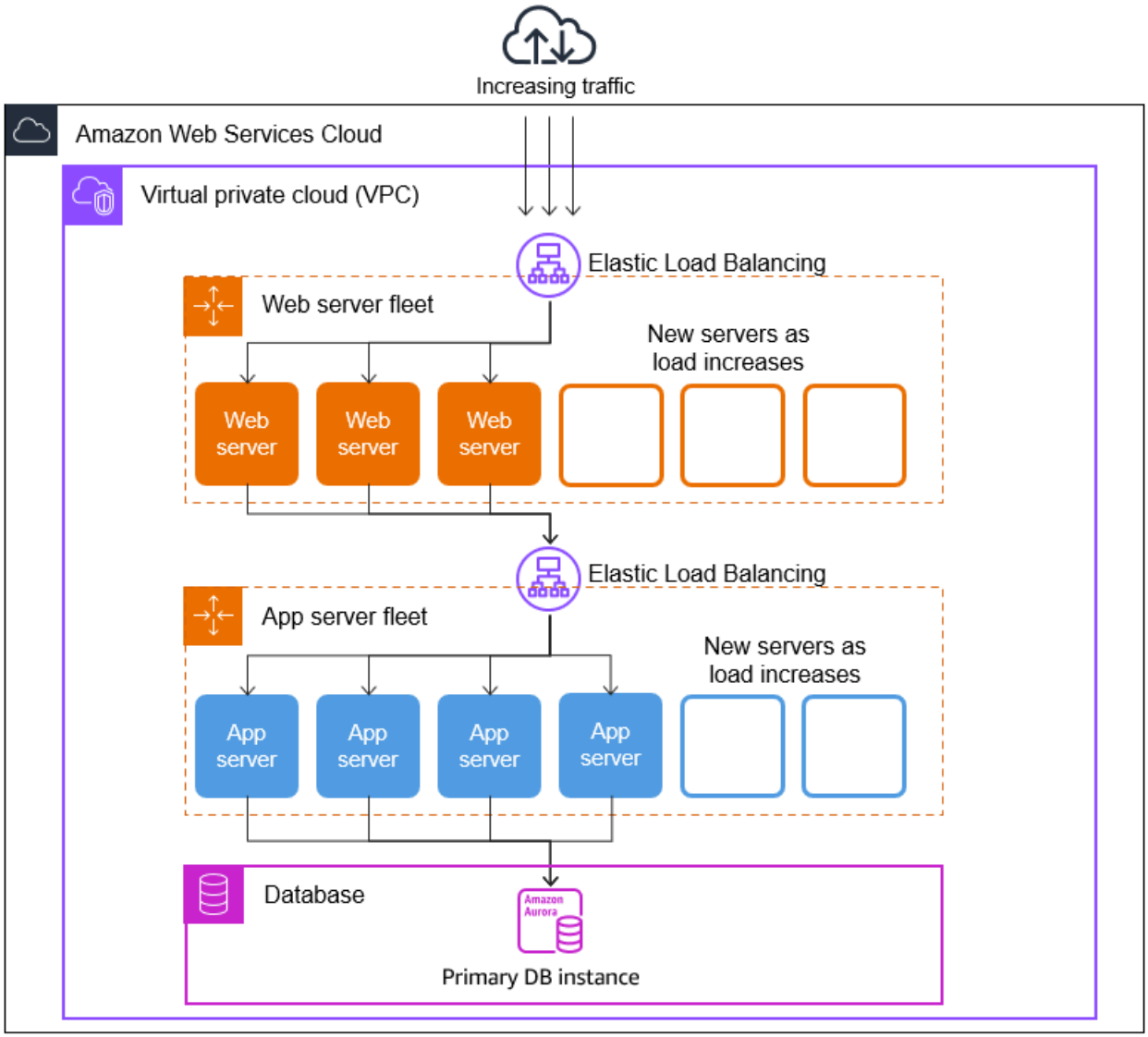
En ajoutant Amazon EC2 Auto Scaling à cette application, vous disposez d'une troisième option. Vous pouvez ajouter de nouvelles instances à l'application uniquement si nécessaire, et la résilier lorsque vous n'en avez plus besoin. Parce qu'Amazon EC2 Auto Scaling utilise des instances EC2, vous ne payez que pour les instances que vous utilisez, quand vous en avez besoin. Vous disposez désormais d'une architecture rentable qui fournit la meilleure expérience client possible tout en minimisant les dépenses.



Exemple : architecture d'application web

Dans un scénario d'application web classique, vous exécutez des copies de l'application simultanément pour couvrir le volume du trafic client. Ces multiples copies de l'applications sont hébergées sur des instances EC2 identiques (serveurs de cloud), chacune gérant des demandes clients.

Amazon EC2 Auto Scaling gère le lancement et la résiliation de ces instances EC2 en votre nom. Vous définissez un ensemble de critères (tels qu'une CloudWatch alarme Amazon) qui détermine le moment où le groupe Auto Scaling lance ou arrête les instances EC2. L'ajout de groupes Auto Scaling à l'architecture de réseau peut aider à rendre l'application plus hautement disponible et tolérante aux pannes.



Vous pouvez créer autant de groupes Auto Scaling que nécessaire. Par exemple, vous pouvez créer un groupe Auto Scaling pour chaque niveau.

Pour répartir le trafic entre les instances des groupes Auto Scaling, vous pouvez introduire un équilibreur de charge dans l'architecture. Pour plus d’informations, consultez Elastic Load Balancing.

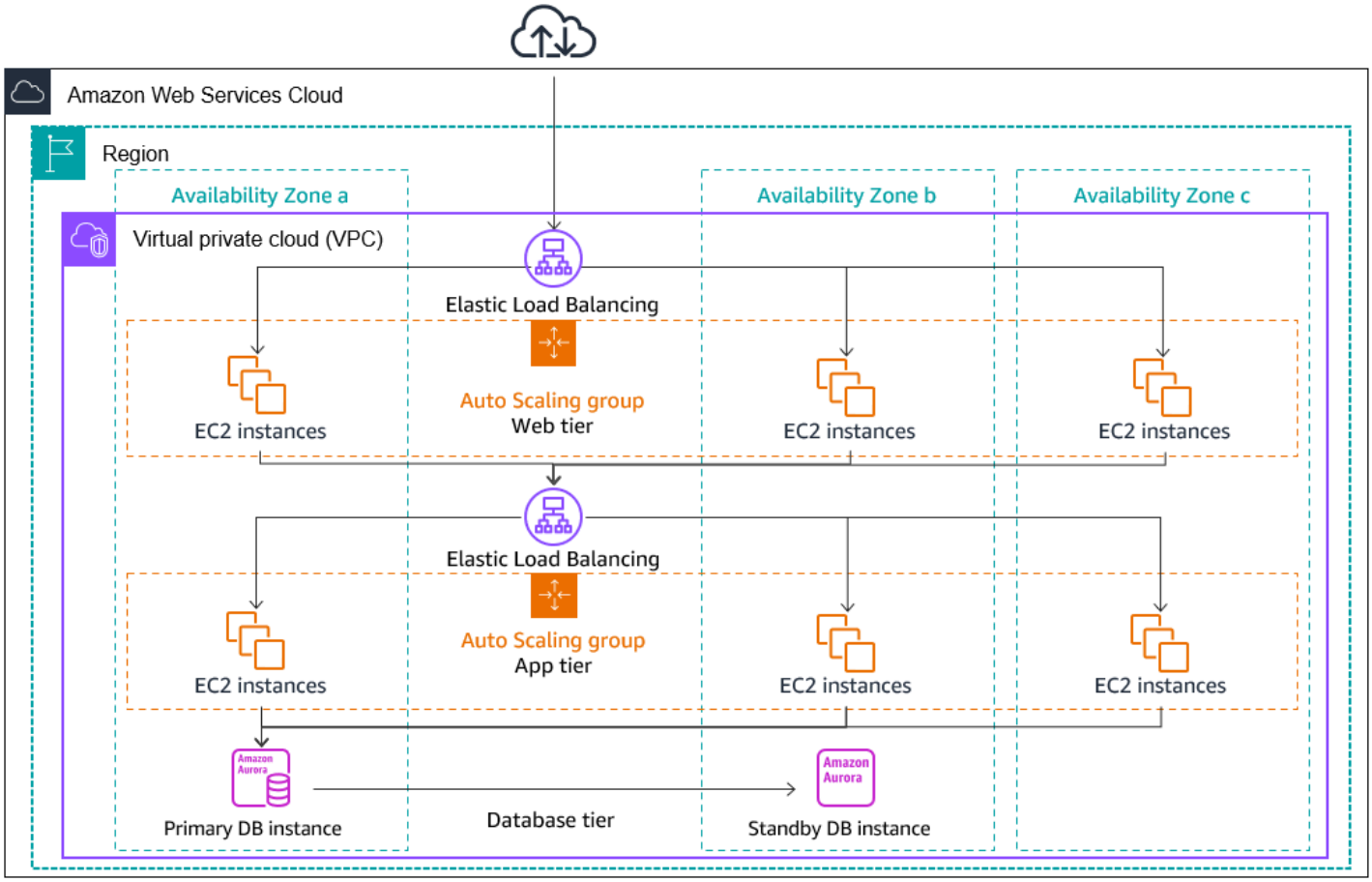
Exemple : répartir les instances dans les zones de disponibilité

Les zones de disponibilité sont des emplacements isolés dans une Région AWS donnée. Chaque Région possède plusieurs zones de disponibilité conçues pour fournir une haute disponibilité pour la Région. Les zones de disponibilité sont indépendantes. Par conséquent, vous augmentez la disponibilité des applications lorsque vous concevez votre application de manière à utiliser plusieurs zones. Pour plus d'informations, consultez Résilience dans Amazon EC2 Auto Scaling.

Une zone de disponibilité est identifiée par le Région AWS code suivi d'une lettre d'identification (par exemple,`us-east-1a`). Si vous créez votre VPC et vos sous-réseaux plutôt que d'utiliser le VPC par défaut, vous pouvez définir un ou plusieurs sous-réseaux dans chaque zone de disponibilité. Chaque sous-réseau doit résider entièrement dans une zone de disponibilité et ne peut pas s'étendre sur plusieurs zones. Pour plus d'informations, veuillez consulter la rubrique Qu'est-ce qu'Amazon VPC ? dans le Guide de l'utilisateur Amazon Virtual Private Cloud.

Lorsque vous créez un groupe Auto Scaling, vous devez choisir le VPC et les sous-réseaux dans lesquels vous allez déployer le groupe Auto Scaling. Amazon EC2 Auto Scaling crée des instances dans les sous-réseaux que vous avez choisis. Chaque instance est ainsi associée à une zone de disponibilité spécifique choisie par Amazon EC2 Auto Scaling. Lorsque les instances sont lancées, Amazon EC2 Auto Scaling essaye de les distribuer uniformément entre les zones pour une disponibilité et une fiabilité élevées.

L'image suivante illustre l'architecture multi-niveaux déployée dans trois zones de disponibilité.



Distribution des instances

Amazon EC2 Auto Scaling essaye automatiquement de maintenir un nombre équivalent d'instances dans chaque zone de disponibilité activée. Pour ce faire, Amazon EC2 Auto Scaling tente de lancer de nouvelles instances dans la zone de disponibilité qui contient le moins d'instances. Si plusieurs sous-réseaux sont choisis pour la zone de disponibilité, Amazon EC2 Auto Scaling sélectionne de manière aléatoire un sous-réseau à partir de la zone de disponibilité. Si la tentative échoue, cependant, Amazon EC2 Auto Scaling tente de lancer des instances dans une autre zone de disponibilité jusqu'à ce qu'il y parvienne.

Si une zone de disponibilité devient non saine ou indisponible, la distribution des instances peut être inégalement distribuée entre les zones de disponibilité. Lorsque la zone de disponibilité est rétablie, Amazon EC2 Auto Scaling rééquilibre automatiquement le groupe Auto Scaling. Pour ce faire, il lance des instances dans les zones de disponibilité activées avec le moins d'instances et en résiliant des instances ailleurs.

Activités de rééquilibrage

Les activités de rééquilibrage se divisent en deux catégories : le rééquilibrage des zones de disponibilité et le rééquilibrage des capacités.

Rééquilibrage des zones de disponibilité

Après certaines actions, un déséquilibrage du groupe Auto Scaling peut avoir lieu entre les zones de disponibilité. Amazon EC2 Auto Scaling compense en rééquilibrant les zones de disponibilité. Les actions suivantes peuvent entraîner une activité de rééquilibrage :

- Vous changez les zones de disponibilité associées au groupe Auto Scaling.
- Vous résiliez ou détachez explicitement des instances ou placez des instances de secours, et cela entraîne le déséquilibrage du groupe.
- Une zone de disponibilité qui auparavant ne disposait pas de suffisamment de capacité récupère et dispose désormais de capacité supplémentaire.
- Une zone de disponibilité qui avait précédemment un prix d'instance Spot supérieur à votre prix maximum a désormais un prix d'instance Spot inférieur à votre prix maximum.

Lors du rééquilibrage, Amazon EC2 Auto Scaling lance de nouvelles instances avant de résilier les anciennes. Ainsi, le rééquilibrage ne compromet pas les performances ou la disponibilité de votre application.

Comme Amazon EC2 Auto Scaling tente de lancer de nouvelles instances avant de résilier les anciennes, le fait d'atteindre la capacité maximale spécifiée ou de s'en approcher peut entraver ou interrompre complètement les activités de rééquilibrage.

Pour contourner ce problème, le système peut temporairement dépasser la capacité maximale spécifiée d'un groupe pendant une activité de rééquilibrage. Par défaut, il peut le faire avec une marge de 10 % ou une instance, selon la valeur la plus élevée. La marge est étendue uniquement si le groupe atteint la capacité maximale ou s'en approche et nécessite un rééquilibrage, L'extension dure uniquement le temps de rééquilibrer le groupe (généralement pendant quelques minutes).

Vous pouvez également établir des seuils pour un groupe Auto Scaling en utilisant une politique de maintenance des instances. Le groupe ne peut ainsi augmenter ou diminuer la capacité que dans cette plage de seuils. Ainsi, vous pouvez contrôler la rapidité avec laquelle votre groupe se rééquilibre. Pour plus d’informations, consultez Politiques de maintenance des instances.

Rééquilibrage de la capacité

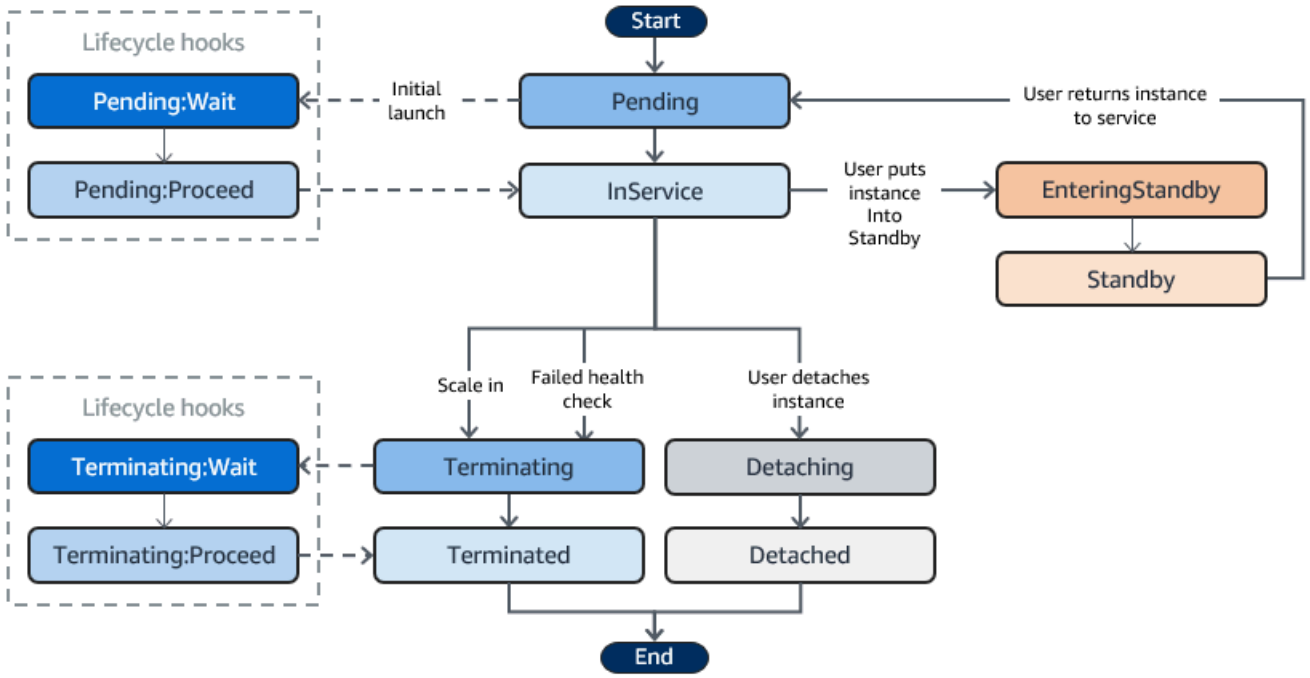
Vous pouvez activer le rééquilibrage des capacités pour vos groupes Auto Scaling lorsque vous utilisez des instances Spot afin qu'Amazon EC2 Auto Scaling tente de lancer une instance Spot chaque fois qu'Amazon EC2 signale un risque élevé d'interruption d'une instance Spot. Dès qu'une nouvelle instance est lancée, une ancienne instance est résiliée. Pour plus d'informations, consultez Utiliser le rééquilibrage de la capacité pour gérer les interruptions Spot Amazon EC2.

Cycle de vie d'une instance Amazon EC2 Auto Scaling

Les instances EC2 dans un groupe Auto Scaling disposent d'un chemin, ou d'un cycle de vie, qui diffère des autres instances EC2. Le cycle de vie commence lorsque le groupe Auto Scaling lance une instance et la met en service. Le cycle de vie se termine lorsque vous résiliez l'instance, ou le groupe Auto Scaling met l'instance hors service et la résilie.

> **Note**
>
> Vous êtes facturé pour les instances dès qu'elles sont lancées, y compris lorsqu'elles ne sont pas encore en service.

L'illustration suivante représente les transitions entre les états de l'instance dans le cycle de vie Amazon EC2 Auto Scaling.



Monter en puissance

Les événements suivants d'augmentation de la taille des instances demandent au groupe Auto Scaling de lancer des instances EC2 et de les attacher au groupe :

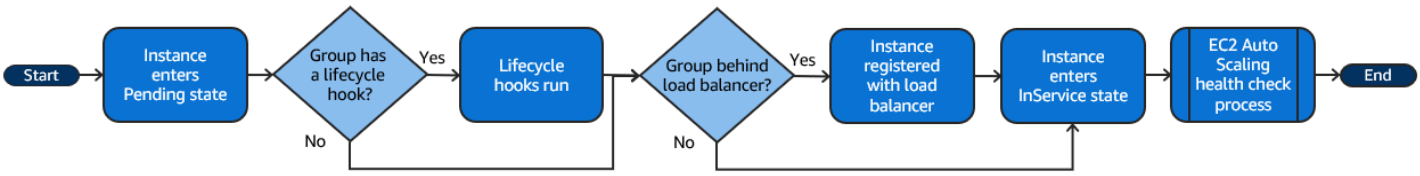
- Vous augmentez manuellement la taille du groupe. Pour plus d'informations, consultez Changer la capacité souhaitée d'un groupe Auto Scaling existant.
- Vous créez une politique de mise à l'échelle pour augmenter automatiquement la taille du groupe en fonction de la hausse spécifiée dans la demande. Pour plus d'informations, consultez Mise à l'échelle dynamique pour Amazon EC2 Auto Scaling.
- Vous configurez la mise à l'échelle selon le calendrier pour augmenter la taille du groupe à un moment spécifique. Pour plus d'informations, consultez Mise à l'échelle planifiée pour Amazon EC2 Auto Scaling.

Lorsqu'un événement de montée en puissance se produit, le groupe Auto Scaling lance le nombre requis d'instances EC2, en utilisant le modèle de lancement qui lui a été attribué. Ces instances démarrent avec l'état `Pending`. Si vous ajoutez un hook de cycle de vie au groupe Auto Scaling, vous pouvez réaliser une action personnalisée. Pour plus d'informations, consultez Hooks de cycle de vie.

Lorsque chaque instance est entièrement configurée et réussit les surveillances de l'état Amazon EC2, elle est attachée au groupe Auto Scaling et passe en statut `InService`. L'instance est décomptée de la capacité souhaitée du groupe Auto Scaling.

Si votre groupe Auto Scaling est configuré pour recevoir le trafic d'un équilibreur de charge Elastic Load Balancing, Amazon EC2 Auto Scaling enregistre automatiquement votre instance auprès de l'équilibreur de charge avant de la marquer comme `InService`.

Ce qui suit résume les étapes d'enregistrement d'une instance auprès d'un équilibreur de charge pour un événement de scale-out.



Instances en service

Les instances restent en statut `InService` jusqu'à ce que l'un des événements suivants se produise :

- Un événement de mise à l'échelle horizontale se produit, et Amazon EC2 Auto Scaling choisit de résilier cette instance pour réduire la taille du groupe Auto Scaling. Pour plus d'informations, consultez Contrôler les instances à scalabilité automatique à résilier pendant une mise à l'échelle horizontale.
- Vous mettez l'instance en statut `Standby`. Pour plus d'informations, consultez Entrer et sortir du mode veille.
- Vous détachez l'instance du groupe Auto Scaling. Pour plus d'informations, consultez Détacher ou attacher des instances.
- L'instance échoue au nombre requis de surveillances de l'état, elle est supprimée du groupe Auto Scaling, résiliée et remplacée. Pour plus d'informations, consultez Surveillance de l'état des instances dans un groupe Auto Scaling.

Mise à l'échelle horizontale

Les événements suivants de diminution de la taille des instances demandent au groupe Auto Scaling de détacher les instances EC2 du groupe et de les résilier.

- Vous diminuez manuellement la taille du groupe. Pour plus d'informations, consultez Changer la capacité souhaitée d'un groupe Auto Scaling existant.
- Vous créez une politique de mise à l'échelle pour diminuer automatiquement la taille du groupe en fonction de la baisse spécifiée de la demande. Pour plus d'informations, consultez Mise à l'échelle dynamique pour Amazon EC2 Auto Scaling.
- Vous configurez la mise à l'échelle selon le calendrier pour diminuer la taille du groupe à un moment spécifique. Pour plus d'informations, consultez Mise à l'échelle planifiée pour Amazon EC2 Auto Scaling.

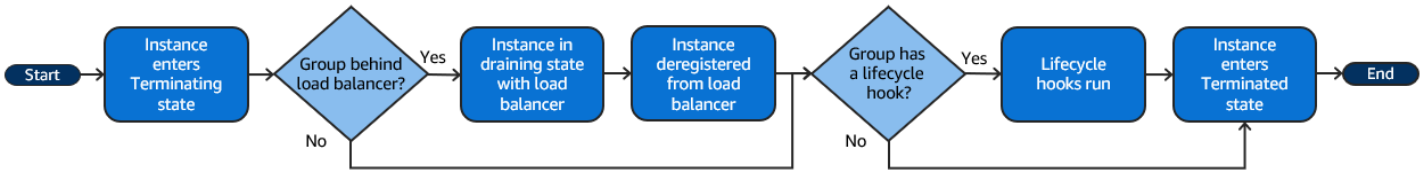
Il est important de créer un événement de diminution de la taille des instances pour chaque événement d'augmentation de la taille des instances que vous créez. Cela garantit que les ressources attribuées à l'application correspondent aussi étroitement que possible à la demande pour ces ressources.

Lorsqu'un événement de diminution de la taille des instances se produit, le groupe Auto Scaling résilie une ou plusieurs instances. Le groupe Auto Scaling utilise sa politique de mise hors service pour déterminer les instances à résilier. Les instances en cours de résiliation du groupe Auto Scaling passent en statut `Terminating`, et ne peuvent pas être remises en service.

Si votre groupe Auto Scaling est configuré pour recevoir du trafic à partir d'un équilibreur de charge Elastic Load Balancing, Amazon EC2 Auto Scaling attend que l'instance se désenregistre de l'équilibreur de charge. L'annulation de l'enregistrement de l'instance garantit que toutes les nouvelles demandes sont redirigées vers d'autres instances du groupe cible de l'équilibreur de charge, tandis que les connexions à l'instance existantes sont autorisées à se poursuivre jusqu'à l'expiration du délai de désinscription.

Si vous ajoutez un hook de cycle de vie au groupe Auto Scaling, vous pouvez réaliser une action personnalisée dans l'instance en cours de résiliation. Pour plus d'informations, consultez Hooks de cycle de vie. Enfin, l'instance est totalement résiliée et passe en statut `Terminated`.

Ce qui suit récapitule les étapes à suivre pour annuler l'enregistrement d'une instance auprès d'un équilibreur de charge pour un événement de scale-in.



Détacher une instance

Vous pouvez détacher une instance du groupe Auto Scaling. Lorsque l'instance est détachée, vous pouvez la gérer séparément du groupe Auto Scaling ou l'attacher à un groupe Auto Scaling différent.

Pour plus d'informations, consultez Détacher ou attacher des instances.

Attacher une instance

Vous pouvez attacher une instance EC2 en cours d'exécution qui répond à certains critères du groupe Auto Scaling. Lorsque l'instance est attachée, elle est gérée dans le cadre du groupe Auto Scaling.

Pour plus d'informations, consultez Détacher ou attacher des instances.

Hooks de cycle de vie

Vous pouvez ajouter un hook de cycle de vie au groupe Auto Scaling afin de pouvoir réaliser des actions personnalisées lorsque des instances sont lancées ou résiliées.

Lorsqu'Amazon EC2 Auto Scaling répond à un événement d'augmentation de la taille des instances, il lance un ou plusieurs instances. Ces instances démarrent avec l'état `Pending`. Si vous ajoutez un hook de cycle de vie `autoscaling:EC2_INSTANCE_LAUNCHING` au groupe Auto Scaling, les instances passent du statut `Pending` au statut `Pending:Wait`. Lorsque vous avez réalisé l'action du cycle de vie, les instances passent en statut `Pending:Proceed`. Lorsque les instances sont entièrement configurées, elles sont attachées au groupe Auto Scaling et passent en statut `InService`.

Lorsqu'Amazon EC2 Auto Scaling répond à un événement mise à l'échelle horizontale, il résilie une ou plusieurs instances. Ces instances sont détachées du groupe Auto Scaling et passent en statut `Terminating`. Si vous ajoutez un hook de cycle de vie `autoscaling:EC2_INSTANCE_TERMINATING` au groupe Auto Scaling, les instances passent du statut `Terminating` au statut `Terminating:Wait`. Lorsque vous avez réalisé l'action du cycle de vie, les instances passent en statut `Terminating:Proceed`. Lorsque les instances sont totalement résiliées, elles passent en statut `Terminated`.

Pour plus d’informations, consultez Hooks de cycle de vie Amazon EC2 Auto Scaling.

Entrer et sortir du mode veille

Vous pouvez mettre n'importe quelle instance se trouvant en statut `InService` en statut `Standby`. Cela vous permet de supprimer l'instance du service, de la dépanner ou d'y apporter des modifications, et de la remettre en service.

Les instances en statut `Standby` continuent d'être gérée par le groupe Auto Scaling. Cependant, elles ne représentent pas une partie active de l'application jusqu'à ce que vous les remettiez en service.

Pour plus d’informations, consultez Supprimer temporairement des instances du groupe Auto Scaling.

Quotas pour les ressources et les groupes Auto Scaling

Vous Compte AWS disposez de quotas par défaut, anciennement appelés limites, pour chaque AWS service. Sauf indication contraire, chaque quota est spécifique à la région. Vous pouvez demander des augmentations pour certains quotas, et d'autres quotas ne peuvent pas être augmentés.

Pour afficher les quotas pour Amazon EC2 Auto Scaling, ouvrez la console Service Quotas. Dans le panneau de navigation, choisissez AWS Services et sélectionnez Amazon EC2 Auto Scaling.

Pour demander une augmentation de quota, consultez Demander une augmentation de quota dans le Guide de l'utilisateur de Service Quotas. Si le quota n'est pas encore disponible dans Service Quotas, utilisez le Auto Scaling Limits form (Formulaire de limites Auto Scaling). Les augmentations de quota sont liées à la région pour laquelle elles ont été demandées.

Toutes les demandes sont soumises à AWS Support. Vous pouvez suivre votre demande dans la console AWS Support .

Ressources Amazon EC2 Auto Scaling

Vous Compte AWS disposez des quotas suivants relatifs au nombre de groupes Auto Scaling et de configurations de lancement que vous pouvez créer.

Ressource	Quota par défaut
Groupes Auto Scaling par Région	500
Lancer les configurations par région	200

Configuration du groupe Auto Scaling

Vous Compte AWS disposez des quotas suivants relatifs à la configuration des groupes Auto Scaling. Elles ne peuvent pas être modifiées.

Ressource	Quota
Politiques de mise à l'échelle par groupe Auto Scaling	50
Actions planifiées par groupe Auto Scaling	125
Ajustements par étape par politique de mise à l'échelle d'étape	20
Hooks de cycle de vie par groupe Auto Scaling	50
Rubriques SNS par groupe Auto Scaling	10

Ressource	Quota
Classic Load Balancers par groupe Auto Scaling	50
Groupes cibles Elastic Load Balancing par groupe Auto Scaling	50
Groupes cibles VPC Lattice par groupe Auto Scaling	5

Opérations de l'API du groupe Auto Scaling

Amazon EC2 Auto Scaling propose des opérations d'API pour apporter des modifications à vos groupes Auto Scaling par lots. Voici les limites de l'API sur le nombre maximum d'éléments (membres maximum du tableau) qui sont autorisés dans une seule opération. Elles ne peuvent pas être modifiées.

Opération	Nombre maximum de membres de tableau
AttachInstances	20 ID d'instance
AttachLoadÉquilibreurs	10 équilibreurs de charge
AttachLoadBalancerTargetGroupes	10 groupes cibles
BatchDeleteScheduledAction	50 actions planifiées
BatchPutScheduledUpdateGroupAction	50 actions planifiées
DetachInstances	20 ID d'instance
DetachLoadÉquilibreurs	10 équilibreurs de charge
DetachLoadBalancerTargetGroupes	10 groupes cibles
EnterStandby	20 ID d'instance
ExitStandby	20 ID d'instance
SetInstanceProtection	50 ID d'instance

Limitation des demandes pour l'API Amazon EC2 Auto Scaling

Les demandes d'API Amazon EC2 Auto Scaling sont limitées à l'aide d'un schéma de bucket à jetons afin de maintenir la bande passante du service. Pour plus d'informations, consultez le taux de demandes d'API dans le manuel Amazon EC2 Auto Scaling API Reference.

Taux de résiliation EC2

Amazon EC2 Auto Scaling détermine dynamiquement le nombre d'opérations de résiliation d'une instance EC2 qu'il peut effectuer lors de la mise à l'échelle horizontale de votre groupe Auto Scaling. Cela signifie que le nombre d'instances résiliées en même temps peut varier d'un groupe Auto Scaling à un autre. Ces variations sont dues à des considérations externes, telles que savoir si Amazon EC2 Auto Scaling doit annuler l'enregistrement des instances auprès d'un équilibreur de charge ou pas.

Autres services

Les quotas pour d'autres services, tels qu'Amazon EC2 et Amazon VPC, peuvent avoir un impact sur vos groupes Auto Scaling. Vous pouvez les utiliser Service Quotas pour mettre à jour les quotas pour les instances EC2 et les autres ressources de votre Compte AWS. Dans la Service Quotas console, vous pouvez consulter tous vos quotas de service disponibles et demander leur augmentation. Pour plus d'informations, consultez Demande d'augmentation de quotas dans le Guide de l'utilisateur Service Quotas .

Pour les quotas spécifiques aux modèles de lancement, consultez la section Restrictions relatives aux modèles de lancement dans le guide de l'utilisateur Amazon EC2.

Configuration pour utiliser Amazon EC2 Auto Scaling

Avant de commencer à utiliser Amazon EC2 Auto Scaling, exécutez les tâches suivantes.

Tâches

- Préparer l'utilisation d'Amazon EC2
- Préparez-vous à utiliser AWS CLI

Préparer l'utilisation d'Amazon EC2

Si vous n'avez jamais utilisé Amazon EC2 auparavant, exécutez les tâches décrites dans la documentation Amazon EC2. Pour plus d'informations, consultez Configuration avec Amazon EC2 dans le guide de l'utilisateur Amazon EC2 ou Configuration avec Amazon EC2 dans le guide de l'utilisateur Amazon EC2.

Préparez-vous à utiliser AWS CLI

Vous pouvez utiliser les outils de ligne de AWS commande pour émettre des commandes sur la ligne de commande de votre système afin d'exécuter Amazon EC2 Auto Scaling et d'autres AWS tâches.

Pour utiliser le AWS Command Line Interface (AWS CLI), téléchargez, installez et configurez la version 1 ou 2 du AWS CLI. La même fonctionnalité Amazon EC2 Auto Scaling est disponible dans les versions 1 et 2. Pour installer la version 1 de l' AWS CLI , consultez Installation, mise à jour et désinstallation de l' AWS CLI dans le Guide l'utilisateur AWS CLI version 1. Pour installer la AWS CLI version 2, reportez-vous à la section Installation ou mise à jour de la dernière version du AWS CLI Guide de l'utilisateur de la AWS CLI version 2.

AWS CloudShell vous permet de ne pas installer le AWS CLI dans votre environnement de développement et de l'utiliser à la AWS Management Console place. En plus d'éviter l'installation, vous n'avez pas besoin de configurer les informations d'identification et de spécifier une région. Votre AWS Management Console session fournit ce contexte au AWS CLI. Vous pouvez l'utiliser AWS CloudShell dans pris en charge Régions AWS. Pour plus d’informations, consultez Créez des groupes Auto Scaling depuis la ligne de commande en utilisant AWS CloudShell.

Pour plus d'informations, consultez update-auto-scaling-group dans le guide de référence des commandes AWS CLI .

Commencer avec Amazon EC2 Auto Scaling

Pour démarrer avec Amazon EC2 Auto Scaling, vous pouvez suivre les didacticiels qui vous présentent le service.

Rubriques

- Tutoriel : Créez votre premier groupe Auto Scaling
- Didacticiel : configurer une application redimensionnée et à charge équilibrée

Pour des didacticiels supplémentaires axés sur des outils spécifiques permettant de gérer le cycle de vie des instances dans un groupe Auto Scaling, consultez les rubriques suivantes :

- Didacticiel : configurer un hook de cycle de vie qui appelle une fonction Lambda. Ce didacticiel explique comment utiliser Amazon EventBridge pour créer des règles qui invoquent des fonctions Lambda en fonction d'événements survenant dans les instances de votre groupe Auto Scaling.
- Tutoriel : configurer les données utilisateur pour récupérer l'état du cycle de vie cible via des métadonnées d'instance. Ce didacticiel explique comment utiliser le service de métadonnées d’instance (IMDS) pour invoquer une action depuis l’instance elle-même.

Avant de créer un groupe Auto Scaling à utiliser avec votre application, étudiez minutieusement votre application lorsqu'elle fonctionne dans l' AWS Cloud. Éléments à prendre en compte :

- Nombre de zones de disponibilité que le groupe Auto Scaling doit couvrir.
- Type de ressources existantes pouvant être utilisés, comme les groupes de sécurité ou les Amazon Machine Images (AMI).
- Voulez-vous mettre à l'échelle pour augmenter ou réduire la capacité, ou voulez-vous seulement vous assurer qu'un nombre spécifique de serveurs soit toujours en cours d'exécution ? Souvenez-vous qu'Amazon EC2 Auto Scaling peut faire les deux en même temps.
- Quelles sont les métriques les plus pertinentes pour les performances de l'application.
- Le temps nécessaire au lancement et au provisionnement d'un serveur.

Mieux vous comprenez l'application, plus l'architecture Auto Scaling est efficace.

Tutoriel : Créez votre premier groupe Auto Scaling

Ce didacticiel fournit une introduction pratique à Amazon EC2 Auto Scaling via le. AWS Management Console Vous allez créer un modèle de lancement qui définit vos instances EC2 et un groupe Auto Scaling contenant une seule instance. Après avoir lancé votre groupe Auto Scaling, vous allez mettre fin à l'instance et vérifier qu'elle a été retirée du service et remplacée. Pour maintenir un nombre constant d'instances, Amazon EC2 Auto Scaling détecte et répond automatiquement aux contrôles d'intégrité et d'accessibilité d'Amazon EC2.

Lorsque vous vous inscrivez AWS, vous pouvez commencer à utiliser Amazon EC2 Auto Scaling gratuitement en utilisant le niveau AWS gratuit. Vous pouvez utiliser l'offre gratuite pour lancer et utiliser une instance `t2.micro` gratuitement pendant 12 mois (dans les régions où `t2.micro` n'est pas disponible, vous pouvez utiliser une instance `t3.micro` avec l'offre gratuite). Si vous lancez une instance qui ne fait pas partie de l'offre gratuite, les frais d'utilisation standard d'Amazon EC2 vous seront facturés pour l'instance. Pour plus d'informations, consultez Tarification Amazon EC2.

Tâches

Préparer la procédure détaillée

Cette procédure détaillée suppose que vous avez déjà lancé des instances EC2 et créé une paire de clés ainsi qu'un groupe de sécurité. Pour plus d'informations, consultez la section Configuration avec Amazon EC2 dans le guide de l'utilisateur Amazon EC2.

Pour commencer à utiliser Amazon EC2 Auto Scaling, vous pouvez utiliser le VPC par défaut pour votre. Compte AWS Le VPC par défaut inclut un sous-réseau public par défaut dans chaque zone de disponibilité et une passerelle Internet qui est attachée à votre VPC. Vous pouvez afficher vos VPC sur la page de vos VPC de la console Amazon Virtual Private Cloud (Amazon VPC).

Étape 1 : créer un modèle de lancement

Au cours de cette étape, vous créez un modèle de lancement qui spécifie le type d'instance EC2 qu'Amazon EC2 Auto Scaling crée pour vous. Indiquez les informations nécessaires, notamment l'ID d'Amazon Machine Image (AMI) à utiliser, le type d'instance, les paires de clés et les groupes de sécurité.

Pour créer un modèle de lancement

1. Ouvrez la console Amazon EC2 et accédez à la page des modèles de lancement.
2. Dans la barre de navigation en haut, sélectionnez une Région AWS. Le modèle de lancement et le groupe Auto Scaling que vous créez sont liés à la Région que vous spécifiez.
3. Choisissez Create launch template (Créer un modèle de lancement).
4. Pour Launch template name (Nom du modèle de lancement), saisissez **`my-template-for-auto-scaling`**.
5. Sous Guide Auto Scaling, activez la case à cocher.
6. Pour Application and OS Images (Amazon Machine Image) (Images d'applications et de systèmes d'exploitation [Amazon Machine Image]), choisissez une version Amazon Linux 2 (HVM) dans la liste Quick Start (Démarrage rapide). L'AMI (Amazon Machine Image) sert de modèle de configuration de base pour vos instances.
7. Pour Instance type (Type d'instance), choisissez une configuration matérielle qui soit compatible avec l'AMI que vous avez spécifiée.
8. (Facultatif) Pour Key pair (login) (Paire de clés [connexion]), choisissez une paire de clés existante. Les paires de clés servent à se connecter aux instances Amazon EC2 via SSH. La connexion à une instance n'est pas incluse dans ce didacticiel. Par conséquent, vous n'avez pas besoin de spécifier de paire de clés sauf si vous avez l'intention de vous connecter à votre instance à l'aide du protocole SSH.
9. Pour Network settings (Paramètres réseau), développez Advanced network configuration (Configuration réseau avancée) et procédez comme suit :

    a. Choisissez Add network interface (Ajouter une interface réseau) pour ajouter une interface réseau primaire.

    b. Pour Attribuer automatiquement une adresse IP publique, spécifiez si votre instance reçoit une adresse IPv4 publique. Par défaut, Amazon EC2 attribue une adresse IPv4 publique si l'instance EC2 est lancée dans un sous-réseau par défaut ou si l'instance est lancée dans

un sous-réseau configuré pour attribuer automatiquement une adresse IPv4 publique. Si vous n'avez pas besoin de vous connecter à votre instance, choisissez Disable.

c. Pour l'ID du groupe de sécurité, choisissez un groupe de sécurité dans le même VPC que vous prévoyez d'utiliser comme VPC pour votre groupe Auto Scaling. Si vous ne spécifiez pas de groupe de sécurité lorsque vous lancez une instance, celle-ci est automatiquement associée au groupe de sécurité par défaut pour le VPC.

d. Pour Supprimer à la fin, choisissez Oui pour supprimer l'interface réseau lorsque l'instance est supprimée.

10. Choisissez Create launch template (Créer un modèle de lancement).
11. Sur la page de confirmation, choisissez Create Auto Scaling group (Créer un groupe Auto Scaling).

Étape 2 : créer un groupe Auto Scaling à instance unique

Suivez la procédure ci-dessous pour continuer là où vous vous êtes arrêté après avoir créé un modèle de lancement.

Pour créer un groupe Auto Scaling

1. Dans la page Choisir un modèle de lancement ou une configuration, entrez **`my-first-asg`** comme Nom du groupe Auto Scaling.
2. Choisissez Suivant.

   La page Choisir les options de lancement d'une instance apparaît, vous permettant de choisir les paramètres réseau VPC que le groupe Auto Scaling doit utiliser et vous propose des options pour lancer des instances On-Demand et Spot.
3. Dans la section Réseau, maintenez le VPC défini sur le VPC par défaut de votre choix Région AWS, ou sélectionnez votre propre VPC. Le VPC par défaut est automatiquement configuré pour fournir une connectivité Internet à votre instance. Ce VPC inclut un sous-réseau public dans chaque zone de disponibilité de la région.
4. Pour Availability Zones and subnets (Zones de disponibilité et sous-réseaux), choisissez un sous-réseau pour chaque zone de disponibilité que vous voulez inclure. Utilisez les sous-réseaux dans plusieurs zones de disponibilité pour une haute disponibilité. Pour plus d'informations, consultez Considérations à prendre en compte lors du choix des sous-réseaux VPC.

5. Dans la section Instance type requirements (Exigences relatives au type d'instance), utilisez le paramètre par défaut pour simplifier cette étape. (Ne remplacez pas le modèle de lancement.) Pour ce didacticiel, vous lancerez une seule instance à la demande en utilisant le type d'instance spécifié dans votre modèle de lancement.
6. Conservez le reste des valeurs par défaut de ce didacticiel et choisissez Skip to review (Ignorer pour vérifier).

   > **Note**
   >
   > La taille initiale du groupe est déterminée par sa capacité désirée. La valeur par défaut est 1 instance.

7. Sur la page Review (Vérification), vérifiez les informations, puis choisissez Create Auto Scaling group (Créer un groupe Auto Scaling).

Étape 3 : vérifier votre groupe Auto Scaling

Maintenant que vous avez créé un groupe Auto Scaling, vous êtes prêt à vérifier que ce dernier a lancé une instance EC2.

> **Tip**
>
> Dans la procédure suivante, vous consultez les sections Activity history (Historique des activités), et Instances pour le groupe Auto Scaling. Dans les deux sections, les colonnes nommées doivent déjà être affichées. Pour afficher les colonnes masquées ou modifier le nombre de lignes affichées, cliquez sur l'icône en forme de roue dentée dans le coin supérieur droit de chaque section pour ouvrir les préférences modeales, mettez à jour les paramètres au besoin et cliquez sur Confirm (Confirmer).

Pour vérifier que votre groupe Auto Scaling a lancé une instance EC2

1. Ouvrez la page des groupes Auto Scaling de la console Amazon EC2.
2. Activez la case à cocher en regard du groupe Auto Scaling que vous venez de créer.

   Un volet fractionné s'ouvre en bas de la page Auto Scaling groups (Groupes Auto Scaling). Le premier onglet disponible est l'onglet Details (Détails) qui affiche des informations sur le groupe Auto Scaling.

3. Choisissez le deuxième onglet, Activity (Activité). Sous Historique des activités, vous pouvez afficher la progression des activités associées au groupe Auto Scaling. La colonne Status (État) affiche l'état actuel de votre instance. Lorsqu'une instance est en cours de lancement, son statut est `Not yet in service`. Le statut passe à `Successful`, après le lancement de l'instance. Vous pouvez également utiliser le bouton d'actualisation pour consulter le statut actuel de l'instance.
4. Sous l'onglet Instance management (Gestion des instances), sous Instances, vous pouvez afficher le statut de l'instance.
5. Vérifiez que votre instance a été lancée correctement. Il suffit de peu de temps pour lancer une instance.

   - La colonne Lifecycle (Cycle de vie) affiche l'état de votre instance. Initialement, votre instance est à l'état `Pending`. Lorsqu'une instance est prête à recevoir du trafic, son statut passe à `InService`.
   - La colonne État de santé affiche le résultat des tests de santé effectués par Amazon EC2 Auto Scaling sur votre instance.

Étape 4 : résilier une instance de votre groupe Auto Scaling

Ces étapes permettent d'en savoir plus sur la façon dont Amazon EC2 Auto Scaling fonctionne, et en particulier sur la façon dont il lance de nouvelles instances, le cas échéant. La taille minimale du groupe Auto Scaling que vous avez créé dans ce didacticiel est de une instance. Par conséquent, si vous mettez fin à l'instance en cours d'exécution, Amazon EC2 Auto Scaling doit lancer une nouvelle instance pour la remplacer.

1. Ouvrez la page des groupes Auto Scaling de la console Amazon EC2.
2. Cochez la case située en regard de votre groupe Auto Scaling.
3. Dans l'onglet Instance management (Gestion des instances) sous Instances, sélectionnez l'ID de l'instance.

   Cela vous amène à la page Instances de la console Amazon EC2, où vous pouvez résilier l'instance.
4. Choisissez Actions, Instance State (État de l'instance), Terminate (Résilier). Lorsque vous êtes invité à confirmer, choisissez Yes, Terminate (Oui, résilier).
5. Dans le volet de navigation, sous Auto Scaling, choisissez Auto Scaling Groups (Groupes Auto Scaling). Sélectionnez votre groupe Auto Scaling, puis choisissez l'onglet Activity (Activité).

Lorsque vous mettez fin à une instance à partir de la page Instances, il faut une minute ou deux après la fin de l'instance pour qu'une nouvelle instance soit lancée. Dans l'historique d'activité, lorsque la mise à l'échelle démarre, vous observez une entrée pour la résiliation de la première instance et une autre pour le lancement d'une nouvelle instance. Utilisez le bouton d'actualisation jusqu'à ce que les nouvelles entrées apparaissent.

6. Dans l'onglet Instance management (Gestion des instances), la section Instances affiche uniquement la nouvelle instance.
7. Dans le panneau de navigation, sous Instances, choisissez Instances. Cette page affiche l'instance mise hors service et celle en cours d'exécution.

Étape 5 : étapes suivantes

Passez à l'étape suivante si vous souhaitez supprimer l'infrastructure de base que vous venez de créer. Sinon, vous pouvez utiliser cette infrastructure comme base et essayer une ou plusieurs des actions suivantes :

- Se connecter à votre instance Linux à l'aide du Gestionnaire de session ou SSH Pour plus d'informations, consultez les sections Connexion à votre instance Linux à l'aide du gestionnaire de session et Connexion à votre instance Linux depuis Linux ou macOS via SSH dans le guide de l'utilisateur Amazon EC2.
- Configurez une notification Amazon SNS pour vous avertir chaque fois que votre groupe Auto Scaling lance ou résilier des instances. Pour plus d'informations, consultez Options de notification Amazon SNS.
- Mettez manuellement à l'échelle la capacité de votre groupe Auto Scaling pour tester la notification SNS. Pour plus d'informations, consultez Changer la capacité souhaitée de votre groupe Auto Scaling.

Vous pouvez également commencer à vous familiariser avec les concepts de mise à l'échelle automatique en consultant Politiques de suivi des objectifs de la mise à l'échelle. Si la charge de votre application change, votre groupe Auto Scaling peut augmenter (ajouter des instances) et mettre des instances à l'échelle horizontale (exécuter moins d'instances) automatiquement en ajustant la capacité souhaitée du groupe entre les limites de capacité minimale et maximale. Pour plus d'informations sur le paramétrage de ces limites, consultez Définissez des limites de mise à l'échelle pour votre groupe Auto Scaling.

Étape 6 : Nettoyer

Vous pouvez soit supprimer votre infrastructure de dimensionnement, soit supprimer uniquement votre groupe Auto Scaling et conserver votre modèle de lancement pour une utilisation ultérieure.

Si vous avez lancé une instance qui ne fait pas partie de l'offre gratuite AWS, vous devez mettre fin à votre instance pour éviter d'avoir à payer des frais supplémentaires. Lorsque vous résiliez l'instance, les données qui y sont associées sont également supprimées.

Pour supprimer votre groupe Auto Scaling

1. Ouvrez la page des groupes Auto Scaling de la console Amazon EC2.
2. Cochez la case située en regard de votre groupe Auto Scaling (`my-first-asg`).
3. Sélectionnez Delete (Supprimer).
4. Lorsque vous êtes invité à confirmer l'opération, saisissez **`delete`** pour confirmer la suppression du groupe Auto Scaling spécifié, puis choisissez Delete (Supprimer).

   Une icône de chargement dans la colonne Name (Nom) indique que le groupe Auto Scaling est en cours de suppression. Lorsque la suppression s'est produite, les colonnes Desired (Souhaité), Min et Max affichent `0` instances du groupe Auto Scaling. Quelques minutes sont nécessaires pour résilier l'instance et supprimer le groupe. Actualisez la liste pour afficher l'état actuel.

Passez cette procédure si vous souhaitez conserver le modèle de lancement.

Pour supprimer votre modèle de lancement

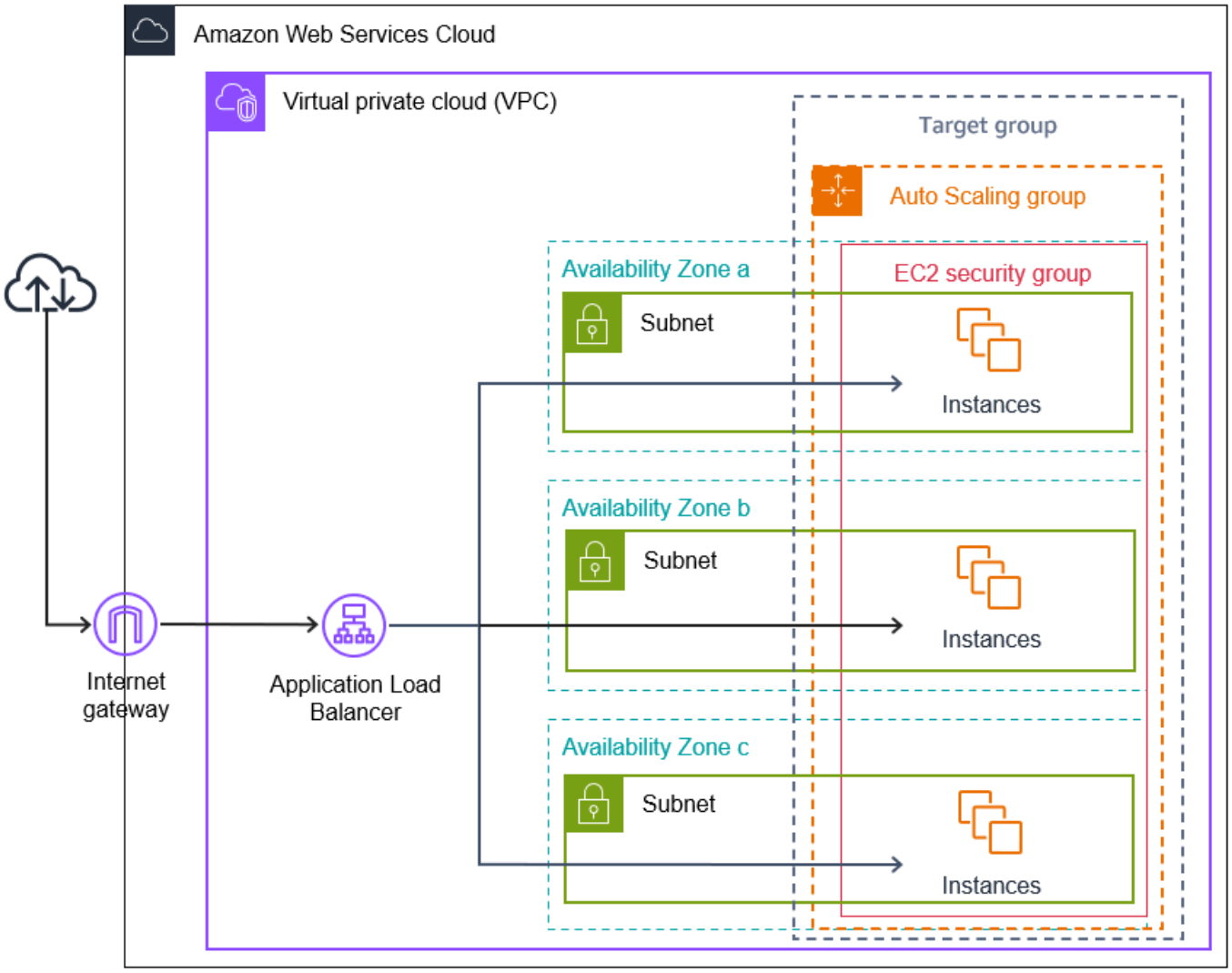
1. Ouvrez la page des modèles de lancement de la console Amazon EC2.
2. Sélectionnez votre modèle de lancement (`my-template-for-auto-scaling`)
3. Choisissez Actions, puis Supprimer le modèle.
4. Lorsque vous êtes invité à confirmer l'opération, saisissez **`Delete`** pour confirmer la suppression du modèle de lancement spécifié, puis choisissez Delete (Supprimer).

Didacticiel : configurer une application redimensionnée et à charge équilibrée

> **Important**
>
> Avant d'explorer ce didacticiel, nous vous recommandons de consulter d'abord le didacticiel d'introduction suivant : Créez votre premier groupe Auto Scaling.

L'enregistrement de votre groupe Auto Scaling avec un équilibreur de charge Elastic Load Balancing vous aide à configurer une application à charge équilibrée. Elastic Load Balancing fonctionne avec Amazon EC2 Auto Scaling pour répartir le trafic entrant sur vos instances Amazon EC2 saines. Cela augmente l'évolutivité et la disponibilité de votre application. Vous pouvez activer Elastic Load Balancing dans plusieurs zones de disponibilité pour augmenter la tolérance aux pannes de vos applications.

Dans ce didacticiel, nous couvrons les étapes de base pour la configuration d'une application à charge équilibrée lors de la création du groupe Auto Scaling. Une fois que vous avez terminé, votre architecture doit ressembler au schéma suivant :



Elastic Load Balancing prend en charge différents types d'équilibreurs de charge. Nous vous recommandons d'utiliser un Application Load Balancer pour ce didacticiel.

Pour plus d'informations sur l'introduction d'un équilibreur de charge dans votre architecture, consultez Utiliser Elastic Load Balancing pour répartir le trafic sur les instances dans votre groupe Auto Scaling..

Tâches

- Prérequis
- Étape 1 : configurer un modèle de lancement ou d'une configuration de lancement
- Étape 2 : créer un groupe Auto Scaling
- Étape 3 : vérifier que votre équilibreur de charge est attaché
- Étape 4 : étapes suivantes

Prérequis

- Un équilibreur de charge et un groupe cible. Assurez-vous de choisir les mêmes zones de disponibilité pour l'équilibreur de charge que celles que vous prévoyez d'utiliser pour votre groupe Auto Scaling. Pour plus d'informations, consultez Prise en main d'Elastic Load Balancing dans le Guide de l'utilisateur Elastic Load Balancing.
- Un groupe de sécurité pour votre modèle de lancement ou votre configuration du lancement. Le groupe de sécurité doit autoriser l'accès à partir de l'équilibreur de charge sur le port de l'écouteur (généralement le port 80 pour le trafic HTTP) et le port que vous souhaitez que Elastic Load Balancing utilise pour effectuer des surveillances de l'état. Pour plus d'informations, consultez la documentation pertinente :
  - Groupes de sécurité cibles dans le Guide de l'utilisateur des Application Load Balancers
  - Groupes de sécurité cibles dans le Guide de l'utilisateur des Network Load Balancers

  Le cas échéant, si vos instances doivent avoir des adresses IP publiques, vous pouvez autoriser le trafic SSH pour la connexion aux instances.
- (Facultatif) Rôle IAM qui accorde à votre application l'accès à AWS.
- (Facultatif) Une Amazon Machine Image (AMI) définie en tant que modèle source pour vos instances Amazon EC2. Pour en créer une maintenant, lancez une instance. Spécifiez le rôle IAM (si vous en avez créé un) ainsi que les scripts de configuration dont vous avez besoin comme données utilisateur. Connectez-vous à l'instance et personnalisez-la. Par exemple, vous pouvez procéder à l'installation des logiciels et des applications, à la copie des données et à l'attachement des volumes EBS supplémentaires. Testez vos applications sur votre instance pour vous assurer qu'elle est correctement configurée. Enregistrez cette configuration mise à jour en tant qu'AMI personnalisée. Vous pouvez résilier l'instance si vous n'en avez pas besoin ultérieurement. Les instances lancées à partir de cette nouvelle AMI incluront les personnalisations apportées lors de sa création.
- Un Virtual Private Cloud (VPC). Ce didacticiel fait référence au VPC par défaut, mais vous pouvez utiliser le vôtre. Si vous utilisez votre propre VPC, assurez-vous qu'il dispose d'un sous-réseau mappé à chaque zone de disponibilité de la région dans laquelle vous travaillez. Au minimum, vous devez disposer de deux sous-réseaux publics disponibles pour créer l'équilibreur de charge. Vous

devez également disposer de deux sous-réseaux privés ou deux sous-réseaux publics pour créer votre groupe Auto Scaling et l'enregistrer auprès de l'équilibreur de charge.

Étape 1 : configurer un modèle de lancement ou d'une configuration de lancement

Utilisez un modèle de lancement ou une configuration de lancement pour ce didacticiel.

Rubriques

Sélectionnez ou créez un modèle de lancement

Si vous possédez déjà un modèle de lancement que vous souhaiteriez utiliser, sélectionnez-le grâce à la procédure suivante.

Pour sélectionner un modèle de lancement existant

1. Ouvrez la page des modèles de lancement de la console Amazon EC2.
2. Dans la barre de navigation située en haut de l'écran, choisissez la région dans laquelle l'équilibreur de charge a été créé.
3. Sélectionnez un modèle de lancement.
4. Choisissez Actions, Create Auto Scaling group (Créer un groupe Auto Scaling).

Sinon, pour créer un nouveau modèle de lancement, utilisez la procédure suivante.

Pour créer un modèle de lancement

1. Ouvrez la page des modèles de lancement de la console Amazon EC2.
2. Dans la barre de navigation située en haut de l'écran, choisissez la région dans laquelle l'équilibreur de charge a été créé.
3. Choisissez Créer un modèle de lancement.
4. Saisissez un nom et une description pour la version initiale du modèle de lancement.

5. Pour Application and OS Images (Amazon Machine Image) (Images d'applications et de systèmes d'exploitation [Amazon Machine Image]), sélectionnez l'ID de l'AMI pour vos instances. Vous pouvez effectuer une recherche parmi toutes les AMI disponibles ou sélectionner une AMI depuis la liste Recent (Récent) ou Quick Start (Démarrage rapide). Si vous ne voyez pas l'AMI dont vous avez besoin, choisissez Browser more AMIs (Parcourir plus d'AMI) pour parcourir le catalogue complet des AMI.
6. Pour Instance type (Type d'instance), sélectionnez une configuration matérielle pour vos instances qui soit compatible avec l'AMI que vous avez spécifiée.
7. (Facultatif) Pour Key pair (login) (Paire de clés [connexion]), choisissez la paire de clés à utiliser lors de la connexion à vos instances.
8. Pour Network settings (Paramètres réseau), développez Advanced network configuration (Configuration réseau avancée) et procédez comme suit :

    a. Choisissez Add network interface (Ajouter une interface réseau) pour ajouter une interface réseau primaire.

    b. Pour Attribuer automatiquement une adresse IP publique, spécifiez si vos instances reçoivent des adresses IPv4 publiques. Par défaut, Amazon EC2 attribue une adresse IPv4 publique si l'instance EC2 est lancée dans un sous-réseau par défaut ou si l'instance est lancée dans un sous-réseau configuré pour attribuer automatiquement une adresse IPv4 publique. Si vous n'avez pas besoin de vous connecter à vos instances, vous pouvez choisir Désactiver pour empêcher les instances de votre groupe de recevoir du trafic directement depuis Internet. Dans ce cas, elles recevront le trafic uniquement de l'équilibreur de charge.

    c. Pour Security group ID (ID du groupe de sécurité), spécifiez un groupe de sécurité pour vos instances à partir du même VPC que l'équilibreur de charge.

    d. Pour Delete on termination (Supprimer à la résiliation), choisissez Yes. Cela supprime l'interface réseau lorsque le groupe Auto Scaling est mis à l'échelle et lorsque l'instance à laquelle l'interface réseau est attachée est résiliée.
9. (Facultatif) Pour distribuer en toute sécurité les informations d'identification à vos instances, pour Advanced details (Détails avancés), IAM instance profile (Profil d'instance IAM), saisissez l'Amazon Resource Name (ARN) de votre rôle IAM.
10. (Facultatif) Pour spécifier des données utilisateur ou un script de configuration pour vos instances, collez-les dans Advanced details (Détails avancés), User data (Données utilisateur).
11. Choisissez Create launch template (Créer un modèle de lancement).
12. Sur la page de confirmation, choisissez Create Auto Scaling group (Créer un groupe Auto Scaling).

Sélectionner ou créer une configuration de lancement

> **Note**
>
> Nous vous déconseillons vivement d'utiliser des configurations de lancement dans les nouvelles applications, car il s'agit d'une fonctionnalité héritée pour laquelle aucun investissement n'est prévu. En outre, les nouveaux comptes créés le 1er juin 2023 ou après cette date n'auront pas la possibilité de créer de nouvelles configurations de lancement via la console. Pour plus d'informations, consultez Configurations de lancement.

Pour sélectionner une configuration de lancement existante

1. Ouvrez la page des configurations de lancement de la console Amazon EC2.
2. Dans la barre de navigation en haut de l'écran, choisissez la région dans laquelle l'équilibreur de charge a été créé.
3. Sélectionnez une configuration de lancement.
4. Choisissez Actions, Create Auto Scaling group (Créer un groupe Auto Scaling).

Sinon, pour créer une nouvelle configuration de lancement, utilisez la procédure suivante.

Pour créer une configuration du lancement

1. Ouvrez la page des configurations de lancement de la console Amazon EC2. Lorsque vous êtes invité à confirmer, choisissez Afficher les configurations de lancement pour confirmer que vous souhaitez consulter la page Configurations de lancement.
2. Dans la barre de navigation en haut de l'écran, choisissez la région dans laquelle l'équilibreur de charge a été créé.
3. Choisissez Create launch configuration (Créer une configuration de lancement) et entrez un nom pour votre configuration.
4. Pour Amazon machine Image (AMI), saisissez l'ID de l'AMI pour vos instances en tant que critères de recherche.
5. Pour Instance type (Type d'instance), sélectionnez une configuration matérielle pour l'instance.
6. Sous Additional configuration (Configuration supplémentaire), prêtez attention aux champs suivants :

a. (Facultatif) Pour distribuer de manière sécurisée les informations d'identification à votre instance EC2, pour IAM instance profile (Profil d'instance IAM), sélectionnez votre rôle IAM. Pour plus d'informations, consultez Rôle IAM pour les applications qui s'exécutent sur des instances Amazon EC2.

b. (Facultatif) Pour spécifier des données utilisateur ou un script de configuration pour votre instance, collez-les dans Détails avancés, Données utilisateur.

c. (Facultatif) Pour Détails avancés, Type d'adresse IP, conservez la valeur par défaut. Lorsque vous créez votre groupe Auto Scaling, vous pouvez attribuer une adresse IP publique aux instances de votre groupe Auto Scaling en utilisant des sous-réseaux dont l'attribut d'adressage IP public est activé, tels que les sous-réseaux par défaut dans le VPC par défaut. Sinon, si vous n'avez pas besoin de vous connecter à vos instances, vous pouvez choisir N'affecter une adresse IP publique à aucune instance afin d'empêcher les instances de votre groupe de recevoir du trafic directement à partir d'Internet. Dans ce cas, elles recevront le trafic uniquement de l'équilibreur de charge.

7. Pour Security groups (Groupes de sécurité), choisissez un groupe de sécurité existant dans le même VPC que l'équilibreur de charge. Si vous ne désélectionnez pas Create a new security group (Créer un groupe de sécurité), une règle SSH par défaut est configurée pour les instances Amazon EC2 s'exécutant sur les systèmes d'exploitation Linux. Une règle RDP par défaut est configurée pour les instances Amazon EC2 s'exécutant sous Windows.

8. Pour Key pair (login) (Paire de clés [connexion]), choisissez une option sous Key pair options (Options de la paire de clés).

   Si vous avez déjà configuré une paire de clés d'instance Amazon EC2, vous pouvez la choisir ici.

   Si vous ne disposez pas déjà d'une paire de clés d'instance Amazon EC2, choisissez Create a new key pair (Créer une nouvelle paire de clés) et attribuez-lui un nom facilement identifiable. Choisissez Download key pair (Télécharger une paire de clés) pour télécharger la paire de clés sur votre ordinateur.

   > **Important**
   >
   > Ne choisissez pas Proceed without a key pair (Continuer sans paire de clés) si vous avez besoin de vous connecter aux instances.

9. Sélectionnez la case à cocher de confirmation, puis choisissez Create launch configuration (Créer une configuration de lancement).

10. Activez la case à cocher en regard du nom de votre nouvelle configuration du lancement et choisissez Actions, Create Auto Scaling group (Créer un groupe Auto Scaling).

Étape 2 : créer un groupe Auto Scaling

Utilisez la procédure suivante pour reprendre là où vous en étiez après avoir créé ou sélectionné votre modèle de lancement ou votre configuration de lancement.

Pour créer un groupe Auto Scaling

1. Dans la page Choisir un modèle de lancement ou une configuration, dans Auto Scaling group name (Nom du groupe Auto Scaling), entrez un nom pour le groupe Auto Scaling.
2. [Modèle de lancement uniquement] Pour Launch template (Modèle de lancement), indiquez si le groupe Auto Scaling utilise la version par défaut, la version la plus récente ou une version spécifique du modèle de lancement lors de l'évolutivité horizontale.
3. Choisissez Suivant.

   La page Choisir les options de lancement de l'instance (Choisir les options de lancement d'instance) s'affiche, vous permettant de choisir les paramètres de réseau VPC que vous voulez que le groupe Auto Scaling utilise et vous donnant des options pour le lancement des instances à la demande et Spot (si vous avez choisi un modèle de lancement).
4. Dans la section Network (Réseau), pour VPC, choisissez le VPC que vous avez utilisé pour votre équilibreur de charge. Si vous choisissez le VPC par défaut, il est automatiquement configuré pour fournir une connectivité Internet à vos instances. Ce VPC inclut un sous-réseau public dans chaque zone de disponibilité de la région.
5. Pour Availability Zones and subnets (Zones de disponibilité et sous-réseaux), choisissez un ou plusieurs sous-réseaux dans chaque zone de disponibilité que vous souhaitez inclure, en fonction des zones de disponibilité dans lesquelles se trouve l'équilibreur de charge. Pour plus d'informations, consultez Considérations à prendre en compte lors du choix des sous-réseaux VPC.
6. [Modèle de lancement uniquement] Dans la section Exigences relatives au type d'instance, utilisez le paramètre par défaut pour simplifier cette étape. (Ne remplacez pas le modèle de lancement.) Pour ce didacticiel, vous lancerez uniquement des instances à la demande en utilisant le type d'instance spécifié dans votre modèle de lancement.
7. Choisissez Next (Suivant) pour accéder à la page Configurer les options avancées.

8. Pour attacher le groupe à un équilibreur de charge existant, dans la section Répartition de charge, choisissez Attach to an existing load balancer (Attacher à un équilibreur de charge existant). Vous pouvez choisir Choose from your load balancer target groups (Choisir parmi les groupes cibles de votre équilibreur de charge) ou Choose from Classic Load Balancers (Choisir parmi les Classic Load Balancers). Vous pouvez ensuite choisir le nom d'un groupe cible pour l'Application Load Balancer ou le Network Load Balancer que vous avez créé ou choisir le nom d'un Classic Load Balancer.
9. (Facultatif) Pour utiliser les surveillances de l'état Elastic Load Balancing, pour Health checks (Surveillances de l'état), choisissez ELB sous Health check type (Type de surveillance de l'état).
10. Lorsque vous avez terminé la configuration du groupe Auto Scaling, choisissez Skip to review (Ignorer pour vérification).
11. Sur la page Review (Vérifier), passez en revue les détails de votre groupe Auto Scaling. Vous pouvez choisir Edit (Modifier) pour effectuer des changements. Lorsque vous avez terminé, choisissez Create Auto Scaling group (Créer un groupe Auto Scaling).

Après avoir créé le groupe Auto Scaling avec l'équilibreur de charge attaché, l'équilibreur de charge enregistre automatiquement les nouvelles instances au fur et à mesure qu'elles sont en ligne. À ce stade, vous n'avez qu'une seule instance, il n'y a donc pas grand-chose à enregistrer. Toutefois, vous pouvez ajouter des instances supplémentaires en mettant à jour la capacité souhaitée du groupe. Pour step-by-step obtenir des instructions, voirChanger la capacité souhaitée de votre groupe Auto Scaling.

Étape 3 : vérifier que votre équilibreur de charge est attaché

Pour vérifier que votre équilibreur de charge est attaché

1. Dans la page des groupes Auto Scaling de la console Amazon EC2, cochez la case située en regard de votre groupe Auto Scaling.
2. Dans l'onglet Details (Détails), Load balancing (Répartition de charge) affiche les groupes cibles d'équilibrage de charge attachés ou les Classic Load Balancers.
3. Dans l'onglet Activity, au niveau d'Activity history (Historique de l'activité), vous pouvez vérifier que vos instances ont été lancées correctement. La colonne Status indique si le groupe Auto Scaling a réussi le lancement des instances. Si vos instances ne parviennent pas à se lancer, vous trouverez des idées de dépannage pour des problèmes courants de lancement d'instance dans Résoudre les problèmes d'Amazon EC2 Auto Scaling.

4. Dans l'onglet Instance management (Gestion des instances) sous Instances, vous pouvez vérifier que vos instances sont prêtes à recevoir le trafic. Initialement, vos instances sont à l'état `Pending`. Lorsqu'une instance est prête à recevoir du trafic, son statut passe à `InService`. La colonne Health Status (État de santé) affiche le résultat des surveillances de l'état Amazon EC2 Auto Scaling des instances. Bien qu'une instance puisse être marquée comme saine, l'équilibreur de charge n'envoie le trafic qu'aux instances qui passent les surveillances d'état de l'équilibreur de charge.
5. Vérifiez que vos instances sont enregistrées auprès de l'équilibreur de charge. Ouvrez la page des groupes cibles de la console Amazon EC2. Sélectionnez votre groupe cible, puis cliquez sur l'onglet Targets (Cibles). Si l'état de vos instances est `initial`, c'est probablement parce qu'ils sont encore en train d'être enregistrés ou qu'ils subissent encore des surveillances de l'état. Lorsque l'état de vos instances indique `healthy`, elles sont prêtes à être utilisées.

Étape 4 : étapes suivantes

Maintenant que vous avez terminé ce didacticiel, vous pouvez en savoir plus :

- Amazon EC2 Auto Scaling détermine si une instance est saine en fonction du statut des surveillances de l'état que votre groupe Auto Scaling utilise. Si vous activez les contrôles de santé de l'équilibreur de charge et qu'une instance échoue aux tests de santé, votre groupe Auto Scaling considère que l'instance est défectueuse et la remplace. Pour plus d'informations, consultez Surveillance de l'état.
- Vous pouvez étendre votre application à une zone de disponibilité supplémentaire dans la même région afin d'augmenter la tolérance aux pannes en cas d'interruption de service. Pour plus d'informations, consultez Ajouter de zones de disponibilité.
- Vous pouvez configurer votre groupe Auto Scaling pour qu'il utilise une politique de suivi des objectifs et d'échelonnement. Cela augmente ou diminue automatiquement le nombre d'instances à mesure que la demande sur vos instances change. Cela permet au groupe de gérer les modifications de la quantité de trafic que votre application reçoit. Pour plus d'informations, consultez Politiques de suivi des objectifs de la mise à l'échelle.

Étape 5 : nettoyer

Une fois que vous avez fini avec les ressources que vous avez créées dans le cadre de ce didacticiel, vous devez les nettoyer pour éviter des frais inutiles.

Pour supprimer votre groupe Auto Scaling

1. Ouvrez la page des groupes Auto Scaling de la console Amazon EC2.
2. Cochez la case située en regard de votre groupe Auto Scaling.
3. Sélectionnez Delete (Supprimer).
4. Lorsque vous êtes invité à confirmer l'opération, saisissez **`delete`** pour confirmer la suppression du groupe Auto Scaling spécifié, puis choisissez Delete (Supprimer).

   Une icône de chargement dans la colonne Name (Nom) indique que le groupe Auto Scaling est en cours de suppression. Lorsque la suppression s'est produite, les colonnes Desired (Souhaité), Min et Max affichent `0` instances du groupe Auto Scaling. Quelques minutes sont nécessaires pour résilier l'instance et supprimer le groupe. Actualisez la liste pour afficher l'état actuel.

Passez cette procédure si vous souhaitez conserver le modèle de lancement.

Pour supprimer votre modèle de lancement

1. Ouvrez la page des modèles de lancement de la console Amazon EC2.
2. Sélectionnez votre modèle de lancement.
3. Choisissez Actions, puis Delete template (Supprimer le modèle).
4. Lorsque vous êtes invité à confirmer l'opération, saisissez **`Delete`** pour confirmer la suppression du modèle de lancement spécifié, puis choisissez Delete (Supprimer).

Passez cette procédure si vous souhaitez conserver la configuration du lancement.

Pour supprimer la configuration du lancement

1. Ouvrez la page des configurations de lancement de la console Amazon EC2.
2. Sélectionnez votre configuration de lancement.
3. Choisissez Actions, Delete launch configuration (Supprimer la configuration du lancement).
4. Lorsque vous êtes invité à confirmer l’opération, choisissez Supprimer.

Ignorez la procédure suivante si vous souhaitez conserver l'équilibreur de charge pour une utilisation ultérieure.

Pour supprimer l'équilibreur de charge

1. Ouvrez la page des équilibreurs de charge de la console Amazon EC2.
2. Sélectionnez l'équilibreur de charge et choisissez Actions, Delete (Supprimer).
3. Lorsque vous êtes invité à confirmer l'opération, choisissez Oui, supprimer.

Pour supprimer votre groupe cible

1. Ouvrez la page des groupes cibles de la console Amazon EC2.
2. Sélectionnez le groupe cible et choisissez Actions, Delete (Supprimer).
3. Lorsque vous êtes invité à confirmer l’opération, choisissez Yes, Delete.

Ressources connexes

Vous pouvez ainsi créer et provisionner des déploiements d' AWS infrastructure de manière prévisible et répétée, en utilisant des fichiers modèles pour créer et supprimer un ensemble de ressources en une seule unité (une pile). AWS CloudFormation Pour plus d’informations, consultez le Guide de l’utilisateur AWS CloudFormation.

Pour une démonstration vous expliquant comment utiliser un modèle de pile pour alimenter un groupe Auto Scaling et Application Load Balancer, consultez la section Procédure : Création d’une application redimensionnée et équilibrée de charge dans le Guide de l’utilisateur AWS CloudFormation . Utilisez la démonstration et le modèle en exemple comme point de départ pour créer des modèles similaires répondant à vos besoins.

Modèles de lancement Amazon EC2 Auto Scaling

Un modèle de lancement est semblable à une configuration de lancement, en ce sens qu'il indique les informations sur la configuration de l'instance. Il comprend l'ID de l'Amazon Machine Image (AMI), le type d'instance, une paire de clés, les groupes de sécurité et les autres paramètres utilisés pour lancer des instances EC2. Toutefois, la définition d'un modèle de lancement plutôt qu'une configuration de lancement permet de disposer de plusieurs versions d'un modèle de lancement.

Avec la gestion des versions des modèles de lancement, vous pouvez créer un sous-ensemble de l'ensemble complet de paramètres. Ensuite, vous pouvez le réutiliser pour créer d'autres versions du même modèle de lancement. Par exemple, vous pouvez créer un modèle de lancement qui définit une configuration de base sans AMI ou script de données utilisateur. Après avoir créé votre modèle de lancement, vous pouvez créer une nouvelle version et ajouter l'AMI et les données utilisateur contenant la dernière version de votre application pour tester. Cela donne deux versions du modèle de lancement. Le stockage d'une configuration de base vous aide à maintenir les paramètres généraux de configuration requis. Vous pouvez créer une nouvelle version de votre modèle de lancement à partir de la configuration de base quand vous le souhaitez. Vous pouvez également supprimer les versions utilisées pour tester votre application lorsque vous n'en avez plus besoin.

Nous vous recommandons d'utiliser des modèles de lancement pour vous assurer que vous accédez aux fonctions et améliorations les plus récentes. Toutes les fonctionnalités Amazon EC2 Auto Scaling ne sont pas disponibles lorsque vous utilisez des configurations de lancement. Par exemple, vous ne pouvez pas créer un groupe Auto Scaling qui lance à la fois des instances Spot et des instances à la demande, ou qui spécifie plusieurs types d'instance. Vous devez utiliser un modèle de lancement pour configurer ces fonctions. Pour plus d’informations, consultez Groupes Auto Scaling combinant plusieurs types d'instances et options d'achat.

Avec les modèles de lancement, vous pouvez également utiliser des fonctions plus récentes d'Amazon EC2. Cela inclut les paramètres de Systems Manager (AMI ID), la génération actuelle de volumes EBS Provisioned IOPS (io2), le balisage des volumes EBS, les instances T2 Unlimited, les réservations de capacité et les hôtes dédiésCapacity Blocks, pour n'en citer que quelques-uns.

Lorsque vous créez un modèle de lancement, tous les paramètres sont facultatifs. Toutefois, si un modèle de lancement ne spécifie pas d'AMI, vous ne pouvez pas ajouter l'AMI lorsque vous créez votre groupe Auto Scaling. Si vous spécifiez une AMI mais aucun type d'instance, vous pouvez ajouter un ou plusieurs types d'instance lorsque vous créez votre groupe Auto Scaling.

Table des matières

Autorisations d'utilisation des modèles de lancement

Les procédures de cette section supposent que vous disposez déjà des autorisations nécessaires pour créer des modèles de lancement. Pour plus d'informations sur la manière dont un administrateur vous accorde des autorisations, consultez la section Contrôler l'accès pour lancer des modèles avec des autorisations IAM dans le guide de l'utilisateur Amazon EC2.

Notez que si vous ne disposez pas des autorisations suffisantes pour utiliser et créer les ressources spécifiées dans un modèle de lancement, vous recevez un message d’erreur indiquant que vous n’êtes pas autorisé à utiliser le modèle de lancement lorsque vous essayez de le spécifier pour un groupe Auto Scaling. Pour plus d’informations, consultez Résoudre les problèmes d'Amazon EC2 Auto Scaling : modèles de lancement.

Pour des exemples de politiques IAM qui vous permettent d'appeler les opérations `CreateAutoScalingGroup`, `UpdateAutoScalingGroup`, et d'`RunInstances`API à l'aide d'un modèle de lancement, consultezSupport de modèle de lancement.

Opérations API prises en charge par les modèles de lancement

Pour obtenir la liste des opérations d’API prises en charge par les modèles de lancement, consultez les actions Amazon EC2 dans la Référence d’API Amazon EC2.

Créer un modèle de lancement pour un groupe Auto Scaling

Avant de pouvoir créer un groupe Auto Scaling à l'aide d'un modèle de lancement, vous devez créer un modèle de lancement contenant les informations de configuration pour lancer une instance, comme l'ID Amazon Machine Image (AMI).

Utilisez la procédure suivante pour créer un modèle de lancement :

Table des matières

> **Important**
>
> Les paramètres du modèle de lancement ne sont pas entièrement validés lorsque vous créez le modèle de lancement. Si vous spécifiez des valeurs incorrectes pour les paramètres ou si vous n'utilisez pas de combinaisons de paramètres prises en charge, aucune instance ne peut se lancer à l'aide de ce modèle de lancement. Veillez à spécifier les valeurs correctes pour les paramètres et à utiliser les combinaisons de paramètres prises en charge. Par exemple, pour lancer des instances avec une AMI AWS Graviton ou Graviton2 basée sur Arm, vous devez spécifier un type d'instance compatible avec Arm. Pour plus d'informations, consultez la section Restrictions relatives aux modèles de lancement dans le guide de l'utilisateur Amazon EC2.

Créer votre modèle de lancement (console)

Les étapes suivantes décrivent comment configurer un modèle de lancement de base :

- Spécifiez l'Amazon Machine Image (AMI) utilisée pour le lancement des instances.
- Choisissez un type d'instance qui est compatible avec l'AMI que vous spécifiez.

- Spécifiez la paire de clés à utiliser lors de la connexion à des instances, par exemple, à l'aide de SSH.
- Ajoutez un ou plusieurs groupes de sécurité pour autoriser l'accès réseau aux instances.
- Indiquez si vous souhaitez attacher des volumes supplémentaires à chaque instance.
- Ajoutez des balises personnalisées (paires clé-valeur) pour les instances et les volumes.

Pour créer un modèle de lancement

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Dans le volet de navigation, sous Instances, choisissez Modèles de lancement.
3. Choisissez Create launch template (Créer un modèle de lancement). Saisissez un nom et une description pour la version initiale du modèle de lancement.
4. (Facultatif) Sous Conseils Auto Scaling, cochez la case pour obtenir des conseils d'Amazon EC2 pour créer un modèle à utiliser avec Amazon EC2 Auto Scaling.
5. Sous Launch template contents (Contenu du modèle de lancement), remplissez chaque champ requis et tous les champs facultatifs requis.

   a. Application and OS Images (Amazon Machine Image) (Images d'applications et de systèmes d'exploitation [Amazon Machine Image]) : choisissez l'ID de l'AMI pour vos instances. Vous pouvez effectuer une recherche parmi toutes les AMI disponibles ou sélectionner une AMI depuis la liste Recent (Récent) ou Quick Start (Démarrage rapide). Si vous ne voyez pas l'AMI dont vous avez besoin, choisissez Browser more AMIs (Parcourir plus d'AMI) pour parcourir le catalogue complet des AMI.

      Pour choisir une AMI personnalisée, vous devez d'abord créer votre AMI à partir d'une instance personnalisée. Pour plus d'informations, consultez la section Création d'une AMI dans le guide de l'utilisateur Amazon EC2.

   b. Pour Instance type (Type d'instance), choisissez un type d'instance unique compatible avec l'AMI que vous avez spécifié.

      Autrement, pour utiliser la sélection du type d'instance basée sur des attributs, choisissez Avancé, Spécifier les attributs de type d'instance, puis spécifiez les options suivantes :

      - Number of vCPUs (Nombre de vCPU) : saisissez le nombre minimum et maximum de vCPUs. Pour n'indiquer aucune limite, saisissez un minimum de 0 et laissez le champ maximum vide.

- Amount of memory (MiB) (Quantité de mémoire [MiB]) : saisissez la quantité minimale et maximale de mémoire, en MiB. Pour n'indiquer aucune limite, saisissez un minimum de 0 et laissez le champ maximum vide.
- Développez Optional instance type attributes (Attributs de type d'instance en option) et choisissez Add attribute (Ajouter un attribut) pour limiter davantage les types d'instances pouvant être utilisées pour atteindre la capacité souhaitée. Pour plus d'informations sur chaque attribut, consultez InstanceRequirementsRequest dans le manuel Amazon EC2 API Reference.
- Types d'instance résultants : vous pouvez afficher les types d'instance qui correspondent aux exigences de calcul spécifiées, telles que les vCPU, la mémoire et le stockage.
- Pour exclure les types d'instance, choisissez Add Attribut (Ajouter un attribut). Dans la liste Attribute (Attributs), choisissez Excluded instance types (Types d'instances exclues). À partir de la liste Attribute value (Valeur d'attribut), sélectionnez les types d'instances à exclure.

c. Key pair (login) (Paire de clés [login]) : pour Key pair name (Nom de la paire de clés), choisissez une paire de clés existante ou choisissez Create new key pair (Créer une paire de clés) pour en créer une. Pour plus d'informations, consultez les paires de clés Amazon EC2 et les instances Linux dans le guide de l'utilisateur Amazon EC2.

d. Network settings (Paramètres réseau) : pour Firewall (security groups) (Pare-feu [groupes de sécurité]), utilisez un ou plusieurs groupes de sécurité, ou laissez ce champ vide et configurez un ou plusieurs groupes de sécurité comme faisant partie de l'interface réseau. Pour plus d'informations, veuillez consulter la section Amazon EC2 security groups for Linux instances (français non garanti) dans le Guide de l'utilisateur Amazon EC2.

   Si vous ne spécifiez aucun groupe de sécurité dans votre modèle de lancement, Amazon EC2 utilise le groupe de sécurité par défaut pour le VPC dans lequel votre groupe Auto Scaling lancera des instances. Par défaut, ce groupe de sécurité n'autorise pas le trafic entrant provenant de réseaux externes. Pour de plus amples informations, veuillez consulter la rubrique Groupes de sécurité par défaut pour votre VPC dans le Guide de l'utilisateur Amazon VPC.

e. Effectuez l'une des actions suivantes :

   - Modifier les paramètres par défaut de l'interface réseau. Par exemple, vous pouvez activer ou désactiver la fonction d'adressage IPv4 public, qui remplace le paramètre d'attribution automatique des adresses IPv4 publiques sur le sous-réseau. Pour plus

d'informations, consultez Modifier les paramètres par défaut de l'interface réseau (console).

- Ignorez cette étape pour conserver les paramètres par défaut de l'interface réseau.

f. Effectuez l'une des actions suivantes :

- Modifier la configuration du stockage. Pour plus d'informations, consultez Modifier la configuration du stockage (console).
- Ignorez cette étape pour conserver la configuration de stockage par défaut.

g. Pour le champ Resource tags (Balises de ressource), spécifiez les balises en fournissant les combinaisons clé et valeur. Si vous spécifiez des balises d'instance dans votre modèle de lancement, puis que vous avez choisi de propager les balises de votre groupe Auto Scaling à ses instances, toutes les balises sont fusionnées. Si la même clé d'identification est spécifiée pour une identification dans votre modèle de lancement et une identification dans votre groupe Auto Scaling, la valeur de l'identification du groupe est prioritaire.

6. (Facultatif) Configurer les paramètres avancés. Par exemple, vous pouvez choisir un rôle IAM que votre application peut utiliser lorsqu'elle accède à d'autres ressources AWS ou spécifier les données utilisateur d'instance pouvant être utilisées pour effectuer des tâches de configuration automatisées courantes après le démarrage d'une instance. Pour plus d'informations, consultez Créer un modèle de lancement à l'aide de paramètres avancés.
7. Lorsque vous êtes prêt à créer votre modèle de lancement, choisissez Create launch template (Créer un modèle de lancement).
8. Pour créer un groupe Auto Scaling, choisissez Créer un groupe Auto Scaling dans la page de confirmation.

Modifier les paramètres par défaut de l'interface réseau (console)

Les interfaces réseau fournissent une connectivité à d'autres ressources de votre VPC et à Internet. Pour plus d'informations, consultez Fournir une connectivité réseau pour vos instances Auto Scaling à l'aide d'Amazon VPC.

Cette section explique comment modifier les paramètres par défaut de l'interface réseau. Cela vous permet, par exemple, de définir si vous souhaitez attribuer une adresse IP publique à chaque instance au lieu d'utiliser par défaut le paramètre d'attribution automatique d'adresses IPv4 publiques sur le sous-réseau.

Considérations et restrictions

Lorsque vous modifiez les paramètres par défaut de l'interface réseau, gardez à l'esprit les considérations et limitations suivantes :

- Vous devez configurer les groupes de sécurité comme faisant partie de l'interface réseau, et non dans la section Groupes de sécurité du modèle. Vous ne pouvez pas spécifier de groupes de sécurité aux deux endroits.
- Vous ne pouvez pas attribuer d'adresses IP privées secondaires, appelées adresses IP secondaires, à une interface réseau.
- Si vous spécifiez un ID d'interface réseau existant, vous ne pouvez lancer qu'une seule instance. Pour ce faire, vous devez utiliser le AWS CLI ou un SDK pour créer le groupe Auto Scaling. Lorsque vous créez le groupe, vous devez spécifier la zone de disponibilité, mais pas l'ID de sous-réseau. En outre, vous pouvez spécifier une interface réseau existante uniquement si elle possède un index de périphérique de 0.
- Vous ne pouvez pas attribuer automatiquement une adresse IPv4 publique si vous spécifiez plus d'une interface réseau. Vous ne pouvez pas non plus spécifier les index de périphériques dupliqués sur les interfaces réseau. Les interfaces réseau primaire et secondaire résident toutes deux dans le même sous-réseau.
- Lorsqu'une instance est lancée, une adresse privée est automatiquement attribuée à chaque interface réseau. L'adresse provient de la plage CIDR du sous-réseau dans lequel l'instance est lancée. Pour plus d'informations sur la spécification des blocs d'adresses CIDR (ou plages d'adresses IP) pour votre VPC ou votre sous-réseau, consultez le Guide de l'utilisateur Amazon VPC.

Pour modifier les paramètres par défaut de l'interface réseau

1. Sous Network settings (Paramètres réseau) (Paramètres réseau), développez Advanced network configuration (Configuration réseau avancée).
2. Choisissez Add network interface (Ajouter une interface réseau) pour configurer l'interface réseau primaire, en faisant attention aux champs suivants :

    a. Device index (Index de périphérique) : conservez la valeur par défaut de 0 pour appliquer vos modifications à l'interface réseau principale (eth0).

    b. Network interface (Interface réseau) : conservez la valeur par défaut, New interface (Nouvelle interface), afin qu'Amazon EC2 Auto Scaling crée automatiquement une interface

réseau lorsqu'une instance est lancée. Vous pouvez également choisir une interface réseau disponible existante avec un index de périphérique égal à 0, mais cela limite votre groupe Auto Scaling à une seule instance.

c. Description : saisissez un nom descriptif.

d. Subnet (Sous-réseau) : conservez le paramètre par défaut Don't include in launch template (Ne pas inclure dans le modèle de lancement).

   Si l'AMI spécifie un sous-réseau pour l'interface réseau, il en résulte une erreur. Nous vous recommandons de désactiver Auto Scaling guidance (Recommandations Auto Scaling) comme solution de contournement. Après avoir fait ce changement, vous ne recevrez plus de message d'erreur. Toutefois, quel que soit l'endroit où le sous-réseau est spécifié, les paramètres de sous-réseau du groupe Auto Scaling ont priorité et ne peuvent pas être remplacés.

e. Auto-assign public IP (Attribuer automatiquement l'adresse IP publique) : indiquez si l'interface réseau associée à un index de périphérique 0 reçoit une adresse IPv4 publique. Par défaut, les instances d'un sous-réseau par défaut se voient attribuer une adresse IPv4 publique, tandis que les instances dans un sous-réseau non par défaut n'en reçoivent pas. Sélectionnez Activer ou Désactiver pour remplacer le paramètre par défaut du sous-réseau.

f. Security groups (Groupes de sécurité) : choisissez un ou plusieurs groupes de sécurité pour l'interface réseau. Chaque groupe de sécurité doit être configuré pour le VPC dans lequel votre groupe Auto Scaling lancera des instances. Pour plus d'informations, veuillez consulter la section Amazon EC2 security groups for Linux instances (français non garanti) dans le Guide de l'utilisateur Amazon EC2.

g. Delete on termination (Supprimer à la résiliation) : choisissez Yes (Oui) pour supprimer l'interface réseau lorsque l'instance est mise hors service, ou choisissez No (Non) pour conserver l'interface réseau.

h. Elastic Fabric Adapter : pour prendre en charge les cas d'utilisation du calcul haute performance (HPC) et du machine learning, changez l'interface réseau en une interface réseau Elastic Fabric Adapter. Pour plus d'informations, consultez Elastic Fabric Adapter dans le guide de l'utilisateur Amazon EC2.

i. Network card index (Index de la carte réseau) : choisissez 0 pour attacher l'interface réseau principale à la carte réseau avec un index de périphérique égal à 0. Si cette option n'est pas disponible, conservez la valeur par défaut, Don't include in launch template (Ne pas inclure dans le modèle de lancement). La connexion de l'interface réseau à une carte réseau spécifique n'est disponible que pour les types d'instance pris en charge. Pour plus

d'informations, consultez la section Cartes réseau dans le guide de l'utilisateur Amazon EC2.

j. ENA Express : pour les types d'instances compatibles avec ENA Express, choisissez Enable pour activer ENA Express ou Disable pour le désactiver. Pour plus d'informations, consultez Améliorer les performances du réseau avec ENA Express sur les instances Linux dans le guide de l'utilisateur Amazon EC2.

k. ENA Express UDP : si vous activez ENA Express, vous pouvez éventuellement l'utiliser pour le trafic UDP. Choisissez Enable pour activer ENA Express UDP ou Disable pour le désactiver.

3. Pour ajouter une interface réseau secondaire, choisissez Add network interface (Ajouter une interface réseau).

Modifier la configuration du stockage (console)

Vous pouvez modifier la configuration de stockage pour les instances lancées à partir d'une AMI d'Amazon EBS ou d'une AMI basée sur le stockage d'instance. Vous pouvez également spécifier des volumes EBS supplémentaires à attacher aux instances. L'AMI comprend un ou plusieurs volumes de stockage, dont le volume racine (Volume 1 (AMI Root)) (Volume 1 [racine AMI]).

Pour modifier la configuration du stockage

1. Dans Configure storage (Configurer le stockage), modifiez la taille ou le type de volume.

   Si la valeur que vous spécifiez pour la taille du volume est en dehors des limites du type de volume, ou inférieure à la taille de l'instantané, un message d'erreur s'affiche. Pour vous aider à résoudre ce problème, ce message donne la valeur minimale ou maximale que le champ peut accepter.

   Seuls les volumes associés à une AMI d'Amazon EBS s'affichent. Pour afficher des informations sur la configuration de stockage d'une instance lancée depuis une AMI basée sur le stockage d'instance, choisissez Show details (Afficher les détails) à partir de la section Instance store volumes (Volumes de stockage d'instance).

   Pour spécifier tous les paramètres de volume EBS, basculez sur la vue Advanced (Avancé) dans le coin supérieur droit.

2. Pour les options avancées, développez le volume que vous souhaitez modifier et configurer-le comme suit :

a. Storage type (Type de stockage) : le type de volume (EBS ou éphémère) à associer à votre instance. Le type de volume de stockage d'instances (éphémère) est disponible uniquement si vous sélectionnez un type d'instance qui le prend en charge. Pour plus d'informations, consultez les volumes Amazon EBS dans le guide de l'utilisateur Amazon EBS et le magasin d'instances Amazon EC2 dans le guide de l'utilisateur Amazon EC2.

b. Device name (Nom du dispositif) : sélectionnez le périphérique dans la liste des noms de périphériques disponibles pour le volume.

c. Snapshot (Instantané) : sélectionnez l'instantané à partir duquel vous souhaitez créer le volume. Vous pouvez rechercher les instantanés partagés et publics disponibles en saisissant un texte dans le champ Snapshot (Instantané).

d. Size (GiB) (Taille (Gio)) : pour les volumes EBS, vous pouvez spécifier une taille de stockage. Si vous avez sélectionné une AMI et une instance éligibles pour l'offre gratuite, n'oubliez pas que pour ne pas dépasser les limites de celle-ci, vous devez veiller à ne pas dépasser 30 GiO de stockage au total. Pour plus d'informations, consultez la section Contraintes relatives à la taille et à la configuration d'un volume EBS dans le guide de l'utilisateur Amazon EBS.

e. Volume type (Type de volume) : pour les volumes EBS, choisissez le type de volume. Pour plus d'informations, consultez les types de volumes Amazon EBS dans le guide de l'utilisateur Amazon EBS.

f. IOPS : si vous avez sélectionné un type de volume SSD d'IOPS provisionnés (`io1` et `io2`) ou SSD polyvalents (`gp3`), alors vous pouvez saisir le nombre d'opérations d'I/O par seconde (IOPS) que le volume peut prendre en charge. Ceci est requis pour les volumes io1, io2 et gp3. Il n'est pas pris en charge pour les volumes gp2, st1, sc1 ou standard.

g. Delete on termination (Supprimer à la résiliation) : pour les volumes EBS, choisissez Yes (Oui) pour supprimer le volume lors de la résiliation de l'instance ou No (Non) pour conserver le volume.

h. Encrypted (Chiffré) : si le type d'instance prend en charge le chiffrement EBS, vous pouvez sélectionner Yes (Oui) pour activer le chiffrement du volume. Si vous avez activé le chiffrement par défaut dans cette région, le chiffrement est activé automatiquement. Pour plus d'informations, consultez les sections Cryptage Amazon EBS et Activer le chiffrement par défaut dans le Guide de l'utilisateur Amazon EBS.

   L'effet par défaut qu'entraîne ce paramètre varie en fonction de la source du volume, comme décrit dans le tableau ci-dessous. Dans tous les cas, vous devez être autorisé à utiliser ce qui est spécifié AWS KMS key.

Résultats du chiffrement

Si le paramètre **`Encrypted`** est défini sur...	Et si la source du volume est...	Alors l'état de chiffrement par défaut est...	Remarques
Non	Nouveau volume (vide)	Non chiffré(e)*	N/A
	Instantané non chiffré que vous possédez	Non chiffré(e)*	
	Instantané chiffré que vous possédez	Chiffré par la même clé	
	Instantané non chiffré qui est partagé avec vous	Non chiffré(e)*	
	Instantané chiffré qui est partagé avec vous	Chiffré par clé KMS par défaut	
Oui	Nouveau volume	Chiffré par clé KMS par défaut	Pour utiliser une clé KMS autre que celle par défaut, spécifiez une valeur pour le paramètre KMS Key (Clé KMS).
	Instantané non chiffré que vous possédez	Chiffré par clé KMS par défaut	
	Instantané chiffré que vous possédez	Chiffré par la même clé	
	Instantané non chiffré qui est partagé avec vous	Chiffré par clé KMS par défaut	
	Instantané chiffré qui est partagé avec vous	Chiffré par clé KMS par défaut	

* Si le chiffrement par défaut est activé, tous les nouveaux volumes récemment créés (indépendamment du fait que le paramètre Encrypted [Chiffré] soit défini sur Yes [Oui])

sont chiffrés à l'aide de la clé KMS par défaut. Si vous définissez à la fois les paramètres Encrypted (Chiffré) et KMS Key (Clé KMS), vous pouvez alors spécifier une clé KMS autre que celle par défaut.

   i. KMS Key (Clé KMS) : si vous avez sélectionné Yes (Oui) pour Encrypted (Chiffré), vous devez ensuite sélectionner une clé gérée par le client à utiliser pour chiffrer le volume. Si vous avez activé le chiffrement par défaut dans cette région, la clé gérée par le client par défaut est sélectionnée pour vous. Vous pouvez sélectionner une clé différente ou spécifier l'ARN de n'importe quelle clé gérée par le client que vous avez créée à l'aide de AWS Key Management Service.

3. Pour spécifier d'autres volumes à attacher aux instances lancées par ce modèle de lancement, choisissez Add new volume (Ajouter un nouveau volume).

Créer un modèle de lancement à partir d'une instance existante (console)

Pour créer un modèle de lancement à partir d'une instance existante

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Dans le panneau de navigation, sous Instances, choisissez Instances.
3. Sélectionnez l'instance et choisissez Actions, Image et modèles, Créer un modèle à partir d'une instance.
4. Fournissez un nom et une description.
5. Sous Auto Scaling guidance (Guide Auto Scaling), activez la case à cocher.
6. Ajustez les paramètres si nécessaire, puis choisissez Create launch template (Créer un modèle de lancement)Create launch template.
7. Pour créer un groupe Auto Scaling, choisissez Créer un groupe Auto Scaling dans la page de confirmation.

Ressources connexes

Nous fournissons quelques extraits de modèles JSON et YAML que vous pouvez utiliser pour comprendre comment déclarer des modèles de lancement dans vos AWS CloudFormation modèles de pile. Pour plus d'informations, consultez les AWS CloudFormation sections AWS::EC2::LaunchTemplateet Créer des modèles de lancement avec du Guide de l'AWS CloudFormation utilisateur.

Pour plus d'informations sur les modèles de lancement, consultez la section Lancement d'une instance à partir d'un modèle de lancement dans le guide de l'utilisateur Amazon EC2.

Limites

- Bien que vous puissiez spécifier un sous-réseau dans un modèle de lancement, cela n'est pas nécessaire si vous utilisez le modèle de lancement uniquement pour créer des groupes Auto Scaling. Vous ne pouvez pas spécifier le sous-réseau d'un groupe Auto Scaling en le spécifiant dans un modèle de lancement. Les sous-réseaux du groupe Auto Scaling sont issus de la propre définition de ressource du groupe Auto Scaling.
- Pour d'autres limitations sur les interfaces réseau définies par l'utilisateur, consultez Modifier les paramètres par défaut de l'interface réseau (console).

Créer un modèle de lancement à l'aide de paramètres avancés

Cette rubrique explique comment créer un modèle de lancement avec des paramètres avancés à partir du AWS Management Console.

Pour créer un modèle de lancement à l'aide des paramètres avancés

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Dans le volet de navigation, sous Instances, choisissez Modèles de lancement, puis choisissez Créer un modèle de lancement.
3. Configurez votre modèle de lancement comme décrit dans les rubriques suivantes :
   - Réglages requis
   - Paramètres avancés
4. Choisissez Créer un modèle de lancement.

Réglages requis

Lorsque vous créez un modèle de lancement, vous devez inclure les paramètres obligatoires suivants.

Nom du modèle de lancement

Entrez un nom unique qui décrit le modèle de lancement.

Images d’applications et de systèmes d’exploitation (Amazon Machine Image)

Choisissez l'Amazon Machine Image (AMI) que vous souhaitez utiliser. Vous pouvez rechercher ou parcourir l'AMI que vous souhaitez utiliser. Pour optimiser l'efficacité de la mise à l'échelle, choisissez une AMI personnalisée entièrement configurée pour lancer une instance avec le code de votre application et nécessitant peu de modifications au lancement.

Type d’instance

Choisissez un type d'instance compatible avec votre AMI. Vous pouvez ignorer l'ajout d'un type d'instance à votre modèle de lancement si vous prévoyez d'utiliser plusieurs types d'instances intégrés dans la définition des ressources du groupe Auto Scaling. Un type d'instance n'est requis que si vous ne prévoyez pas de créer un groupe d'instances mixtes.

Paramètres avancés

Les paramètres avancés sont facultatifs. Si vous ne configurez aucun paramètre avancé, les fonctionnalités spécifiques ne seront pas ajoutées à vos instances.

Développez la section Détails avancés pour afficher les paramètres avancés. Les sections suivantes décrivent les paramètres avancés les plus utiles sur lesquels se concentrer lors de la création d'un modèle de lancement pour un groupe Auto Scaling. Pour plus d'informations, consultez les informations avancées dans le guide de l'utilisateur Amazon EC2.

Profil d'instance IAM

Le profil d'instance contient le rôle IAM que vous souhaitez utiliser. Lorsque votre groupe Auto Scaling lance une instance EC2, les autorisations définies dans le rôle IAM associé sont accordées aux applications exécutées sur l'instance. Pour plus d’informations, consultez Rôle IAM pour les applications qui s'exécutent sur des instances Amazon EC2.

Protection de la résiliation

Lorsqu'elle est activée, cette fonctionnalité empêche les utilisateurs de mettre fin à une instance à l'aide de la console Amazon EC2, des commandes CLI et des opérations d'API. La protection contre les interruptions constitue une protection supplémentaire contre les interruptions accidentelles. Cela n'empêche pas Amazon EC2 Auto Scaling de mettre fin à une instance. Pour contrôler les instances auxquelles Amazon EC2 Auto Scaling peut mettre fin, consultez. Utiliser la protection de la taille d'instance

CloudWatch Surveillance détaillée

Vous pouvez activer la surveillance détaillée de vos instances EC2 afin de leur permettre d'envoyer des données métriques à Amazon à CloudWatch intervalles d'une minute. Par défaut, les instances EC2 envoient des données métriques à CloudWatch des intervalles de 5 minutes. Des frais supplémentaires seront facturés. Pour plus d'informations, consultez Configurer la surveillance pour les instances à scalabilité automatique.

Spécification du crédit

Amazon EC2 fournit des instances de performances évolutives, telles que T2, T3 et T3a, qui permettent aux applications de dépasser les performances de base du processeur lorsque cela est nécessaire. Par défaut, ces instances peuvent éclater pendant une durée limitée avant que leur utilisation du processeur ne soit limitée. Vous pouvez éventuellement activer le mode illimité afin que les instances puissent dépasser la ligne de base aussi longtemps que nécessaire. Cela permet aux applications de maintenir des performances élevées du processeur lorsque cela est nécessaire. Des frais supplémentaires peuvent être facturés. Pour plus d'informations, consultez Use an Auto Scaling group to launch a burstable performance instance Unlimited dans le guide de l'utilisateur Amazon EC2.

Nom du groupe de placement

Vous pouvez spécifier un groupe de placement et utiliser une stratégie de cluster ou de partition pour influencer la manière dont vos instances sont physiquement situées dans le centre de AWS données. Pour les petits groupes Auto Scaling, vous pouvez également utiliser la stratégie de spread. Pour plus d'informations, consultez la section Groupes de placement dans le guide de l'utilisateur Amazon EC2.

Certaines considérations doivent être prises en compte lors de l'utilisation de groupes de placement avec des groupes Auto Scaling :

- Si un groupe de placement est spécifié à la fois dans le modèle de lancement et dans le groupe Auto Scaling, le groupe de placement du groupe Auto Scaling est prioritaire. Une fois le groupe créé, le groupe de placement spécifié dans les paramètres du groupe Auto Scaling ne peut pas être modifié.
- Dans AWS CloudFormation, soyez prudent si vous définissez un groupe de placement dans le modèle de lancement. Amazon EC2 Auto Scaling lancera les instances dans le groupe de placement spécifié. Cependant, vous ne CloudFormation recevrez pas de signaux de ces instances si vous en utilisez un UpdatePolicyavec votre groupe Auto Scaling (bien que cela puisse changer à l'avenir).

Option d'achat

Vous pouvez choisir Request Spot Instances pour demander des instances Spot au prix Spot, plafonné au prix à la demande, et choisir Personnaliser pour modifier les paramètres par défaut des instances Spot. Pour un groupe Auto Scaling, vous devez spécifier une demande unique sans date de fin (valeur par défaut). Pour plus d'informations, consultez Demander des instances Spot pour des applications flexibles et tolérantes aux pannes. Ce paramètre peut être utile dans des circonstances particulières, mais en général, il est préférable de ne pas le spécifier et de créer plutôt un groupe d'instances mixtes. Pour plus d'informations, consultez Groupes Auto Scaling combinant plusieurs types d'instances et options d'achat.

Si vous spécifiez une demande d'instance Spot dans votre modèle de lancement, vous ne pouvez pas créer de groupe d'instances mixtes. Si vous essayez d'utiliser un modèle de lancement qui demande des instances Spot avec un groupe d'instances mixtes, le message d'erreur suivant s'affiche : `Incompatible launch template: You cannot use a launch template that is set to request Spot Instances (InstanceMarketOptions) when you configure an Auto Scaling group with a mixed instances policy. Add a different launch template to the group and try again.`

Capacity Reservation

Les réservations de capacité vous permettent de réserver de la capacité pour vos instances Amazon EC2 dans une zone de disponibilité spécifique pour une durée quelconque. Pour plus d'informations, consultez la section Réservations de capacité à la demande dans le guide de l'utilisateur Amazon EC2.

Vous pouvez choisir de lancer des instances dans :

- toute réservation de capacité ouverte (ouverte)
- une réservation de capacité spécifique (cible par ID)
- un groupe de réservations de capacité (cible par groupe)

Pour cibler une réservation de capacité spécifique, le type d'instance indiqué dans votre modèle de lancement doit correspondre au type d'instance de la réservation. Lorsque vous créez votre groupe Auto Scaling, utilisez la même zone de disponibilité que la réservation de capacité. En fonction de Région AWS votre choix, vous pouvez choisir de cibler un bloc de capacité à la place. Pour plus d'informations, consultez Utilisation Capacity Blocks pour les charges de travail liées à l'apprentissage automatique.

Pour cibler un groupe de réservations de capacité, voirUtilisez les réserves de capacité à la demande pour réserver de la capacité dans des zones de disponibilité spécifiques. En ciblant un groupe de réservations de capacité, vous pouvez répartir la capacité sur plusieurs zones de disponibilité afin d'améliorer la résilience.

Location

Amazon EC2 propose trois options pour la location de vos instances EC2 :

- Partagé (partagé) — Plusieurs Comptes AWS peuvent partager le même matériel physique. Il s'agit de l'option de location par défaut lors du lancement d'une instance.
- Instances dédiées (dédiées) : votre instance s'exécute sur du matériel à locataire unique. Aucun autre AWS client ne partage le même serveur physique. Pour plus d'informations, consultez Instances dédiées dans le Guide de l'utilisateur Amazon EC2.
- Hôtes dédiés (hôte dédié) : l'instance s'exécute sur un serveur physique dédié à votre usage. L'utilisation d'hôtes dédiés permet d'apporter plus facilement à EC2 vos propres licences (BYOL) répondant à des exigences matérielles spécifiques et répondant aux cas d'utilisation liés à la conformité. Si vous choisissez cette option, vous devez fournir un groupe de ressources hôtes pour le groupe de ressources hôte Tenancy. Pour plus d'informations, consultez la section sur les hôtes dédiés dans le guide de l'utilisateur Amazon EC2.

Support pour les hôtes dédiés uniquement si vous spécifiez un groupe de ressources hôtes. Vous ne pouvez pas cibler un hôte spécifique ou utiliser l'affinité de placement de l'hôte.

- Si vous essayez d'utiliser un modèle de lancement qui spécifie un ID d'hôte, le message d'erreur suivant s'affiche : `Incompatible launch template: Tenancy host ID is not supported for Auto Scaling.`
- Si vous essayez d'utiliser un modèle de lancement qui spécifie l'affinité de placement de l'hôte, le message d'erreur suivant s'affiche : `Incompatible launch template: Auto Scaling does not support host placement affinity.`

Groupe de ressources de l'hôte locataire

Vous pouvez y apporter vos propres licences AWS et les gérer de manière centralisée. AWS License Manager Un groupe de ressources d'hôtes est un groupe d'hôtes dédiés liés à une configuration de licence License Manager spécifique. Les groupes de ressources hôtes vous permettent de lancer facilement des instances EC2 sur des hôtes dédiés qui répondent à vos besoins en matière de licences logicielles. Il n'est pas nécessaire d'allouer manuellement des hôtes dédiés à l'avance. Ils sont automatiquement créés selon les besoins. Notez que lorsque vous associez une AMI à une configuration de licence, cette AMI ne peut être associée qu'à un

seul groupe de ressources hôtes à la fois. Pour plus d'informations, consultez la section Groupes de ressources Host AWS License Manager dans le Guide de l'utilisateur du License Manager.

Configurations de licence

Ce paramètre vous permet de définir une configuration de licence pour vos instances sans restreindre leur location à des hôtes dédiés. La configuration des licences permet de suivre les licences logicielles déployées sur les instances afin que vous puissiez surveiller l'utilisation et la conformité de vos licences. Pour plus d'informations, consultez la section Création d'une licence autogérée dans le Guide de l'utilisateur du License Manager.

Métadonnées accessibles

Vous pouvez choisir d'activer ou de désactiver l'accès au point de terminaison HTTP du service de métadonnées d'instance. Par défaut, le point de terminaison HTTP est activé. Si vous choisissez de désactiver le point de terminaison, l'accès aux métadonnées de votre instance est désactivé. Vous pouvez spécifier la condition pour exiger IMDSv2 uniquement lorsque le point de terminaison HTTP est activé. Pour plus d'informations, consultez Configurer les options de métadonnées de l'instance dans le guide de l'utilisateur Amazon EC2.

Version des métadonnées

Vous pouvez choisir d'exiger l'utilisation du service de métadonnées d'instance version 2 (IMDSv2) lorsque vous demandez des métadonnées d'instance. Si vous ne spécifiez pas de valeur, la valeur par défaut est de prendre en charge IMDSv1 et IMDSv2. Pour plus d'informations, consultez Configurer les options de métadonnées de l'instance dans le guide de l'utilisateur Amazon EC2.

Limite de sauts de réponse des jetons de métadonnées

Vous pouvez définir le nombre de sauts réseau autorisés pour le jeton de métadonnées. Si vous ne spécifiez pas de valeur, la valeur par défaut est 1. Pour plus d'informations, consultez Configurer les options de métadonnées de l'instance dans le guide de l'utilisateur Amazon EC2.

Données utilisateur

Vous pouvez personnaliser et terminer la configuration de vos instances au moment du lancement en spécifiant des scripts shell ou des directives cloud-init sous forme de données utilisateur. Les données utilisateur s'exécutent au démarrage initial de l'instance, ce qui vous permet d'installer automatiquement des applications, des dépendances ou des personnalisations au moment du lancement. Pour plus d'informations, consultez la section Exécuter des commandes sur votre instance Linux au lancement dans le guide de l'utilisateur Amazon EC2.

Si vous avez des téléchargements volumineux ou des scripts complexes, cela augmente le temps nécessaire pour que l'instance soit prête à être utilisée. Dans ce cas, vous devrez peut-être configurer un hook de cycle de vie pour empêcher une instance d'atteindre l'`InService`état jusqu'à ce qu'elle soit entièrement provisionnée. Pour plus d'informations sur l'ajout d'un hook de cycle de vie à votre groupe Auto Scaling, consultezHooks de cycle de vie Amazon EC2 Auto Scaling.

Demander des instances Spot pour des applications flexibles et tolérantes aux pannes

Dans votre modèle de lancement, vous pouvez éventuellement demander des Instances Spot sans date de fin ni durée. Les instances Spot Amazon EC2 sont disponibles à des réductions importantes par rapport au prix EC2 à la demande. Les instances Spot constituent un choix économique si vous êtes flexible quant au moment où vos applications s'exécutent et à la possibilité de les interrompre. Pour plus d'informations sur la création d'un modèle de lancement qui demande des Instances Spot, consultez Créer un modèle de lancement à l'aide de paramètres avancés.

> **Important**
>
> Les instances Spot sont généralement utilisées pour compléter les instances à la demande. Dans ce scénario, vous pouvez spécifier les mêmes paramètres que ceux qui sont utilisés pour lancer des instances Spot dans le cadre des paramètres de votre groupe Auto Scaling. Lorsque vous spécifiez les paramètres dans le groupe Auto Scaling, vous pouvez demander de lancer des instances Spot uniquement après avoir lancé un certain nombre d'instances à la demande, puis continuer à lancer une combinaison d'instances à la demande et d'instances Spot au fur et à mesure que le groupe évolue. Pour plus d'informations, consultez Groupes Auto Scaling combinant plusieurs types d'instances et options d'achat.

Cette rubrique décrit comment lancer uniquement des Instances Spot dans votre groupe Auto Scaling en spécifiant les paramètres dans un modèle de lancement, plutôt que dans le groupe Auto Scaling lui-même. Les informations de cette rubrique s'appliquent également aux groupes Auto Scaling qui demandent des instances Spot avec une configuration du lancement. La différence est qu'une configuration du lancement requiert un prix maximum, mais pour les modèles de lancement, le prix maximum est facultatif.

Lorsque vous créez un modèle de lancement pour lancer uniquement des instances Spot, gardez à l'esprit les considérations suivantes :

- prix Spot. Vous ne payez que le prix Spot actuel pour les instances Spot que vous lancez. Cette tarification change lentement au fil du temps en fonction des tendances à long terme de l'offre et de la demande. Pour plus d'informations, consultez les sections Instances Spot et tarification et économies dans le guide de l'utilisateur Amazon EC2.
- Paramétrage du prix maximum. Vous pouvez éventuellement inclure un prix maximum par heure pour les Instances Spot dans votre modèle de lancement. Si votre prix maximum dépasse le prix Spot actuel, le service Spot d'Amazon EC2 satisfait votre demande immédiatement si la capacité est disponible. Si le prix de l'instance Spot dépasse votre prix maximum pour une instance en cours d'exécution dans votre groupe Auto Scaling, il résilie votre instance.

  > **Warning**
  >
  > Votre application peut ne pas fonctionner si vous ne recevez pas d'Instances Spot, par exemple lorsque votre prix maximum est trop bas. Pour profiter des Instances Spot disponibles le plus longtemps possible, définissez votre prix maximum proche du prix à la demande.

- Équilibrage sur toutes les zones de disponibilité. Si vous spécifiez plusieurs zones de disponibilité, Amazon EC2 Auto Scaling répartit les demandes ponctuelles sur ces zones de disponibilité. Si votre prix maximum pour les instances Spot est trop faible dans une zone de disponibilité et ne permet pas de satisfaire les demandes, Amazon EC2 Auto Scaling vérifie si les demandes ont été satisfaites dans les autres zones de disponibilité. Si c'est le cas, Amazon EC2 Auto Scaling annule les demandes qui ont échoué et les répartit entre les zones de disponibilité qui ont des demandes satisfaites. Si le prix dans une zone de disponibilité n'ayant aucune demande satisfaite baisse suffisamment pour que les demandes futures soient acceptées, Amazon EC2 Auto Scaling procède à un rééquilibrage sur toutes les zones de disponibilité.
- Résiliation d'instance Spot. Les Instances Spot peuvent être résiliées à tout moment. Le service Spot d'Amazon EC2 peut résilier des Instances Spot dans votre groupe Auto Scaling si la disponibilité ou le prix des Instances Spot change. Lors de la mise à l'échelle ou de la surveillance de l'état, Amazon EC2 Auto Scaling peut également résilier les Instances Spot de la même manière qu'il peut résilier les Instances à la demande. Lorsqu'une instance est résiliée, tout stockage est supprimé.
- Maintenir la capacité souhaitée. Lorsqu'une Instance Spot est résiliée, Amazon EC2 Auto Scaling tente de lancer une autre Instance Spot pour maintenir la capacité souhaitée pour le groupe. Si le

prix Spot actuel est inférieur à votre prix maximum, il lance une Instance Spot. Si la demande d'une Instance Spot n'aboutit pas, il continue à essayer.

- Modification du prix maximum. Pour modifier votre prix maximum, créez un nouveau modèle de lancement ou mettez à jour un modèle de lancement existant avec le nouveau prix maximum, puis associez-le à votre groupe Auto Scaling. Les Instances Spot existantes continuent à s'exécuter tant que le prix maximum spécifié dans le modèle de lancement utilisé pour ces instances est supérieur au prix Spot actuel. Si vous n'avez pas défini de prix maximum, le prix maximum par défaut est le prix à la demande.

Utilisation Capacity Blocks pour les charges de travail liées à l'apprentissage automatique

Capacity Blocksvous aider à réserver des instances de GPU très recherchées à une date future afin de prendre en charge vos charges de travail de courte durée liées à l'apprentissage automatique (ML).

Pour un aperçu de leur fonctionnement Capacity Blocks et de leur fonctionnement, consultez le manuel Capacity Blocksd'apprentissage automatique dans le guide de l'utilisateur Amazon EC2.

Pour commencer à utiliserCapacity Blocks, vous devez créer une réservation de capacité dans une zone de disponibilité spécifique. Capacity Blockssont livrés sous forme `targeted` de réservations de capacité dans une seule zone de disponibilité. Lorsque vous créez votre modèle de lancement, spécifiez l'ID de réservation et le type d'instance du Capacity Block. Mettez ensuite à jour votre groupe Auto Scaling pour utiliser le modèle de lancement que vous avez créé et la zone de disponibilité du Capacity Block. Lorsque votre réservation de bloc de capacité commence, utilisez le dimensionnement planifié pour lancer le même nombre d'instances que votre réservation de bloc de capacité.

> **Important**
>
> Capacity Blocksne sont disponibles que pour certains types d'instances Amazon EC2 et. Régions AWS Pour plus d'informations, consultez la section Conditions requises dans le guide de l'utilisateur Amazon EC2.

Table des matières

- Spécifier un bloc de capacité dans votre modèle de lancement
- Limites
- Ressources connexes

Directives opérationnelles

Voici les directives opérationnelles de base que vous devez suivre lorsque vous utilisez un bloc de capacité avec un groupe Auto Scaling.

- Effectuez une mise à l'échelle horizontale à zéro de votre groupe Auto Scaling plus de 30 minutes avant l'heure de fin de la réservation du bloc de capacité. Amazon EC2 mettra fin à toutes les instances encore en cours d'exécution 30 minutes avant l'heure de fin du bloc de capacité.
- Nous vous recommandons d'utiliser le dimensionnement planifié pour augmenter (ajouter des instances) et augmenter (supprimer des instances) aux heures de réservation appropriées. Pour plus d'informations, consultez Mise à l'échelle planifiée pour Amazon EC2 Auto Scaling.
- Ajoutez des hooks de cycle de vie selon les besoins pour effectuer un arrêt optimal de votre application dans les instances lors de la mise à l'échelle. Laissez suffisamment de temps pour que l'action du cycle de vie soit terminée avant qu'Amazon EC2 ne commence à résilier de force vos instances 30 minutes avant l'heure de fin de la réservation du bloc de capacité. Pour plus d'informations, consultez Hooks de cycle de vie Amazon EC2 Auto Scaling.
- Assurez-vous que le groupe Auto Scaling pointe vers la bonne version du modèle de lancement pendant toute la durée de la réservation. Nous vous recommandons de pointer vers une version spécifique du modèle de lancement plutôt que vers la version `$Default` ou `$Latest`.

> **Note**
>
> Si vous laissez une instance de Capacity Block en cours d'exécution jusqu'à la fin de la réservation et qu'Amazon EC2 la récupère, les activités de dimensionnement de votre groupe Auto Scaling indiquent qu'elle était `taken out of service in response to an EC2 health check that indicated it had been terminated or stopped` « », même si elle a été volontairement récupérée à la fin du Capacity Block. De même, Amazon EC2 Auto Scaling tentera de remplacer l'instance de la même manière que pour toute instance dont le bilan de santé échoue. Pour plus d'informations, consultez Surveillance de l'état des instances dans un groupe Auto Scaling.

Spécifier un bloc de capacité dans votre modèle de lancement

Pour créer un modèle de lancement qui cible un bloc de capacité spécifique pour votre groupe Auto Scaling, appliquez l'une des méthodes suivantes :

Console

Pour spécifier un bloc de capacité dans votre modèle de lancement (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Dans la barre de navigation supérieure, sélectionnez l' Région AWS endroit où vous avez créé votre bloc de capacité.
3. Dans le volet de navigation, sous Instances, choisissez Modèles de lancement.
4. Choisissez Créer un modèle de lancement, puis créez le modèle de lancement. Indiquez l'ID d'Amazon Machine Image (AMI), le type d'instance et tout autre paramètre du modèle de lancement, le cas échéant.
5. Développez la section Détails avancés pour afficher les paramètres avancés.
6. Pour l'option d'achat, choisissez Blocs de capacité.
7. Pour la réservation de capacité, choisissez Cible par ID, puis pour Réservation de capacité - Cible par ID, choisissez l'ID de réservation de capacité d'un bloc de capacité existant.
8. Lorsque vous avez terminé, choisissez Créer un modèle de lancement.

   Pour obtenir de l'aide sur la création d'un groupe Auto Scaling avec un modèle de lancement, consultezCréer un groupe Auto Scaling avec un modèle de lancement.

AWS CLI

Pour spécifier un bloc de capacité dans votre modèle de lancement (AWS CLI)

Utilisez la commande create-launch-template suivante pour créer un modèle de lancement spécifiant un ID de réservation de bloc de capacité. Remplacez chaque *espace réservé à la saisie de l'utilisateur* par vos propres informations.

```
aws ec2 create-launch-template --launch-template-name my-template-for-capacity-block
  \
  --version-description AutoScalingVersion1 --region us-east-2 \
  --launch-template-data file://config.json
```

> **Tip**
>
> Si cette commande génère une erreur, assurez-vous d'avoir mis à jour AWS CLI localement la dernière version.

Contenu de `config.json`.

```
{
    "ImageId": "ami-04d5cc9b88example",
    "InstanceType": "p4d.24xlarge",
    "SecurityGroupIds": [
        "sg-903004f88example"
    ],
    "KeyName": "MyKeyPair",
    "InstanceMarketOptions": {
        "MarketType": "capacity-block"
    },
    "CapacityReservationSpecification": {
        "CapacityReservationTarget": {
            "CapacityReservationId": "cr-02168da1478b509e0"
        }
    }
}
```

Voici un exemple de sortie.

```
{
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-068f72b724example",
        "LaunchTemplateName": "my-template-for-capacity-block",
        "CreateTime": "2023-10-27T15:12:44.000Z",
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
        "DefaultVersionNumber": 1,
        "LatestVersionNumber": 1
    }
}
```

Vous pouvez utiliser la commande describe-launch-template-versions suivante pour vérifier l'ID de réservation du bloc de capacité associé au modèle de lancement.

```
aws ec2 describe-launch-template-versions --launch-template-names my-template-for-capacity-block \
  --region us-east-2
```

Voici un exemple de sortie d'un modèle de lancement indiquant une réservation de bloc de capacité.

```
{
    "LaunchTemplateVersions": [
        {
            "LaunchTemplateId": "lt-068f72b724example",
            "LaunchTemplateName": "my-template-for-capacity-block",
            "VersionNumber": 1,
            "CreateTime": "2023-10-27T15:12:44.000Z",
            "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
            "DefaultVersion": true,
            "LaunchTemplateData": {
                "ImageId": "ami-04d5cc9b88example",
                "InstanceType": "p5.48xlarge",
                "SecurityGroupIds": [
                    "sg-903004f88example"
                ],
                "KeyName": "MyKeyPair",
                "InstanceMarketOptions": {
                    "MarketType": "capacity-block"
                },
                "CapacityReservationSpecification": {
                    "CapacityReservationTarget": {
                        "CapacityReservationId": "cr-02168da1478b509e0"
                    }
                }
            }
        }
    ]
}
```

Limites

- Support pour n'Capacity Blocksest disponible que si votre groupe Auto Scaling possède une configuration compatible. Les groupes d'instances mixtes et les groupes chauds ne sont pas pris en charge.

- Vous ne pouvez cibler qu'un seul bloc de capacité à la fois.

Ressources connexes

- Pour connaître les conditions préalables et les recommandations relatives à l'utilisation des instances P5, consultez la section Commencer avec les instances P5 dans le guide de l'utilisateur Amazon EC2.
- Amazon EKS prend en charge l'utilisation Capacity Blocks pour prendre en charge vos charges de travail d'apprentissage automatique (ML) de courte durée sur les clusters Amazon EKS. Pour plus d'informations, consultez la section Capacity Blocksconsacrée au ML dans le guide de l'utilisateur Amazon EKS.
- Vous pouvez l'utiliser Capacity Blocks avec les types d'instances et les régions pris en charge. Cependant, les réservations de capacité à la demande offrent la flexibilité nécessaire pour réserver de la capacité pour d'autres types d'instances et régions. Pour consulter un didacticiel expliquant comment utiliser l'option de réservation de capacité à la demande, consultezUtilisez les réserves de capacité à la demande pour réserver de la capacité dans des zones de disponibilité spécifiques.

Migrez vos groupes Auto Scaling pour lancer des modèles

À partir de 2023, vous ne pouvez plus appeler `CreateLaunchConfiguration` avec les nouveaux types d'instances Amazon EC2 publiés après le 31 décembre 2022. Pour plus d'informations, consultez Configurations de lancement.

Pour migrer vos groupes Auto Scaling des configurations de lancement vers les modèles de lancement, consultez les étapes suivantes.

> **Important**
>
> Avant de poursuivre, assurez-vous de disposer des autorisations requises pour utiliser les modèles de lancement. Pour plus d'informations, consultez Autorisations d'utilisation des modèles de lancement.

Étape 1 : Trouver les groupes Auto Scaling utilisant des configurations de lancement

Pour déterminer si des groupes Auto Scaling utilisent toujours des configurations de lancement, exécutez la commande describe-auto-scaling-groups suivante à l'aide de l' AWS CLI. Remplacez *REGION* par votre Région AWS.

```
aws autoscaling describe-auto-scaling-groups --region REGION \
  --query 'AutoScalingGroups[?LaunchConfigurationName!=`null`]'
```

Voici un exemple de sortie.

```
[
    {
 "AutoScalingGroupName": "group-1",
 "AutoScalingGroupARN": "arn",
 "LaunchConfigurationName": "my-launch-config",
 "MinSize": 1,
 "MaxSize": 5,
 "DesiredCapacity": 2,
 "DefaultCooldown": 300,
 "AvailabilityZones": [
           "us-west-2a",
           "us-west-2b",
           "us-west-2c"
       ],
 "LoadBalancerNames": [],
 "TargetGroupARNs": [],
 "HealthCheckType": "EC2",
 "HealthCheckGracePeriod": 300,
 "Instances": [
           {
               "ProtectedFromScaleIn": false,
               "AvailabilityZone": "us-west-2a",
               "LaunchConfigurationName": "my-launch-config",
               "InstanceId": "i-05b4f7d5be44822a6",
               "InstanceType": "t3.micro",
               "HealthStatus": "Healthy",
               "LifecycleState": "InService"
           },
           {
               "ProtectedFromScaleIn": false,
```

```
                "AvailabilityZone": "us-west-2b",
                "LaunchConfigurationName": "my-launch-config",
                "InstanceId": "i-0c20ac468fa3049e8",
                "InstanceType": "t3.micro",
                "HealthStatus": "Healthy",
                "LifecycleState": "InService"
            }
 ],
 "CreatedTime": "2023-03-09T22:15:11.611Z",
 "SuspendedProcesses": [],
 "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
 "EnabledMetrics": [],
 "Tags": [
            {
  "ResourceId": "group-1",
  "ResourceType": "auto-scaling-group",
  "Key": "environment",
  "Value": "production",
  "PropagateAtLaunch": true
            }
        ],
 "TerminationPolicies": [
     "Default"
 ],
 "NewInstancesProtectedFromScaleIn": false,
 "ServiceLinkedRoleARN": "arn",
       "TrafficSources": []
    },

    ... additional groups ...

]
```

Sinon, pour tout supprimer sauf les noms des groupes Auto Scaling avec les noms de leurs configurations de lancement et de leurs balises respectives dans la sortie, exécutez la commande suivante :

```
aws autoscaling describe-auto-scaling-groups --region REGION \
  --query 'AutoScalingGroups[?LaunchConfigurationName!=`null`].{AutoScalingGroupName:
 AutoScalingGroupName, LaunchConfigurationName: LaunchConfigurationName, Tags: Tags}'
```

Voici un exemple de sortie.

```
[
    {
        "AutoScalingGroupName": "group-1",
        "LaunchConfigurationName": "my-launch-config",
        "Tags": [
            {
                "ResourceId": "group-1",
                "ResourceType": "auto-scaling-group",
                "Key": "environment",
                "Value": "production",
                "PropagateAtLaunch": true
            }
        ]
    },

    ... additional groups ...

]
```

Pour plus d'informations sur le filtrage, consultez la section Filtrage AWS CLI de la sortie dans le guide de AWS Command Line Interface l'utilisateur.

Étape 2 : Copier une configuration de lancement vers un modèle de lancement

Vous pouvez copier une configuration de lancement vers un modèle de lancement à l'aide de la procédure suivante. Ensuite, vous pouvez l'ajouter à votre groupe Auto Scaling.

La copie de plusieurs configurations de lancement entraîne des modèles de lancement portant le même nom. Pour modifier le nom donné à un modèle de lancement pendant le processus de copie, vous devez copier les configurations de lancement une par une.

> **Note**
>
> La fonction de copie est uniquement disponible depuis la console.

Pour copier une configuration de lancement vers un modèle de lancement (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.

2. Dans le volet de navigation à gauche, sous Auto Scaling, choisissez Groupes Auto Scaling.
3. Sélectionnez Configurations de lancement en haut de la page. Lorsque vous êtes invité à confirmer, choisissez Afficher les configurations de lancement pour confirmer que vous souhaitez consulter la page Configurations de lancement.
4. Sélectionnez la configuration de lancement que vous souhaitez copier et choisissez Copy to launch template, Copy selected (Copier vers le modèle de lancement, Copier la sélection). Cette action configure un nouveau modèle de lancement avec le même nom et les mêmes options que la configuration de lancement que vous avez sélectionnée.
5. Pour le New launch template name (Nom du nouveau modèle de lancement), vous pouvez utiliser le nom de la configuration de lancement (par défaut) ou saisir un nouveau nom. Les noms des modèles de lancement doivent être uniques.
6. (Facultatif) Sélectionnez Créer un groupe Auto Scaling à l'aide du nouveau modèle.

   Vous pouvez passer cette étape pour terminer de copier la configuration de lancement. Vous n'avez pas besoin de créer un groupe Auto Scaling.
7. Choisissez Copier.

Pour copier toutes les configurations de lancement dans les modèles de lancement (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Dans le panneau de navigation, sous Auto Scaling, choisissez Configurations de lancement.
3. Choisissez Copier vers le modèle de lancement, Copier tout. Cette opération copie chaque configuration de lancement de la région actuelle vers un nouveau modèle de lancement avec le même nom et les mêmes options.
4. Choisissez Copier.

Étape 3 : Mettre à jour un groupe Auto Scaling pour utiliser un modèle de lancement

Après avoir créé un modèle de lancement, vous êtes prêt à l'ajouter à votre groupe Auto Scaling.

Pour mettre à jour un groupe Auto Scaling pour utiliser un modèle de lancement (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.

2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre dans la partie inférieure de la page avec des informations sur le groupe sélectionné.
3. Sous l'onglet Details (Détails), choisissez Launch configuration (Configuration du lancement), Edit (Modifier).
4. Choisissez Switch to launch template (Basculer vers un modèle de lancement).
5. Pour Launch Template (Modèle de lancement), sélectionnez votre modèle de lancement.
6. Pour Version, sélectionnez la version appropriée du modèle de lancement. Après avoir créé des versions d'un modèle de lancement, vous pouvez indiquer si le groupe Auto Scaling utilise la version par défaut ou la version la plus récente du modèle de lancement lors de l'augmentation.
7. Choisissez Mettre à jour.

Pour mettre à jour un groupe Auto Scaling pour utiliser un modèle de lancement (AWS CLI)

La commande update-auto-scaling-group suivante met à jour le groupe Auto Scaling spécifié pour utiliser la version initiale du modèle de lancement spécifié.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-template LaunchTemplateName=my-template-for-auto-scaling,Version='1'
```

Pour obtenir plus d'exemples d'utilisation de commandes CLI pour mettre à jour un groupe Auto Scaling afin d'utiliser un modèle de lancement, consultez Mettre à jour un groupe Auto Scaling pour utiliser un modèle de lancement.

Étape 4 : Remplacer vos instances

Une fois que vous avez remplacé la configuration de lancement par un modèle de lancement, toutes les nouvelles instances utiliseront le nouveau modèle de lancement. Les instances existantes ne sont pas affectées.

Pour mettre à jour les instances existantes, vous pouvez démarrer une actualisation d'instance pour remplacer les instances de votre groupe Auto Scaling plutôt que de remplacer manuellement quelques instances à la fois. Pour plus d'informations, consultez Utiliser une actualisation d'instance pour mettre à jour les instances d'un groupe Auto Scaling. Si le groupe est important, une actualisation de l'instance peut être particulièrement utile.

Vous pouvez également autoriser la mise à l'échelle automatique pour remplacer progressivement les instances existantes par de nouvelles instances en fonction des politiques de résiliation du groupe, ou vous pouvez les résilier. La résiliation manuelle oblige votre groupe Auto Scaling à lancer de nouvelles instances pour maintenir la capacité souhaitée du groupe. Pour plus d'informations, consultez la section Résiliation d'une instance dans le guide de l'utilisateur Amazon EC2.

Informations supplémentaires

Pour plus d'informations, consultez Amazon EC2 Auto Scaling n'ajoutera plus la prise en charge des nouvelles fonctionnalités EC2 aux configurations de lancement sur le AWS Compute Blog.

Pour consulter une rubrique expliquant comment migrer des AWS CloudFormation piles depuis des configurations de lancement vers des modèles de lancement, consultezMigrer AWS CloudFormation les piles vers les modèles de lancement.

Migrer AWS CloudFormation les piles vers les modèles de lancement

Vous pouvez migrer vos modèles de AWS CloudFormation stack existants des configurations de lancement vers les modèles de lancement. Pour ce faire, ajoutez un modèle de lancement directement à un modèle de pile existant, puis associez le modèle de lancement au groupe Auto Scaling dans le modèle de pile. Utilisez ensuite le modèle modifié pour mettre à jour votre pile.

Lors de la migration vers des modèles de lancement, cette rubrique vous permet de gagner du temps en fournissant des instructions pour réécrire les configurations de lancement dans vos modèles de CloudFormation pile en tant que modèles de lancement. Pour plus d'informations sur la migration des configurations de lancement vers des modèles de lancement, consultez Migrez vos groupes Auto Scaling pour lancer des modèles.

Rubriques

- Trouver les groupes Auto Scaling qui utilisent une configuration de lancement
- Mettre à jour une pile pour utiliser un modèle de lancement
- Comprendre les comportements de mise à jour des ressources d'une pile
- Suivre la migration
- Référence de mappage de la configuration du lancement

Trouver les groupes Auto Scaling qui utilisent une configuration de lancement

Pour trouver les groupes Auto Scaling qui utilisent une configuration de lancement

- Utilisez la commande describe-auto-scaling-groups suivante pour répertorier les noms des groupes Auto Scaling qui utilisent des configurations de lancement dans la région spécifiée. Incluez l'`--filters`option permettant de restreindre les résultats aux groupes associés à une CloudFormation pile (en filtrant par la clé de `aws:cloudformation:stack-name` balise).

  ```
  aws autoscaling describe-auto-scaling-groups --region REGION \
    --filters Name=tag-key,Values=aws:cloudformation:stack-name \
    --query 'AutoScalingGroups[?LaunchConfigurationName!
  =`null`].AutoScalingGroupName'
  ```

  Voici un exemple de sortie.

  ```
  [
      "{stack-name}-group-1",
      "{stack-name}-group-2",
      "{stack-name}-group-3"
  ]
  ```

  Vous pouvez trouver d'autres AWS CLI commandes utiles pour trouver des groupes Auto Scaling dans lesquels migrer et filtrer la sortieMigrez vos groupes Auto Scaling pour lancer des modèles.

> **Important**
>
> Si les ressources de votre pile ont `AWSEB` dans leur nom, cela signifie qu'elles ont été créées via AWS Elastic Beanstalk. Dans ce cas, vous devez mettre à jour l'environnement Beanstalk pour demander à Elastic Beanstalk de supprimer la configuration de lancement et la remplacer par un modèle de lancement.

Mettre à jour une pile pour utiliser un modèle de lancement

Suivez les étapes de cette section afin d'effectuer les étapes suivantes :

- Réécrivez la configuration de lancement en tant que modèle de lancement en utilisant les propriétés de modèle de lancement équivalentes.
- Associer le nouveau modèle de lancement avec le groupe Auto Scaling.
- Déployez ces mises à jour.

Pour modifier le modèle de pile et mettre à jour la pile

1. Suivez les mêmes procédures générales pour modifier le modèle de pile décrites dans la section Modification d'un modèle de pile du Guide de l'utilisateur AWS CloudFormation .
2. Réécrivez la configuration de lancement en tant que modèle de lancement. Consultez l'exemple suivant:

   Exemple : Configuration de lancement simple

```
---
Resources:
  myLaunchConfig:
    Type: AWS::AutoScaling::LaunchConfiguration
    Properties:
      ImageId: ami-02354e95b3example
      InstanceType: t3.micro
      SecurityGroups:
        - !Ref EC2SecurityGroup
      KeyName: MyKeyPair
      BlockDeviceMappings:
        - DeviceName: /dev/xvda
          Ebs:
            VolumeSize: 150
            DeleteOnTermination: true
      UserData:
        Fn::Base64: !Sub |
          #!/bin/bash -xe
          yum install -y aws-cfn-bootstrap
          /opt/aws/bin/cfn-signal -e $? --stack ${AWS::StackName} --resource myASG
 --region ${AWS::Region}
```

   Exemple : L'équivalent du modèle de lancement

```
---
Resources:
```

```
  myLaunchTemplate:
    Type: AWS::EC2::LaunchTemplate
    Properties:
      LaunchTemplateName: !Sub ${AWS::StackName}-launch-template
      LaunchTemplateData:
        ImageId: ami-02354e95b3example
        InstanceType: t3.micro
        SecurityGroupIds:
          - Ref! EC2SecurityGroup
        KeyName: MyKeyPair
        BlockDeviceMappings:
          - DeviceName: /dev/xvda
            Ebs:
              VolumeSize: 150
              DeleteOnTermination: true
        UserData:
          Fn::Base64: !Sub |
            #!/bin/bash -x
            yum install -y aws-cfn-bootstrap
            /opt/aws/bin/cfn-signal -e $? --stack ${AWS::StackName} --resource
myASG --region ${AWS::Region}
```

Pour obtenir des informations de référence sur toutes les propriétés prises en charge par Amazon EC2, consultez le guide AWS::EC2::LaunchTemplatede l'AWS CloudFormation utilisateur.

Notez que le modèle de lancement inclut la propriété `LaunchTemplateName` avec une valeur de `!Sub ${AWS::StackName}-launch-template`. Cela est nécessaire si vous souhaitez que le nom du modèle de lancement inclue le nom de la pile.

3. Si la propriété **`IamInstanceProfile`** est présente dans votre configuration de lancement, vous devez la convertir en structure et spécifier le nom ou l'ARN du profil d'instance. Pour obtenir un exemple, consultez AWS::EC2::LaunchTemplate.
4. Si les propriétés **`AssociatePublicIpAddress`**, **`InstanceMonitoring`**, ou **`PlacementTenancy`** sont présentes dans votre configuration de lancement, vous devez les convertir en structure. Pour obtenir des exemples, consultez AWS::EC2::LaunchTemplate.

   Il existe une exception lorsque la valeur de la propriété `MapPublicIpOnLaunch` sur les sous-réseaux que vous avez utilisés pour votre groupe Auto Scaling correspond à la valeur de la propriété `AssociatePublicIpAddress` dans votre configuration de lancement. Dans ce cas, vous pouvez ignorer la propriété `AssociatePublicIpAddress`. La propriété

`AssociatePublicIpAddress` est uniquement utilisée pour remplacer la propriété `MapPublicIpOnLaunch` afin de modifier si les instances reçoivent une adresse IPv4 publique au lancement.

5. Vous pouvez copier les groupes de sécurité de la propriété **`SecurityGroups`** vers l'un des deux emplacements de votre modèle de lancement. Normalement, vous copiez les groupes de sécurité dans la propriété `SecurityGroupIds`. Toutefois, si vous créez une structure `NetworkInterfaces` dans votre modèle de lancement pour spécifier la propriété `AssociatePublicIpAddress`, vous devez plutôt copier les groupes de sécurité dans la propriété `Groups` de l'interface réseau.
6. Si une ou plusieurs structures `BlockDeviceMapping` sont présentes dans votre configuration de lancement avec **`NoDevice`** défini sur `true`, vous devez spécifier une chaîne vide pour `NoDevice` dans votre modèle de lancement pour qu'Amazon EC2 omette le périphérique.
7. Si la propriété **`SpotPrice`** est présente dans votre configuration de lancement, nous vous recommandons de l'omettre de votre modèle de lancement. Votre instance Spot sera lancée au prix Spot en vigueur. Ce prix ne dépassera jamais le prix À la demande.

   Pour demander des instances Spot, vous disposez de deux options qui s'excluent mutuellement :

   - La première consiste à utiliser la structure `InstanceMarketOptions` de votre modèle de lancement (non recommandé). Pour plus d'informations, consultez AWS::EC2::LaunchTemplate InstanceMarketOptionsle guide de AWS CloudFormation l'utilisateur.
   - L'autre option consiste à ajouter une structure `MixedInstancesPolicy` à votre groupe Auto Scaling. Cela vous donne plus d'options quant à la manière dont vous faites la demande. Une demande d'instance Spot dans votre modèle de lancement prend en charge la sélection d'un seul type d'instance par groupe Auto Scaling. Cependant, une politique d'instances mixtes prend en charge la sélection de plusieurs types d'instances par groupe Auto Scaling. Les demandes d'instance Spot ont l'avantage d'offrir le choix entre plusieurs types d'instances. Pour plus d'informations, voir AWS::AutoScaling::AutoScaling MixedInstancesPolicyAWS::AutoScaling::AutoScalingGroupe dans le Guide de AWS CloudFormation l'utilisateur.
8. Supprimez la **`LaunchConfigurationName`** propriété de la ressource AWS::AutoScaling::AutoScalingAWS::AutoScaling::AutoScalingGroup . Ajoutez le modèle de lancement à sa place.

Dans les exemples suivants, la fonction intrinsèque Ref obtient l'ID de la AWS::EC2::LaunchTemplateressource avec l'ID logiquemyLaunchTemplate. La GetAttfonction obtient le dernier numéro de version (par exemple,1) du modèle de lancement de la `Version` propriété.

Exemple : Sans politique d'instances mixtes

```
---
Resources:
  myASG:
    Type: AWS::AutoScaling::AutoScalingGroup
    Properties:
      LaunchTemplate:
        LaunchTemplateId: !Ref myLaunchTemplate
        Version: !GetAtt myLaunchTemplate.LatestVersionNumber
      ...
```

Exemple : Avec une politique d'instances mixtes

```
---
Resources:
  myASG:
    Type: AWS::AutoScaling::AutoScalingGroup
    Properties:
      MixedInstancesPolicy:
        LaunchTemplate:
          LaunchTemplateSpecification:
            LaunchTemplateId: !Ref myLaunchTemplate
            Version: !GetAtt myLaunchTemplate.LatestVersionNumber
      ...
```

Pour obtenir des informations de référence sur toutes les propriétés prises en charge par Amazon EC2 Auto Scaling, consultez AWS::AutoScaling::AutoScaling dans le guide de l'AWS CloudFormation utilisateur.

9. Lorsque vous êtes prêt à déployer ces mises à jour, suivez les CloudFormation procédures pour mettre à jour la pile avec votre modèle de pile modifié. Pour plus d'informations, consultez Modification d'un modèle de pile dans le Guide de l'utilisateur AWS CloudFormation .

Comprendre les comportements de mise à jour des ressources d’une pile

CloudFormation met à jour les ressources de pile en comparant les modifications entre le modèle mis à jour que vous fournissez et les configurations de ressources que vous avez décrites dans la version précédente de votre modèle de pile. Les configurations de ressources qui n’ont pas changé restent inchangées pendant le processus de mise à jour.

CloudFormation prend en charge l'UpdatePolicyattribut pour les groupes Auto Scaling. Lors d'une mise à jour, si `UpdatePolicy` ce paramètre est défini sur`AutoScalingRollingUpdate`, CloudFormation remplace `InService` les instances une fois que vous avez effectué les étapes de cette procédure. S'il `UpdatePolicy` est défini sur`AutoScalingReplacingUpdate`, CloudFormation remplace le groupe Auto Scaling et son pool de chaleur (s'il en existe un).

Si vous n'avez pas spécifié d'`UpdatePolicy`attribut pour votre groupe Auto Scaling, l'exactitude du modèle de lancement est vérifiée, mais aucune modification CloudFormation n'est déployée sur les instances du groupe Auto Scaling. Toutes les nouvelles instances utilisent votre modèle de lancement, mais les instances existantes continuent à s’exécuter selon la configuration de lancement initiale (bien qu’elle n’existe pas). L’exception est lorsque vous modifiez vos options d’achat, par exemple en ajoutant une politique d’instances mixtes. Dans ce cas, votre groupe Auto Scaling remplace progressivement les instances existantes par de nouvelles instances correspondant aux nouvelles options d’achat.

Suivre la migration

Suivre la migration

1. Dans la console AWS CloudFormation, choisissez la pile que vous avez mise à jour et choisissez l'onglet Events afin d'afficher les événements de la pile.
2. Pour mettre à jour la liste des événements avec les événements les plus récents, cliquez sur le bouton d'actualisation de la CloudFormation console.
3. Pendant la mise à jour de votre pile, vous remarquerez plusieurs événements pour chaque mise à jour des ressources. Si vous voyez une exception dans la colonne Status reason indiquant un problème lors de la tentative de création du modèle de lancement, consultez Résoudre les problèmes d'Amazon EC2 Auto Scaling : modèles de lancement pour rechercher les causes potentielles.
4. (Facultatif) En fonction de votre utilisation de l’attribut `UpdatePolicy`, vous pouvez suivre la progression de votre groupe Auto Scaling depuis la page des groupes Auto Scaling de la console Amazon EC2. Sélectionnez le groupe Auto Scaling. Sous l’onglet Activité sous

Historique de l'activité, la colonne État indique si votre groupe Auto Scaling a réussi à lancer ou à résilier des instances, ou si l'activité de mise à l'échelle est toujours en cours.

5. Lorsque la mise à jour de la pile est terminée, CloudFormation émet un événement de `UPDATE_COMPLETE` pile. Pour de plus amples informations, veuillez consulter Surveillance de la progression d'une mise à jour de pile dans le Guide de l'utilisateur AWS CloudFormation .
6. Une fois la mise à jour de la pile terminée, ouvrez la page des modèles de lancement et la page des configurations de lancement de la console Amazon EC2. Vous remarquerez qu'un nouveau modèle de lancement est créé et que la configuration de lancement est supprimée.

Référence de mappage de la configuration du lancement

À des fins de référence, le tableau suivant répertorie toutes les propriétés de niveau supérieur de la AWS::AutoScaling::LaunchConfigurationressource avec leur propriété correspondante dans la AWS::EC2::LaunchTemplateressource.

Propriété source de la configuration de lancement	Propriété cible du modèle de lancement
`AssociatePublicIpAddress`	`NetworkInterfaces.AssociatePublicIpAddress`
`BlockDeviceMappings`	`BlockDeviceMappings`
`ClassicLinkVPCId`	Non disponible[1]
`ClassicLinkVPCSecurityGroups`	Non disponible[1]
`EbsOptimized`	`EbsOptimized`
`IamInstanceProfile`	Soit `IamInstanceProfile.Arn` soit `IamInstanceProfile.Name`, mais pas les deux.
`ImageId`	`ImageId`
`InstanceId`	`InstanceId`
`InstanceMonitoring`	`Monitoring.Enabled`

Propriété source de la configuration de lancement	Propriété cible du modèle de lancement
`InstanceType`	`InstanceType`
`KernelId`	`KernelId`
`KeyName`	`KeyName`
`LaunchConfigurationName`	`LaunchTemplateName`
`MetadataOptions`	`MetadataOptions`
`PlacementTenancy`	`Placement.Tenancy`
`RamDiskId`	`RamDiskId`
`SecurityGroups`	Soit `SecurityGroupIds` soit `NetworkInterfaces.Groups`, mais pas les deux.
`SpotPrice`	`InstanceMarketOptions.SpotOptions.MaxPrice`
`UserData`	`UserData`

[1] Les `ClassicLinkVPCSecurityGroups` propriétés `ClassicLinkVPCId` et ne peuvent pas être utilisées dans un modèle de lancement car EC2-Classic n'est plus disponible.

Exemples de création et de gestion de modèles de lancement à l'aide du AWS CLI

Vous pouvez créer et gérer des modèles de lancement via le AWS Management Console, AWS Command Line Interface (AWS CLI) ou les SDK. Cette section présente des exemples de création et de gestion de modèles de lancement pour Amazon EC2 Auto Scaling à partir du. AWS CLI

Table des matières

- Spécifier des balises qui balisent les instances au lancement
- Spécifier un rôle IAM à transmettre aux instances
- Attribuer des adresses IP publiques
- Spécifier un script de données utilisateur qui configure les instances au lancement
- Spécifier un mappage de périphérique de stockage en mode bloc
- Spécifier les hôtes dédiés pour obtenir des licences logicielles auprès de fournisseurs externes
- Spécifier une interface réseau existante
- Créer plusieurs interfaces réseau
- Gérer vos modèles de lancement
- Mettre à jour un groupe Auto Scaling pour utiliser un modèle de lancement

Exemple d'utilisation

```
{
    "LaunchTemplateName": "my-template-for-auto-scaling",
    "VersionDescription": "test description",
    "LaunchTemplateData": {
        "ImageId": "ami-04d5cc9b88example",
        "InstanceType": "t2.micro",
        "SecurityGroupIds": [
            "sg-903004f88example"
        ],
        "KeyName": "MyKeyPair",
        "Monitoring": {
            "Enabled": true
        },
        "Placement": {
            "Tenancy": "dedicated"
        },
        "CreditSpecification": {
            "CpuCredits": "unlimited"
        },
        "MetadataOptions": {
            "HttpTokens": "required",
            "HttpPutResponseHopLimit": 1,
            "HttpEndpoint": "enabled"
        }
    }
```

```
}
```

Créer un modèle de lancement de base

Pour créer un modèle de lancement de base, utilisez la commande create-launch-template comme suit, avec ces modifications :

- Remplacez `ami-04d5cc9b88example` par l'ID d'AMI à partir duquel lancer les instances.
- Remplacez `t2.micro` par un type d'instance compatible avec l'AMI que vous avez spécifiée.

Cet exemple crée un modèle de lancement avec le nom *my-template pour auto-scaling*. Si les instances créées par ce modèle de lancement sont lancées dans un VPC par défaut, elles reçoivent une adresse IP publique par défaut. Si les instances sont lancées sur un VPC personnalisé, elles ne reçoivent pas d'adresse IP publique par défaut.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data
 '{"ImageId":"ami-04d5cc9b88example","InstanceType":"t2.micro"}'
```

Pour plus d'informations sur l'utilisation de guillemets avec les paramètres au format JSON, consultez Utilisation de guillemets avec des chaînes dans l' AWS CLI dans le Guide de l'utilisateur AWS Command Line Interface .

Sinon, vous pouvez spécifier les paramètres au format JSON dans un fichier de configuration.

L'exemple suivant crée un modèle de lancement de base, référençant un fichier de configuration pour les valeurs de paramètre de modèle de lancement.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data file://config.json
```

Contenu de `config.json` :

```
{
    "ImageId":"ami-04d5cc9b88example",
    "InstanceType":"t2.micro"
}
```

Spécifier des balises qui balisent les instances au lancement

L'exemple suivant ajoute une balise (par exemple,`purpose=webserver`) aux instances au lancement.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data '{"TagSpecifications":[{"ResourceType":"instance","Tags":
[{"Key":"purpose","Value":"webserver"}]}],"ImageId":"ami-04d5cc9b88example","InstanceType":"t2.
```

> **Note**
>
> Si vous spécifiez des balises d'instance dans votre modèle de lancement, puis que vous avez choisi de propager les balises de votre groupe Auto Scaling à ses instances, toutes les balises sont fusionnées. Si la même clé d'identification est spécifiée pour une identification dans votre modèle de lancement et une identification dans votre groupe Auto Scaling, la valeur de l'identification du groupe est prioritaire.

Spécifier un rôle IAM à transmettre aux instances

L'exemple suivant montre comment spécifier le nom du profil d'instance associé au rôle IAM à transmettre aux instances lors du lancement. Pour plus d'informations, consultez Rôle IAM pour les applications qui s'exécutent sur des instances Amazon EC2.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
--launch-template-data '{"IamInstanceProfile":{"Name":"my-instance-
profile"},"ImageId":"ami-04d5cc9b88example","InstanceType":"t2.micro"}'
```

Attribuer des adresses IP publiques

L'exemple create-launch-template suivant configure le modèle de lancement pour affecter des adresses publiques aux instances lancées dans un VPC autre que le VPC par défaut.

> **Note**
>
> Lorsque vous spécifiez une interface réseau, spécifiez une valeur pour `Groups` correspondant aux groupes de sécurité du VPC dans lequel votre groupe Auto Scaling

lancera des instances. Spécifiez les sous-réseaux VPC en tant que propriétés du groupe Auto Scaling.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data '{"NetworkInterfaces":
[{"DeviceIndex":0,"AssociatePublicIpAddress":true,"Groups":
["sg-903004f88example"],"DeleteOnTermination":true}],"ImageId":"ami-04d5cc9b88example","Instanc
```

Spécifier un script de données utilisateur qui configure les instances au lancement

L'exemple suivant spécifie un script de données utilisateur sous la forme d'une chaîne codée en base64 qui configure les instances au lancement. La commande create-launch-template nécessite des données utilisateur encodées en base64.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
--launch-template-data
 '{"UserData":"IyEvYmluL2Jhc...","ImageId":"ami-04d5cc9b88example","InstanceType":"t2.micro"}'
```

Spécifier un mappage de périphérique de stockage en mode bloc

L'exemple create-launch-template suivant crée un modèle de lancement avec un mappage de périphérique de stockage en mode bloc : un volume EBS de 22 gigaoctets mappé à `/dev/xvdcz`. Le volume `/dev/xvdcz` utilise le type de volume SSD à usage général (gp2) et est supprimé lors de la résiliation de l'instance à laquelle il est attaché.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data '{"BlockDeviceMappings":[{"DeviceName":"/dev/xvdcz","Ebs":
{"VolumeSize":22,"VolumeType":"gp2","DeleteOnTermination":true}}],"ImageId":"ami-04d5cc9b88exam
```

Spécifier les hôtes dédiés pour obtenir des licences logicielles auprès de fournisseurs externes

Si vous spécifiez host (hôte), vous pouvez spécifier un groupe de ressources hôte et une configuration de licence License Manager pour obtenir des licences logicielles éligibles auprès de fournisseurs externes. Ensuite, vous pouvez utiliser les licences sur les instances EC2 à l'aide de la commande create-launch-template suivante.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data '{"Placement":
{"Tenancy":"host","HostResourceGroupArn":"arn"},"LicenseSpecifications":
[{"LicenseConfigurationArn":"arn"}],"ImageId":"ami-04d5cc9b88example","InstanceType":"t2.micro"
```

Spécifier une interface réseau existante

L'exemple create-launch-template suivant configure l'interface réseau principale de façon à utiliser une interface réseau existante.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data '{"NetworkInterfaces":
[{"DeviceIndex":0,"NetworkInterfaceId":"eni-
b9a5ac93","DeleteOnTermination":false}],"ImageId":"ami-04d5cc9b88example","InstanceType":"t2.mi
```

Créer plusieurs interfaces réseau

L'exemple create-launch-template suivant ajoute une interface réseau secondaire. L'index de périphérique principal est 0 pour l'interface réseau principale, et l'index de périphérique secondaire est 1 pour l'interface réseau secondaire.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data '{"NetworkInterfaces":[{"DeviceIndex":0,"Groups":
["sg-903004f88example"],"DeleteOnTermination":true},{"DeviceIndex":1,"Groups":
["sg-903004f88example"],"DeleteOnTermination":true}],"ImageId":"ami-04d5cc9b88example","Instanc
```

Si vous utilisez un type d'instance qui prend en charge plusieurs cartes réseau et des adaptateurs Elastic Fabric (EFA), vous pouvez ajouter une interface secondaire à une carte réseau secondaire et

activer EFA à l'aide de la commande create-launch-template. Pour plus d'informations, consultez la section Ajouter un EFA à un modèle de lancement dans le guide de l'utilisateur Amazon EC2.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data '{"NetworkInterfaces":
[{"NetworkCardIndex":0,"DeviceIndex":0,"Groups":
["sg-7c2270198example"],"InterfaceType":"efa","DeleteOnTermination":true},
{"NetworkCardIndex":1,"DeviceIndex":1,"Groups":
["sg-7c2270198example"],"InterfaceType":"efa","DeleteOnTermination":true}],"ImageId":"ami-09d95
```

Warning

Le type d'instance p4d.24xlarge entraîne des coûts plus élevés que les autres exemples de cette section. Pour plus d'informations sur la tarification des instances P4d, consultez Tarification des instances P4d Amazon EC2.

Note

L'attachement de plusieurs interfaces réseau du même sous-réseau à une instance peut introduire un routage asymétrique, en particulier sur les instances utilisant une variante Linux non Amazon. Si vous avez besoin de ce type de configuration, vous devez configurer l'interface réseau secondaire dans le système d'exploitation. Par exemple, consultez Comment puis-je faire fonctionner mon interface réseau secondaire dans mon instance Ubuntu EC2 ? dans le AWS Knowledge Center.

Gérer vos modèles de lancement

AWS CLI Il inclut plusieurs autres commandes qui vous aident à gérer vos modèles de lancement.

Table des matières

Lister et décrire vos modèles de lancement

Vous pouvez utiliser deux AWS CLI commandes pour obtenir des informations sur vos modèles de lancement : describe-launch-templates et describe-launch-template-versions.

La commande describe-launch-templates vous permet d'obtenir la liste des modèles de lancement que vous avez créés. Vous pouvez utiliser une option pour filtrer les résultats sur un nom de modèle de lancement, créer une heure, une clé de balise ou une combinaison clé-valeur de balise. Cette commande renvoie des informations récapitulatives sur l'un de vos modèles de lancement, y compris l'identifiant du modèle de lancement, la dernière version et la version par défaut.

L'exemple suivant fournit un résumé du modèle de lancement indiqué.

```
aws ec2 describe-launch-templates --launch-template-names my-template-for-auto-scaling
```

Voici un exemple de réponse.

```
{
    "LaunchTemplates": [
        {
            "LaunchTemplateId": "lt-068f72b729example",
            "LaunchTemplateName": "my-template-for-auto-scaling",
            "CreateTime": "2020-02-28T19:52:27.000Z",
            "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
            "DefaultVersionNumber": 1,
            "LatestVersionNumber": 1
        }
    ]
}
```

Si vous n'utilisez pas le paramètre `--launch-template-names` pour limiter la sortie à un seul modèle de lancement, les informations sur tous vos modèles de lancement sont renvoyées.

La commande describe-launch-template-versions suivante fournit des informations décrivant les versions du modèle de lancement spécifié.

```
aws ec2 describe-launch-template-versions --launch-template-id lt-068f72b729example
```

Voici un exemple de réponse.

```
{
    "LaunchTemplateVersions": [
        {
            "VersionDescription": "version1",
            "LaunchTemplateId": "lt-068f72b729example",
            "LaunchTemplateName": "my-template-for-auto-scaling",
            "VersionNumber": 1,
            "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
            "LaunchTemplateData": {
                "TagSpecifications": [
                    {
                        "ResourceType": "instance",
                        "Tags": [
                            {
                                "Key": "purpose",
                                "Value": "webserver"
                            }
                        ]
                    }
                ],
                "ImageId": "ami-04d5cc9b88example",
                "InstanceType": "t2.micro",
                "NetworkInterfaces": [
                    {
                        "DeviceIndex": 0,
                        "DeleteOnTermination": true,
                        "Groups": [
                            "sg-903004f88example"
                        ],
                        "AssociatePublicIpAddress": true
                    }
                ]
            },
            "DefaultVersion": true,
            "CreateTime": "2020-02-28T19:52:27.000Z"
        }
    ]
}
```


Créer une version d’un modèle de lancement

La commande create-launch-template version suivante crée une nouvelle version du modèle de lancement basée sur la version 1 du modèle de lancement et spécifie un autre ID d'AMI.

```
aws ec2 create-launch-template-version --launch-template-id lt-068f72b729example --
version-description version2 \
  --source-version 1 --launch-template-data "ImageId=ami-c998b6b2example"
```

Pour définir la version par défaut du modèle de lancement, utilisez la commande modify-launch-template.

Supprimer une version d'un modèle de lancement

La commande delete-launch-template-versions supprime la version du modèle de lancement spécifiée.

```
aws ec2 delete-launch-template-versions --launch-template-id lt-068f72b729example --
versions 1
```

Supprimer un modèle de lancement

Si vous n'avez plus besoin d'un modèle de lancement, vous pouvez le supprimer à l'aide de la commande delete-launch-template. La suppression d'un modèle de lancement entraîne celle de toutes ses versions.

```
aws ec2 delete-launch-template --launch-template-id lt-068f72b729example
```

Mettre à jour un groupe Auto Scaling pour utiliser un modèle de lancement

Vous pouvez utiliser la commande update-auto-scaling-group pour ajouter un modèle de lancement à un groupe Auto Scaling existant.

Mettre à jour un groupe Auto Scaling pour utiliser la dernière version d'un modèle de lancement

La commande update-auto-scaling-group suivante met à jour le groupe Auto Scaling spécifié pour utiliser la dernière version du modèle de lancement spécifié.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-template LaunchTemplateId=lt-068f72b729example,Version='$Latest'
```

Mettre à jour un groupe Auto Scaling pour utiliser une version spécifique d'un modèle de lancement

La commande update-auto-scaling-group suivante met à jour le groupe Auto Scaling spécifié pour utiliser une version spécifique du modèle de lancement spécifié.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-template LaunchTemplateName=my-template-for-auto-scaling,Version='2'
```

Utiliser des AWS Systems Manager paramètres plutôt que des ID d'AMI dans les modèles de lancement

Cette section explique comment créer un modèle de lancement qui spécifie un AWS Systems Manager paramètre faisant référence à un identifiant Amazon Machine Image (AMI). Vous pouvez utiliser un paramètre stocké dans le vôtre Compte AWS, un paramètre partagé par un autre Compte AWS ou un paramètre public pour une AMI publique gérée par AWS.

Les paramètres Systems Manager vous permettent de mettre à jour vos groupes Auto Scaling pour utiliser de nouveaux ID d'AMI sans avoir à créer de nouveaux modèles de lancement ou de nouvelles versions de modèles de lancement chaque fois qu'un ID d'AMI change. Ces ID peuvent changer régulièrement, notamment lorsqu'une AMI reçoit des mises à jour logicielles ou du système d'exploitation.

Vous pouvez créer, mettre à jour ou supprimer vos propres paramètres de Systems Manager à l'aide du Parameter Store, une fonctionnalité de AWS Systems Manager. Vous devez créer un paramètre Systems Manager avant de pouvoir l'utiliser dans un modèle de lancement. Pour commencer, créez un paramètre avec le type de données `aws:ec2:image`, et saisissez pour sa valeur l'ID d'une AMI. L'ID d'AMI se présente sous la forme `ami-<identifier>`, par exemple, `ami-123example456`. L'ID d'AMI correct dépend du type d'instance et de la Région AWS dans laquelle vous lancez le groupe Auto Scaling.

Pour plus d'informations sur la création d'un paramètre valide pour un ID d'AMI, consultez la section Création des paramètres de Systems Manager.

Créez un modèle de lancement qui spécifie un paramètre pour l'AMI

Pour créer un modèle de lancement qui spécifie un paramètre pour l'AMI, appliquez l'une des méthodes suivantes :

Console

Pour créer un modèle de lancement à l'aide d'un AWS Systems Manager paramètre

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Dans le panneau de navigation, choisissez Modèles de lancement, puis Créer un modèle de lancement.
3. Pour Nom du modèle de lancement, entrez un nom descriptif pour le modèle.
4. Sous Application and OS Images (Amazon Machine Image) (Images d'applications et de systèmes d'exploitation (Amazon Machine Image)), choisissez Browse more AMIs (Parcourir plus d'AMI).
5. Sélectionnez le bouton fléché à droite de la barre de recherche, puis choisissez Spécifier une valeur personnalisée/un paramètre Systems Manager.
6. Dans la boîte de dialogue Spécifier une valeur personnalisée ou un paramètre Systems Manager, procédez comme suit :

   a. Pour ID d'AMI ou chaîne de paramètres Systems Manager, saisissez le nom du paramètre Systems Manager en utilisant l'un des formats suivants :

      Pour référencer un paramètre public :

      - **resolve:ssm:*public-parameter***

      Pour référencer un paramètre stocké dans le même compte :

      - **resolve:ssm:*parameter-name***
      - **resolve:ssm:*parameter-name*:*version-number***
      - **resolve:ssm:*parameter-name*:*label***

      Pour référencer un paramètre partagé par un autre Compte AWS :

      - **resolve:ssm:*parameter-ARN***
      - **resolve:ssm:*parameter-ARN*:*version-number***
      - **resolve:ssm:*parameter-ARN*:*label***

   b. Choisissez Enregistrer.

7. Configurez tout autre paramètre de modèle de lancement selon vos besoins, puis choisissez Créer un modèle de lancement. Pour plus d'informations, consultez Créer un modèle de lancement pour un groupe Auto Scaling.

AWS CLI

Pour créer un modèle de lancement qui spécifie un paramètre de Systems Manager, vous pouvez utiliser l'un des exemples de commandes suivants. Remplacez chaque *espace réservé à la saisie de l'utilisateur* par vos propres informations.

Exemple : créer un modèle de lancement qui spécifie un paramètre public AWS appartenant à l'utilisateur

Utilisez la syntaxe suivante : `resolve:ssm:`*`public-parameter`*, où `resolve:ssm` est le préfixe standard et *`public-parameter`* le chemin et le nom du paramètre public.

Dans cet exemple, le modèle de lancement utilise un paramètre public AWS fourni pour lancer des instances à l'aide de la dernière AMI Amazon Linux 2 configurée pour votre profil. Région AWS

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling
 --version-description version1 \
  --launch-template-data file://config.json
```

Contenu de `config.json` :

```
{
    "ImageId":"resolve:ssm:/aws/service/ami-amazon-linux-latest/amzn2-ami-hvm-
x86_64-gp2",
    "InstanceType":"t2.micro"
}
```

Voici un exemple de réponse.

```
{
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-089c023a30example",
        "LaunchTemplateName": "my-template-for-auto-scaling",
        "CreateTime": "2022-12-28T19:52:27.000Z",
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
        "DefaultVersionNumber": 1,
```

```
        "LatestVersionNumber": 1
    }
}
```

Exemple : créer un modèle de lancement qui spécifie un paramètre stocké dans le même compte

Utilisez la syntaxe suivante : `resolve:ssm:`*`parameter-name`*, où `resolve:ssm` est le préfixe standard et *`parameter-name`* le nom du paramètre Systems Manager.

L'exemple suivant crée un modèle de lancement qui obtient l'ID d'AMI à partir d'un paramètre Systems Manager existant nommé *`golden-ami`*.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling \
  --launch-template-data file://config.json
```

Contenu de `config.json` :

```
{
    "ImageId":"resolve:ssm:golden-ami",
    "InstanceType":"t2.micro"
}
```

La version par défaut du paramètre, si elle n'est pas spécifiée, est la dernière version.

L'exemple suivant référence une version spécifique du paramètre *`golden-ami`*. L'exemple utilise la version *`3`* du paramètre *`golden-ami`*, mais vous pouvez utiliser n'importe quel numéro de version valide.

```
{
    "ImageId":"resolve:ssm:golden-ami:3",
    "InstanceType":"t2.micro"
}
```

L'exemple similaire suivant référence l'étiquette de paramètre *`prod`* qui est liée à une version spécifique du paramètre *`golden-ami`*.

```
{
    "ImageId":"resolve:ssm:golden-ami:prod",
    "InstanceType":"t2.micro"
}
```

Voici un exemple de sortie.

```
{
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-068f72b724example",
        "LaunchTemplateName": "my-template-for-auto-scaling",
        "CreateTime": "2022-12-27T17:11:21.000Z",
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
        "DefaultVersionNumber": 1,
        "LatestVersionNumber": 1
    }
}
```

Exemple : créer un modèle de lancement qui spécifie un paramètre partagé par un autre Compte AWS

Utilisez la syntaxe suivante :`resolve:ssm:`*`parameter-ARN`*, où `resolve:ssm` se trouvent le préfixe standard et *`parameter-ARN`* l'ARN du paramètre Systems Manager.

L'exemple suivant crée un modèle de lancement qui obtient l'ID de l'AMI à partir d'un paramètre Systems Manager existant avec l'ARN de*`arn:aws:ssm:us-east-2:123456789012:parameter/MyParameter`*.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling
 --version-description version1 \
  --launch-template-data file://config.json
```

Contenu de `config.json` :

```
{
    "ImageId":"resolve:ssm:arn:aws:ssm:us-east-2:123456789012:parameter/
MyParameter",
    "InstanceType":"t2.micro"
}
```

La version par défaut du paramètre, si elle n'est pas spécifiée, est la dernière version.

L'exemple suivant référence une version spécifique du paramètre *`MyParameter`*. L'exemple utilise la version *`3`* du paramètre *`MyParameter`*, mais vous pouvez utiliser n'importe quel numéro de version valide.

```
{
    "ImageId":"resolve:ssm:arn:aws:ssm:us-east-2:123456789012:parameter/
MyParameter:3",
    "InstanceType":"t2.micro"
}
```

L'exemple similaire suivant référence l'étiquette de paramètre *prod* qui est liée à une version spécifique du paramètre *MyParameter*.

```
{
    "ImageId":"resolve:ssm:arn:aws:ssm:us-east-2:123456789012:parameter/
MyParameter:prod",
    "InstanceType":"t2.micro"
}
```

Voici un exemple de réponse.

```
{
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-00f93d4588example",
        "LaunchTemplateName": "my-template-for-auto-scaling",
        "CreateTime": "2024-01-08T12:43:21.000Z",
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
        "DefaultVersionNumber": 1,
        "LatestVersionNumber": 1
    }
}
```

Pour spécifier un paramètre depuis le magasin de paramètres dans un modèle de lancement, vous devez disposer de l'`ssm:GetParameters`autorisation pour le paramètre spécifié. Toute personne utilisant le modèle de lancement doit également disposer d'une `ssm:GetParameters` autorisation pour que la valeur du paramètre soit validée. Pour plus d'informations, consultez la section Restreindre l'accès aux paramètres de Systems Manager à l'aide de politiques IAM dans le Guide de AWS Systems Manager l'utilisateur.

Vérifiez qu'un modèle de lancement obtient le bon ID d'AMI

Utilisez la commande describe-launch-template-versions et incluez l'`--resolve-alias`option permettant de convertir le paramètre en ID d'AMI réel.

```
aws ec2 describe-launch-template-versions --launch-template-name my-template-for-auto-scaling \
   --versions $Default --resolve-alias
```

L'exemple renvoie l'ID d'AMI pour `ImageId`. Lorsqu'une instance est lancée à l'aide de ce modèle de lancement, l'ID d'AMI est défini sur `ami-0ac394d6a3example`.

```
{
    "LaunchTemplateVersions": [
        {
            "LaunchTemplateId": "lt-089c023a30example",
            "LaunchTemplateName": "my-template-for-auto-scaling",
            "VersionNumber": 1,
            "CreateTime": "2022-12-28T19:52:27.000Z",
            "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
            "DefaultVersion": true,
            "LaunchTemplateData": {
                "ImageId": "ami-0ac394d6a3example",
                "InstanceType": "t2.micro",
            }
        }
    ]
}
```

Ressources connexes

Pour plus de détails sur la spécification d'un paramètre Systems Manager dans votre modèle de lancement, consultez Utiliser un paramètre Systems Manager au lieu d'un ID AMI dans le guide de l'utilisateur Amazon EC2.

Pour plus d'informations sur l'utilisation des paramètres de Systems Manager, consultez les documents de référence suivants dans la documentation de Systems Manager.

- Pour créer des versions de paramètres et des étiquettes, consultez les sections Utilisation des versions de paramètres et Utilisation des étiquettes de paramètres.
- Pour plus d'informations sur la façon de rechercher les paramètres publics de l'AMI pris en charge par Amazon EC2, consultez la section Appeler les paramètres publics de l'AMI.
- Pour plus d'informations sur le partage de paramètres avec d'autres AWS comptes ou via d'autres comptes AWS Organizations, consultez la section Utilisation de paramètres partagés.

- Pour plus d'informations sur le suivi de la création réussie de vos paramètres, consultez la section Prise en charge des paramètres natifs pour les Amazon Machine Image ID.

Limites

Lorsque vous travaillez avec les paramètres de Systems Manager, tenez compte des limites suivantes :

- Amazon EC2 Auto Scaling prend uniquement en charge la spécification d'ID d'AMI comme paramètres.
- La création ou la mise à jour de groupes d'instances mixtes à l'aide d'un modèle de lancement spécifiant un paramètre de Systems Manager n'est actuellement pas prise en charge.
- Si votre groupe Auto Scaling utilise un modèle de lancement qui spécifie un paramètre de Systems Manager, vous ne serez pas en mesure de démarrer une actualisation d'instance avec la configuration souhaitée ou en utilisant la correspondance des sauts.
- À chaque appel visant à créer ou à mettre à jour votre groupe Auto Scaling, Amazon EC2 Auto Scaling définira le paramètre Systems Manager dans le modèle de lancement. Si vous utilisez des paramètres avancés ou des limites de débit plus élevées, les appels fréquents au Parameter Store (c'est-à-dire l'opération `GetParameters`) peuvent augmenter les coûts de Systems Manager, car des frais sont facturés par interaction avec l'API Parameter Store. Pour en savoir plus, consultez AWS Systems Manager Tarification.

Configurations de lancement

> **Important**
>
> Vous ne pouvez pas appeler `CreateLaunchConfiguration` avec les nouveaux types d'instances Amazon EC2 publiés après le 31 décembre 2022. De plus, les nouveaux comptes créés le 1er juin 2023 ou après cette date n'auront pas la possibilité de créer de nouvelles configurations de lancement via la console. À l'avenir, les nouveaux comptes ne pourront pas créer de nouvelles configurations de lancement à l'aide de la console, de l'API, de la CLI et CloudFormation. Migrez vers des modèles de lancement pour vous assurer de ne pas avoir à créer de nouvelles configurations de lancement maintenant ou à l'avenir. Pour plus d'informations sur la migration de vos groupes Auto Scaling vers les modèles de lancement, consultez la section Migrez vos groupes Auto Scaling pour lancer des modèles.

Une configuration du lancement est un modèle de configuration d'instance utilisé par un groupe Auto Scaling pour lancer des instances EC2. Lorsque vous créez une configuration du lancement, fournissez les informations relatives aux instances. Indiquez l'ID d'Amazon Machine Image (AMI), le type d'instance, une paire de clés, un ou plusieurs groupes de sécurité et un mappage de périphérique de stockage en mode bloc. Si vous avez lancé une instance EC2 auparavant, vous avez spécifié les mêmes informations pour la lancer.

Vous pouvez spécifier la configuration du lancement avec plusieurs groupes Auto Scaling. Cependant, vous pouvez uniquement spécifier une seule configuration du lancement pour un groupe Auto Scaling à la fois, et vous ne pouvez pas la modifier après l'avoir créée. Pour modifier la configuration du lancement d'un groupe Auto Scaling, vous devez créer une configuration du lancement, puis mettre à jour votre groupe Auto Scaling avec cette configuration.

Table des matières

Créez une configuration de lancement

> **Important**
>
> Vous ne pouvez pas appeler `CreateLaunchConfiguration` avec les nouveaux types d'instances Amazon EC2 publiés après le 31 décembre 2022. De plus, les nouveaux comptes créés le 1er juin 2023 ou après cette date n'auront pas la possibilité de créer de nouvelles configurations de lancement via la console. À l'avenir, les nouveaux comptes ne pourront pas créer de nouvelles configurations de lancement à l'aide de la console, de l'API, de la CLI et CloudFormation. Migrez vers des modèles de lancement pour vous assurer de ne pas avoir à créer de nouvelles configurations de lancement maintenant ou à l'avenir. Pour plus d'informations sur la migration de vos groupes Auto Scaling vers les modèles de lancement, consultez la section Migrez vos groupes Auto Scaling pour lancer des modèles.

Cette rubrique explique comment créer une configuration de lancement.

Une fois que vous avez créé une configuration de lancement, vous ne pouvez pas la modifier. Vous devez plutôt créer une nouvelle configuration de lancement.

Pour associer une nouvelle configuration de lancement à un groupe Auto Scaling existant, consultezModifier la configuration du lancement pour un groupe Auto Scaling. Pour créer un nouveau groupe Auto Scaling, consultezCréer un groupe Auto Scaling à l'aide d'une configuration du lancement.

Table des matières

Créez une configuration de lancement

Pour créer une configuration du lancement (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Dans la barre de navigation supérieure, sélectionnez votre AWS région.
3. Dans le volet de navigation à gauche, sous Auto Scaling, choisissez Groupes Auto Scaling.

4. Sélectionnez Configurations de lancement en haut de la page. Lorsque vous êtes invité à confirmer, choisissez Afficher les configurations de lancement pour confirmer que vous souhaitez consulter la page Configurations de lancement.
5. Choisissez Create launch configuration (Créer une configuration du lancement) et entrez un nom pour votre configuration du lancement.
6. Choisissez une AMI dans Amazon Machine Image (AMI). Pour trouver une AMI spécifique, vous pouvez rechercher une AMI appropriée, noter son ID et entrer l'ID comme critère de recherche.

   Pour obtenir l'ID de l'AMI Amazon Linux 2 :

   a. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
   b. Dans le volet de navigation à gauche, sous Instances, choisissez Instances, puis choisissez Lancer des instances.
   c. Dans la page Quick Start (Démarrage rapide) de la page Choose an Amazon Machine Image (Sélection d'une Amazon Machine Image (AMI)), notez l'ID de l'AMI en regard d'Amazon Linux 2 AMI (HVM) (AMI Amazon Linux 2 (HVM)).
7. Pour Instance type (Type d'instance), sélectionnez une configuration matérielle pour vos instances.
8. Sous Additional configuration (Configuration supplémentaire), prêtez attention aux champs suivants :

   a. (Facultatif) Pour Purchasing option (Option d'achat), vous pouvez choisir Request Spot Instances (Demander des instances Spot) pour demander des Instances Spot au prix Spot, plafonné au prix des instances à la demande. Le cas échéant, vous pouvez spécifier un prix maximum par heure d'instance pour les instances Spot.

      > **Note**
      >
      > Les instances Spot constituent un choix économique comparées aux instances à la demande, si vous êtes flexible quant au moment où vos applications s'exécutent et à la possibilité qu'elles soient interrompues. Pour plus d'informations, consultez Demander des instances Spot pour des applications flexibles et tolérantes aux pannes.

   b. (Facultatif) Pour IAM instance profile (Profil d'instance IAM), sélectionnez un rôle à associer aux instances. Pour plus d'informations, consultez Rôle IAM pour les applications qui s'exécutent sur des instances Amazon EC2.

c. (Facultatif) Pour la surveillance, choisissez si vous souhaitez autoriser les instances à publier des données métriques à intervalles d'une minute sur Amazon CloudWatch en activant une surveillance détaillée. Des frais supplémentaires seront facturés. Pour plus d'informations, consultez Configurer la surveillance pour les instances à scalabilité automatique.

d. (Facultatif) Pour Advanced details (Détails avancés), User Data (Données utilisateur), vous pouvez spécifier les données utilisateur pour configurer une instance lors du lancement ou pour exécuter un script de configuration après le démarrage de l'instance.

e. (Facultatif) Pour Advanced details (Détails avancés), IP address type (Type d'adresse IP), choisissez si vous souhaitez affecter une adresse IP publique aux instances du groupe. Si vous ne définissez pas de valeur, la valeur par défaut consiste à utiliser les paramètres IP publics d'attribution automatique des sous-réseaux dans lesquels vos instances sont lancées.

9. (Facultatif) Pour Stockage (volumes), si vous n'avez pas besoin de stockage supplémentaire, vous pouvez ignorer cette section. Sinon, pour spécifier les volumes à attacher aux instances en plus des volumes spécifiés par l'AMI, choisissez Add new volume (Ajouter un nouveau volume). Choisissez ensuite les options souhaitées et les valeurs associées pour Devices (Appareils), Snapshot (Instantané), Size (Taille),Volume type (Type de volume), IOPS, Throughput (Débit), Delete on termination (Supprimer à la résiliation) et Encrypted (Chiffré).

10. Pour Security groups (Groupes de sécurité), créez ou sélectionnez le groupe de sécurité à associer aux instances du groupe. Si vous ne désélectionnez pas Create a new security group (Créer un groupe de sécurité), une règle SSH par défaut est configurée pour les instances Amazon EC2 s'exécutant sur les systèmes d'exploitation Linux. Une règle RDP par défaut est configurée pour les instances Amazon EC2 s'exécutant sous Windows.

11. Pour Key pair (login) (Paire de clés [connexion]), choisissez une option sous Key pair options (Options de la paire de clés).

    Si vous avez déjà configuré une paire de clés d'instance Amazon EC2, vous pouvez la choisir ici.

    Si vous ne disposez pas déjà d'une paire de clés d'instance Amazon EC2, choisissez Create a new key pair (Créer une nouvelle paire de clés) et attribuez-lui un nom facilement identifiable. Choisissez Download key pair (Télécharger une paire de clés) pour télécharger la paire de clés sur votre ordinateur.

> **Important**
>
> Ne choisissez pas Proceed without a key pair (Continuer sans paire de clés) si vous avez besoin de vous connecter aux instances.

12. Sélectionnez la case à cocher de confirmation, puis choisissez Create launch configuration (Créer une configuration de lancement).

Pour créer une configuration de lancement à partir d'une configuration de lancement existante (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Dans la barre de navigation supérieure, sélectionnez votre AWS région.
3. Dans le volet de navigation à gauche, sous Auto Scaling, choisissez Groupes Auto Scaling.
4. Sélectionnez Configurations de lancement en haut de la page. Lorsque vous êtes invité à confirmer, choisissez Afficher les configurations de lancement pour confirmer que vous souhaitez consulter la page Configurations de lancement.
5. Sélectionnez la configuration du lancement et cliquez sur Actions, Copy launch configuration (Copier une configuration du lancement). Ainsi, vous créez une nouvelle configuration du lancement avec les mêmes options que l'originale, mais avec la mention « Copie » ajoutée à son nom.
6. Sur la page Copy Launch Configuration (Copier une configuration du lancement), modifiez les options selon vos besoins, puis cliquez sur Create launch configuration (Créer la configuration du lancement).

Pour créer une configuration du lancement avec la ligne de commande

Vous pouvez utiliser l'une des commandes suivantes :

- create-launch-configuration (AWS CLI)
- Nouveautés LaunchConfiguration ()AWS Tools for Windows PowerShell

Configurer les options de métadonnées d’instance

Amazon EC2 Auto Scaling prend en charge la configuration du service de métadonnées d'instance (IMDS) dans les configurations de lancement. Cela vous donne la possibilité d'utiliser des configurations de lancement pour configurer les instances Amazon EC2 dans vos groupes Auto Scaling pour qu'elles nécessitent Instance Metadata Service Version 2 (IMDSv2), qui est une méthode orientée session pour demander des métadonnées d'instance. Pour plus d'informations sur les avantages d'IMDsv2, consultez l'article du blog AWS intitulé Améliorations apportées pour ajouter une défense en profondeur au service de métadonnées d'instance EC2.

Vous pouvez configurer IMDS pour prendre en charge IMDSv2 et IMDSv1 (valeur par défaut) ou pour exiger l'utilisation d'IMDSv2. Si vous utilisez le AWS CLI ou l'un des SDK pour configurer l'IMDS, vous devez utiliser la dernière version du AWS CLI ou du SDK pour exiger l'utilisation d'IMDSv2.

Vous pouvez configurer votre configuration du lancement pour les éléments suivants :

- Imposer l’utilisation d’IMDSv2 lorsqu’il s’agit de demander des métadonnées d’instance
- Spécifier la durée de vie (hop limit) de la réponse PUT
- Désactiver l'accès aux métadonnées d'instance

Vous trouverez plus de détails sur la configuration du service de métadonnées d'instance dans la rubrique suivante : Configuration du service de métadonnées d'instance dans le guide de l'utilisateur Amazon EC2.

Utilisez la procédure suivante pour configurer les options IMDS dans une configuration du lancement. Une fois la configuration du lancement créée, vous pouvez associer celle-ci à votre groupe Auto Scaling. Si vous associez la configuration du lancement à un groupe Auto Scaling existant, la configuration du lancement existante est dissociée du groupe Auto Scaling et les instances existantes devront être remplacées pour utiliser les options IMDS que vous avez spécifiées dans la nouvelle configuration du lancement. Pour plus d’informations, consultez Modifier la configuration du lancement pour un groupe Auto Scaling.

Pour configurer IMDS dans une configuration du lancement (console)

1. Ouvrez la console Amazon EC2 à l’adresse https://console.aws.amazon.com/ec2/.
2. Dans la barre de navigation supérieure, sélectionnez votre AWS région.
3. Dans le volet de navigation à gauche, sous Auto Scaling, choisissez Groupes Auto Scaling.

4. Sélectionnez Configurations de lancement en haut de la page. Lorsque vous êtes invité à confirmer, choisissez Afficher les configurations de lancement pour confirmer que vous souhaitez consulter la page Configurations de lancement.
5. Choisissez Create launch configuration (Créer la configuration du lancement) et créez la configuration du lancement de la manière habituelle. Indiquez l'ID d'Amazon Machine Image (AMI), le type d'instance et éventuellement une paire de clés, un ou plusieurs groupes de sécurité et les volumes EBS ou les volumes de stockage d'instance supplémentaires pour vos instances.
6. Pour configurer les options de métadonnées d'instance pour toutes les instances associées à cette configuration du lancement, dans Additional configuration (Configuration supplémentaire), sous Advanced details (Détails avancés), procédez comme suit :

    a. Pour Metadata accessible (Métadonnées accessibles), choisissez si vous activez ou désactivez l'accès aux métadonnées de l'instance. Par défaut, le point de terminaison HTTP est activé. Si vous choisissez de désactiver le point de terminaison, l'accès aux métadonnées de votre instance est désactivé. Vous pouvez spécifier la condition pour exiger IMDSv2 uniquement lorsque le point de terminaison HTTP est activé.

    b. Pour Metadata version (Version des métadonnées), vous pouvez choisir d'exiger l'utilisation de Service des métadonnées d'instance Version 2 (IMDSv2) lors de la demande de métadonnées d'instance. Si vous ne spécifiez pas de valeur, la valeur par défaut est de prendre en charge IMDSv1 et IMDSv2.

    c. Pour Metadata token response hop limit (Durée de vie de réponse du jeton de métadonnées), vous pouvez définir le nombre autorisé de sauts réseau pour le jeton de métadonnées. Si vous ne spécifiez pas de valeur, la valeur par défaut est 1.
7. Lorsque vous avez terminé, choisissez Create launch configuration (Créer la configuration du lancement).

Pour imposer l'utilisation d'IMDSv2 dans une configuration du lancement à l'aide de la commande AWS CLI

Utilisez la commande create-launch-configuration suivante avec l'option `--metadata-options` définie sur `HttpTokens=required`. Lorsque vous spécifiez une valeur pour `HttpTokens`, vous devez également définir `HttpEndpoint` sur enabled (activé). Comme l'en-tête de jeton sécurisé est défini sur required (obligatoire) pour les demandes de récupération de métadonnées, cette option permet à l'instance d'imposer l'utilisation d'IMDSv2 lors de la demande de métadonnées d'instance.

```
aws autoscaling create-launch-configuration \
  --launch-configuration-name my-lc-with-imdsv2 \
  --image-id ami-01e24be29428c15b2 \
  --instance-type t2.micro \
      ...
  --metadata-options "HttpEndpoint=enabled,HttpTokens=required"
```

Désactivation de l'accès aux métadonnées d'instance

Utilisez la commande create-launch-configuration suivante pour désactiver l'accès aux métadonnées d'instance. Vous pouvez réactiver l'accès ultérieurement à l'aide de la commande modify-instance-metadata-options.

```
aws autoscaling create-launch-configuration \
  --launch-configuration-name my-lc-with-imds-disabled \
  --image-id ami-01e24be29428c15b2 \
  --instance-type t2.micro \
      ...
  --metadata-options "HttpEndpoint=disabled"
```

Créer une configuration du lancement avec une instance EC2

Vous avez également la possibilité de créer une configuration de lancement à l'aide des attributs d'une instance EC2 en cours d'exécution.

Il existe des différences entre la création d'une configuration du lancement à partir de zéro et à partir d'une instance EC2 existante. Lorsque vous créez une configuration du lancement à partir de zéro, vous spécifiez l'ID d'image, le type d'instance, les ressources facultatives (comme les périphériques de stockage), et les paramètres facultatifs (comme la surveillance). Lorsque vous créez une configuration du lancement à partir d'une instance en cours d'exécution, Amazon EC2 Auto Scaling tire ses attributs pour la configuration du lancement de l'instance désignée. Les attributs sont également issus du mappage de périphérique de stockage en mode bloc pour l'AMI à partir de laquelle l'instance a été lancée, en ignorant tout périphérique de stockage en mode bloc supplémentaire ajouté après le lancement.

Lorsque vous créez une configuration du lancement avec une instance en cours d'exécution, vous pouvez remplacer les attributs suivants en les spécifiant ensuite dans le cadre de la même demande : l'AMI, les périphériques de stockage en mode bloc, la paire de clés, le profil d'instance, le type d'instance, le noyau, la surveillance de l'instance, la location de placement, le ramdisk, les groupes

de sécurité, le prix Spot (max), les données utilisateur, si l'instance possède une adresse IP publique, et si l'instance est optimisée pour EBS.

> **Note**
>
> Si l'instance spécifiée possède des propriétés qui ne sont actuellement pas prises en charge par les configurations de lancement, les instances lancées par le groupe Auto Scaling peuvent être différentes de celles de l'instance EC2 d'origine.

> **Important**
>
> L'AMI utilisée pour lancer l'instance spécifiée doit toujours exister.

Rubriques

- Créer une configuration du lancement à partir d'une instance EC2 (AWS CLI)
- Créer une configuration du lancement à partir d'une instance et remplacer les périphériques de stockage en mode bloc (AWS CLI)
- Créer une configuration du lancement et remplacer le type d'instance (AWS CLI)

Créer une configuration du lancement à partir d'une instance EC2 (AWS CLI)

Utilisez la commande create-launch-configuration suivante pour créer une configuration du lancement à partir d'une instance utilisant les mêmes attributs que l'instance. Tous les périphériques de stockage en mode bloc ajoutés après le lancement sont ignorés.

```
aws autoscaling create-launch-configuration --launch-configuration-name my-lc-from-instance --instance-id i-a8e09d9c
```

Vous pouvez utiliser la commande describe-launch-configurations suivante pour décrire la configuration du lancement et vérifier que ses attributs correspondent à ceux de l'instance.

```
aws autoscaling describe-launch-configurations --launch-configuration-names my-lc-from-instance
```

Voici un exemple de réponse.

```
{
    "LaunchConfigurations": [
        {
            "UserData": null,
            "EbsOptimized": false,
            "LaunchConfigurationARN": "arn",
            "InstanceMonitoring": {
                "Enabled": false
            },
            "ImageId": "ami-05355a6c",
            "CreatedTime": "2014-12-29T16:14:50.382Z",
            "BlockDeviceMappings": [],
            "KeyName": "my-key-pair",
            "SecurityGroups": [
                "sg-8422d1eb"
            ],
            "LaunchConfigurationName": "my-lc-from-instance",
            "KernelId": "null",
            "RamdiskId": null,
            "InstanceType": "t1.micro",
            "AssociatePublicIpAddress": true
        }
    ]
}
```

Créer une configuration du lancement à partir d'une instance et remplacer les périphériques de stockage en mode bloc (AWS CLI)

Par défaut, Amazon EC2 Auto Scaling utilise les attributs de l'instance EC2 spécifiée pour créer la configuration du lancement. Toutefois, les périphériques de stockage en mode bloc proviennent de l'AMI utilisée pour lancer l'instance, pas de l'instance. Pour ajouter des périphériques de stockage en mode bloc à la configuration du lancement, remplacez le mappage de périphérique de stockage en mode bloc pour la configuration du lancement.

Utilisez la commande create-launch-configuration suivante pour créer une configuration du lancement avec une instance EC2 et un mappage de périphérique de stockage en mode bloc personnalisé.

```
aws autoscaling create-launch-configuration --launch-configuration-name my-lc-from-instance-bdm --instance-id i-a8e09d9c \
```

```
  --block-device-mappings "[{\"DeviceName\":\"/dev/sda1\",\"Ebs\":{\"SnapshotId\":
\"snap-3decf207\"}},{\"DeviceName\":\"/dev/sdf\",\"Ebs\":{\"SnapshotId\":\"snap-
eed6ac86\"}}]"
```

Utilisez la commande describe-launch-configurations suivante pour décrire la configuration du lancement et vérifier qu'elle utilise le mappage de périphérique de stockage en mode bloc.

```
aws autoscaling describe-launch-configurations --launch-configuration-names my-lc-from-
instance-bdm
```

L'exemple de réponse suivant décrit la configuration du lancement.

```
{
    "LaunchConfigurations": [
        {
            "UserData": null,
            "EbsOptimized": false,
            "LaunchConfigurationARN": "arn",
            "InstanceMonitoring": {
                "Enabled": false
            },
            "ImageId": "ami-c49c0dac",
            "CreatedTime": "2015-01-07T14:51:26.065Z",
            "BlockDeviceMappings": [
                {
                    "DeviceName": "/dev/sda1",
                    "Ebs": {
                        "SnapshotId": "snap-3decf207"
                    }
                },
                {
                    "DeviceName": "/dev/sdf",
                    "Ebs": {
                        "SnapshotId": "snap-eed6ac86"
                    }
                }
            ],
            "KeyName": "my-key-pair",
            "SecurityGroups": [
                "sg-8637d3e3"
            ],
            "LaunchConfigurationName": "my-lc-from-instance-bdm",
            "KernelId": null,
```

```
            "RamdiskId": null,
            "InstanceType": "t1.micro",
            "AssociatePublicIpAddress": true
        }
    ]
}
```

Créer une configuration du lancement et remplacer le type d'instance (AWS CLI)

Par défaut, Amazon EC2 Auto Scaling utilise les attributs de l'instance EC2 spécifiée pour créer la configuration du lancement. En fonction de vos besoins, vous pouvez souhaiter remplacer les attributs de l'instance et utiliser les valeurs dont vous avez besoin. Par exemple, vous pouvez remplacer le type d'instance.

Utilisez la commande create-launch-configuration suivante pour créer une configuration du lancement avec une instance EC2 et un type d'instance (par exemple `t2.medium`) différent de l'instance (par exemple `t2.micro`).

```
aws autoscaling create-launch-configuration --launch-configuration-name my-lc-from-instance-changetype \
  --instance-id i-a8e09d9c --instance-type t2.medium
```

Utilisez la commande describe-launch-configurations suivante pour décrire la configuration du lancement et vérifier que le type d'instance a été remplacé.

```
aws autoscaling describe-launch-configurations --launch-configuration-names my-lc-from-instance-changetype
```

L'exemple de réponse suivant décrit la configuration du lancement.

```
{
    "LaunchConfigurations": [
        {
            "UserData": null,
            "EbsOptimized": false,
            "LaunchConfigurationARN": "arn",
            "InstanceMonitoring": {
                "Enabled": false
            },
            "ImageId": "ami-05355a6c",
```

```
            "CreatedTime": "2014-12-29T16:14:50.382Z",
            "BlockDeviceMappings": [],
            "KeyName": "my-key-pair",
            "SecurityGroups": [
                "sg-8422d1eb"
            ],
            "LaunchConfigurationName": "my-lc-from-instance-changetype",
            "KernelId": "null",
            "RamdiskId": null,
            "InstanceType": "t2.medium",
            "AssociatePublicIpAddress": true
        }
    ]
}
```


Modifier la configuration du lancement pour un groupe Auto Scaling

> **Important**
>
> Nous fournissons des informations sur les configurations de lancement pour les clients qui n'ont pas encore migré des configurations de lancement vers les modèles de lancement. Pour plus d'informations sur la migration de vos groupes Auto Scaling vers les modèles de lancement, consultez la section Migrez vos groupes Auto Scaling pour lancer des modèles.

Cette rubrique décrit comment associer une configuration de lancement différente à votre groupe Auto Scaling.

Une fois que vous avez modifié la configuration de lancement, les nouvelles instances sont lancées à l'aide des nouvelles options de configuration, mais les instances existantes ne sont pas affectées. Pour plus d'informations, consultez Mise à jour des instances Auto Scaling.

Pour remplacer la configuration du lancement pour un groupe Auto Scaling (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Dans le volet de navigation à gauche, sous Auto Scaling, choisissez Groupes Auto Scaling.
3. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.

4. Sous l'onglet Details (Détails) choisissez Launch configuration (Configuration du lancement), Edit (Modifier).
5. Pour Configuration de lancement, choisissez la configuration de lancement.
6. Une fois que vous avez terminé, choisissez Update (Mettre à jour).

Pour modifier la configuration de lancement d'un groupe Auto Scaling à l'aide de la ligne de commande

Vous pouvez utiliser l'une des commandes suivantes :

- update-auto-scaling-group (AWS CLI)
- Mettre à jour en tant que AutoScaling groupe ()AWS Tools for Windows PowerShell

Groupes Auto Scaling

> **Note**
>
> Si vous découvrez les groupes Auto Scaling, suivez les étapes du didacticiel Create your first Auto Scaling group pour commencer et voir comment un groupe Auto Scaling réagit lorsqu'une instance du groupe se termine.

Un groupe Auto Scaling contient un ensemble d'instances EC2 traitées comme un regroupement logique, aux fins de mise à l'échelle et de gestion automatique. Ils vous permettent également d'utiliser des fonctionnalités Amazon EC2 Auto Scaling telles que les remplacements des surveillances de l'état et des politiques de mise à l'échelle. La mise à l'échelle et le maintien automatiques du nombre d'instances dans un groupe Auto-Scaling constitue la fonctionnalité de base du service Amazon EC2 Auto Scaling.

La taille d'un groupe Auto Scaling dépend du nombre d'instances que vous définissez en tant que capacité souhaitée. Vous pouvez ajuster sa taille afin de répondre à la demande, manuellement ou à l'aide de la scalabilité automatique.

Un groupe Auto Scaling démarre en lançant suffisamment d'instances pour atteindre la capacité souhaitée. Le groupe maintient le nombre d'instances en réalisant des surveillances périodiques de l'état sur les instances du groupe. Le groupe Auto Scaling continue à conserver un nombre fixe d'instances, même si une instance devient défectueuse. Si une instance devient défectueuse, le groupe la résilie et en lance une autre pour la remplacer. Pour plus d'informations, consultez Surveillance de l'état des instances dans un groupe Auto Scaling.

Vous pouvez utiliser des politiques de mise à l'échelle pour augmenter ou réduire dynamiquement le nombre d'instances du groupe et répondre ainsi aux changements de conditions. Lorsque la politique de mise à l'échelle est appliquée, le groupe Auto Scaling ajuste la capacité souhaitée du groupe, entre les valeurs de capacité minimum et maximum que vous spécifiez, et lance les instances ou la résilie le cas échéant. Vous pouvez également mettre à niveau selon un calendrier. Pour plus d'informations, consultez Choisissez votre méthode de mise à l'échelle.

Lors de la création d'un groupe Auto Scaling, vous pouvez choisir de lancer les instances à la demande et/ou les instances Spot. Vous pouvez spécifier plusieurs options d'achat pour votre groupe Auto Scaling uniquement lorsque vous utilisez un modèle de lancement. Pour plus d'informations, consultez Groupes Auto Scaling combinant plusieurs types d'instances et options d'achat.

Les instances Spot vous permettent d'accéder à des fonctionnalités non-utilisées de EC2 tout en bénéficiant de remises conséquentes par rapport aux tarifs à la demande. Pour plus d'informations, consultez Amazon EC2 instances Spot. Il existe des différences importantes entre les instances ponctuelles et les instances à la demande :

- Le prix des instances ponctuelles varie en fonction de la demande
- Amazon EC2 peut résilier une instance Spot individuelle au fur et à mesure que la disponibilité ou le prix des instances Spot change

Lorsque votre instance Spot est résiliée, le groupe Auto Scaling tente de lancer une instance de remplacement pour maintenir la capacité souhaitée pour le groupe.

Lorsque les instances sont lancées, si vous avez spécifié plusieurs zones de disponibilité, la capacité souhaitée est distribuée entre ces zones de disponibilité. Si une action de mise à l'échelle se produit, Amazon EC2 Auto Scaling gère automatiquement l'équilibre entre l'ensemble des zones de disponibilité que vous spécifiez.

Table des matières

Créer des groupes Auto Scaling à l’aide de modèles de lancement

Si vous avez créé un modèle de lancement, vous pouvez créer un groupe Auto Scaling qui utilise le modèle de lancement comme modèle de configuration pour ses instances EC2. Le modèle de

lancement spécifie plusieurs informations, notamment l'identifiant d'AMI, le type d'instance, la paire de clés, les groupes de sécurité et le mappage de périphérique de stockage en mode bloc pour les instances. Pour de plus amples informations sur la création de modèles de lancement, veuillez consulter Créer un modèle de lancement pour un groupe Auto Scaling.

Vous devez disposer des autorisations nécessaires pour créer un groupe Auto Scaling. Vous devez également disposer des autorisations nécessaires pour créer un rôle lié à un service qu'Amazon EC2 Auto Scaling utilise pour effectuer les actions en votre nom, s'il n'existe pas encore. Pour obtenir des exemples de politiques IAM qu'un administrateur peut utiliser comme référence pour vous accorder des autorisations, consultez Exemples de politiques basées sur l'identité et Support de modèle de lancement.

Table des matières

Créer un groupe Auto Scaling avec un modèle de lancement

Lorsque vous créez un groupe Auto Scaling, vous devez indiquer les informations nécessaires pour configurer les instances Amazon EC2, les zones de disponibilité et les sous-réseaux VPC pour les instances, la capacité souhaitée et les limites de capacité minimale et maximale.

Pour configurer des instances Amazon EC2 lancées par votre groupe Auto Scaling, vous pouvez spécifier un modèle de lancement ou une configuration du lancement. La procédure suivante montre comment créer un groupe Auto Scaling avec un modèle de lancement.

Prérequis

- Vous devez avoir créé un modèle de lancement. Pour plus d'informations, consultez Créer un modèle de lancement pour un groupe Auto Scaling.

Pour créer un groupe Auto Scaling avec un modèle de lancement (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Dans la barre de navigation en haut de l'écran, choisissez le même Région AWS que celui que vous avez utilisé lors de la création du modèle de lancement.

3. Choisissez Créer un groupe Auto Scaling.
4. Dans la page Choose launch template or configuration (Choisir un modèle de lancement ou une configuration), procédez comme suit :

   a. Pour Auto Scaling group name (Nom du groupe Auto Scaling), saisissez un nom pour votre groupe Auto Scaling.

   b. Dans Launch template (Modèle de lancement), choisissez un modèle de lancement existant.

   c. Pour Version du modèle de lancement, indiquez si le groupe Auto Scaling utilise la version par défaut, la version la plus récente ou une version spécifique du modèle de lancement lors de l'évolutivité horizontale.

   d. Vérifiez que votre modèle de lancement prend en charge toutes les options que vous envisagez d'utiliser, puis choisissez Next (Suivant).
5. Sur la page Choisir les options de lancement d’instance, si vous n’utilisez pas plusieurs types d’instance, vous pouvez ignorer la section Exigences relatives au type d’instance pour utiliser le type d’instance EC2 indiqué dans le modèle de lancement.

   Pour utiliser plusieurs types d’instances, consultez Groupes Auto Scaling combinant plusieurs types d'instances et options d'achat.
6. Sous Network (Réseau), pour VPC, choisissez un VPC. Le groupe Auto Scaling doit être créé dans le même VPC que le groupe de sécurité que vous avez spécifié dans votre modèle de lancement.
7. Pour Availability Zones and subnets (Zones de disponibilité et sous-réseaux), choisissez un ou plusieurs sous-réseaux dans le VPC spécifié. Utilisez les sous-réseaux dans plusieurs zones de disponibilité pour une haute disponibilité. Pour plus d’informations, consultez Considérations à prendre en compte lors du choix des sous-réseaux VPC.
8. Si vous avez créé un modèle de lancement avec un type d'instance spécifié, vous pouvez passer à l'étape suivante pour créer un groupe Auto Scaling qui utilise le type d'instance dans le modèle de lancement.

   Vous pouvez également choisir l'option Remplacer le modèle de lancement si aucun type d'instance n'est spécifié dans votre modèle de lancement ou si vous souhaitez utiliser plusieurs types d'instance pour la scalabilité automatique. Pour plus d’informations, consultez Groupes Auto Scaling combinant plusieurs types d'instances et options d'achat.
9. Choisissez Next (Suivant) pour passer à l'étape suivante.

Vous pouvez également accepter le reste des valeurs par défaut, puis choisir Skip to review (Passer à la révision).

10. (Facultatif) Sur la page Configure advanced options (Configurer les option avancées), configurez les options suivantes, puis choisissez Next (Suivant) :

    a. Sous Paramètres supplémentaires, Surveillance, indiquez si vous souhaitez activer la collecte des métriques de CloudWatch groupe. Ces métriques fournissent des mesures qui peuvent être des indicateurs d'un problème potentiel, comme le nombre d'instances en cours de résiliation ou le nombre d'instances en attente. Pour plus d'informations, consultez Surveillez CloudWatch les métriques de vos groupes et instances Auto Scaling.

    b. Pour Activer le préchauffage de l'instance par défaut, sélectionnez cette option et choisissez le temps de préchauffage de votre application. Si vous créez un groupe Auto Scaling doté d'une politique de dimensionnement, la fonctionnalité de préchauffage de l'instance par défaut améliore les CloudWatch métriques Amazon utilisées pour le dimensionnement dynamique. Pour plus d'informations, consultez Définir la préparation par défaut d'instance d'un groupe Auto Scaling.

11. (Facultatif) Sur la page Configure group size and scaling policies (Configurer les politiques de taille de groupe et de mise à l'échelle), configurez les options suivantes, puis choisissez Next (Suivant) :

    a. Dans Taille du groupe, pour la Capacité souhaitée, entrez le nombre initial d'instances à lancer.

    b. Dans la section Mise à l'échelle, sous Limites de mise à l'échelle, si votre nouvelle valeur pour la capacité souhaitée est supérieure à la capacité minimale souhaitée et à la capacité maximale souhaitée, la capacité maximale souhaitée est automatiquement augmentée à la nouvelle valeur de capacité souhaitée. Vous pouvez modifier ces limites si nécessaire. Pour plus d'informations, consultez Définissez des limites de mise à l'échelle pour votre groupe Auto Scaling.

    c. Pour le dimensionnement automatique, indiquez si vous souhaitez créer une politique de dimensionnement de suivi des cibles. Vous pouvez également élaborer cette politique après avoir créé votre groupe Auto Scaling.

       Si vous choisissez la politique de dimensionnement de suivi des cibles, suivez les instructions dans Création d'une politique de suivi des cibles et d'échelonnement pour créer la politique.

d. Pour la politique de maintenance des instances, indiquez si vous souhaitez créer une politique de maintenance des instances. Vous pouvez également élaborer cette politique après avoir créé votre groupe Auto Scaling. Pour créer une politique, suivez les instructions fournies dans Définir une politique de maintenance des instances.

e. Sous Instance scale-in protection (Protection contre la diminution en charge des instances), choisissez si vous souhaitez activer la protection contre la diminution de la taille d'instance. Pour plus d'informations, consultez Utiliser la protection de la taille d'instance.

12. (Facultatif) Pour recevoir des notifications, dans Add notification (Ajouter une notification), configurez la notification, puis choisissez Next (Suivant). Pour plus d'informations, consultez Options de notification Amazon SNS pour Amazon EC2 Auto Scaling.
13. (Facultatif) Pour ajouter des balises, choisissez Add tag (Ajouter une balise), fournissez une clé de balise et une valeur pour chaque balise, puis choisissez Next (Suivant). Pour plus d'informations, consultez Baliser des groupes et des instances Auto Scaling.
14. Sur la page Review, sélectionnez Create Auto Scaling group (Créer un groupe Auto Scaling).

Pour créer un groupe Auto Scaling avec la ligne de commande

Vous pouvez utiliser l'une des commandes suivantes :

- create-auto-scaling-group (AWS CLI)
- Nouveautés-AS AutoScalingGroup ()AWS Tools for Windows PowerShell

Créer un groupe Auto Scaling avec l'Amazon EC2 Launch Wizard

La procédure suivante montre comment créer un groupe Auto Scaling en utilisant l'assistant de Launch instance (lancement d'instance) dans la console Amazon EC2. Cette option remplit automatiquement un modèle de lancement avec certains détails de configuration de l'assistant de lancement d'instance.

> **Note**
>
> L'assistant ne remplit pas le groupe Auto Scaling avec le nombre d'instances que vous spécifiez ; il remplit uniquement le modèle de lancement avec l'ID Amazon Machine Image (AMI) et le type d'instance. Utilisez l'assistant de Create Auto Scaling group (création de groupe Auto Scaling) pour spécifier le nombre d'instances à lancer.

Une AMI fournit les informations nécessaires à la configuration d'une instance. Lorsque vous avez besoin de plusieurs instances configurées de manière identique, il est possible de lancer plusieurs instances à partir d'une même AMI. Nous vous recommandons d'utiliser une AMI personnalisée sur laquelle votre application est déjà installée pour éviter que vos instances ne soient résiliées si vous redémarrez une instance appartenant à un groupe Auto Scaling. Pour utiliser une AMI personnalisée avec Amazon EC2 Auto Scaling, vous devez d'abord créer votre AMI à partir d'une instance personnalisée, puis utiliser l'AMI pour créer un modèle de lancement pour votre groupe Auto Scaling.

Prérequis

- Vous devez avoir créé une AMI personnalisée Région AWS là où vous prévoyez de créer le groupe Auto Scaling. Pour plus d'informations, consultez la section Création d'une AMI dans le guide de l'utilisateur Amazon EC2.

Utiliser une AMI personnalisée comme modèle

Dans cette section, vous utilisez l'assistant de lancement Amazon EC2 pour remplir automatiquement un modèle de lancement avec votre AMI personnalisée. Vous pouvez également configurer le modèle de lancement à partir de zéro ou pour une description plus détaillée des paramètres que vous pouvez configurer dans votre modèle de lancement, consultez Créer votre modèle de lancement (console).

Pour utiliser une AMI personnalisée comme modèle

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Dans la barre de navigation en haut de l'écran, le courant Région AWS est affiché. Sélectionnez une région dans laquelle vous souhaitez lancer votre groupe Auto Scaling.
3. Dans le panneau de navigation, sélectionnez Instances.
4. Choisissez Launch instance (Lancer une instance), puis effectuez les opérations suivantes :
    a. Sous Name and tags (Nom et balises), laissez le champ Name (Nom) vide. Le nom ne fait pas partie des données utilisées pour créer un modèle de lancement.
    b. Sous Application and OS Images (Amazon Machine Image) (Images d'applications et de systèmes d'exploitation [Amazon Machine Image]), choisissez Browse more AMIs (Parcourir plus d'AMI) pour parcourir le catalogue complet des AMI.

c. Choisissez My AMIs (Mes AMI), recherchez l'AMI que vous avez créée, puis choisissez Select (Sélectionner).

d. Pour Instance type (Type d'Instance), choisissez un type d'instance.

> **Note**
>
> Choisissez le même type d'instance que celui que vous avez utilisé lorsque vous avez créé l'AMI ou un type plus puissant.

e. Sur le côté droit de l'écran, sous Summary (Récapitulatif), pour Number of instances (Nombre d'instances), saisissez le nombre de votre choix. Le numéro que vous saisissez ici n'est pas important. Vous indiquez le nombre d'instances que vous souhaitez lancer lorsque vous créez le groupe Auto Scaling.

   Sous le champ Number of instances (Nombre d'instances), un message s'affiche indiquant When launching more than 1 instance, consider EC2 Auto Scaling (Lorsque vous lancez plus d'une instance, envisagez EC2 Auto Scaling).

f. Cliquez sur le texte du lien hypertexte consider EC2 Auto Scaling (envisagez EC2 Auto Scaling).

g. Dans la boîte de dialogue de confirmation Launch into Auto Scaling Group (Lancer dans le groupe Auto Scaling), choisissez Continue (Continuer) pour accéder à la page Create launch template (Créer un modèle de lancement) avec l'AMI et le type d'instance que vous avez sélectionnés dans l'assistant de lancement d'instance déjà remplis.

Après avoir choisi Continue (Continuer), la page Create launch template (Créer un modèle de lancement) s'ouvre. Procédez comme suit pour terminer la création d'un modèle de lancement.

Pour créer un modèle de lancement

1. Sous Launch template name and description (Nom et description du modèle de lancement), saisissez le nom et une description du nouveau modèle de lancement.
2. (Facultatif) Sous Key pair (login) (Paire de clés [connexion]), pour Key pair name (Nom de la paire de clés), choisissez le nom de la paire de clés précédemment créée à utiliser lors de la connexion aux instances, par exemple, en utilisant SSH.
3. (Facultatif) Sous Network settings (Paramètres réseau), pour Security groups (Groupes de sécurité), choisissez un ou plusieurs groupes de sécurité précédemment créés.

4. (Facultatif) Sous Configure storage (Configurer le stockage), mettez à jour la configuration du stockage. La configuration de stockage par défaut est déterminée par l'AMI et le type d'instance.
5. Lorsque vous avez terminé de configurer le modèle de lancement, sélectionnez Create launch template (Créer un modèle de lancement).
6. Sur la page de confirmation, choisissez Créer la configuration du lancement.

Créer un groupe Auto Scaling

> **Note**
>
> Le reste de cette rubrique décrit la procédure de base pour créer un groupe Auto Scaling. Pour plus de description des paramètres que vous pouvez configurer pour votre groupe Auto Scaling, consultez Créer un groupe Auto Scaling avec un modèle de lancement.

Après avoir choisi Create Auto Scaling group (Créer un groupe Auto Scaling), l'assistant Create Auto Scaling group (Créer un groupe Auto Scaling) s'ouvre. Suivez cette procédure pour créer un groupe Auto Scaling.

Pour créer un groupe Auto Scaling

1. Dans la page Choose launch template or configuration (Choisir un modèle de lancement ou une configuration), entrez un nom pour le groupe Auto Scaling.
2. Le modèle de lancement que vous avez créé est déjà sélectionné pour vous.

   Pour Version du modèle de lancement, indiquez si le groupe Auto Scaling utilise la version par défaut, la version la plus récente ou une version spécifique du modèle de lancement lors de l'évolutivité horizontale.
3. Choisissez Next (Suivant) pour passer à l'étape suivante.
4. Sur la page Choisir les options de lancement d’instance, si vous n’utilisez pas plusieurs types d’instance, vous pouvez ignorer la section Exigences relatives au type d’instance pour utiliser le type d’instance EC2 indiqué dans le modèle de lancement.

   Pour utiliser plusieurs types d’instances, consultez Groupes Auto Scaling combinant plusieurs types d'instances et options d'achat.

5. Sous Network (Réseau), pour VPC, choisissez un VPC. Le groupe Auto Scaling doit être créé dans le même VPC que le groupe de sécurité que vous avez spécifié dans votre modèle de lancement.

   > **Tip**
   >
   > Si vous n'avez pas spécifié de groupe de sécurité dans votre modèle de lancement, vos instances sont lancées avec un groupe de sécurité par défaut du VPC que vous spécifiez. Par défaut, ce groupe de sécurité n'autorise pas le trafic entrant provenant de réseaux externes.

6. Pour Availability Zones and subnets (Zones de disponibilité et sous-réseaux), choisissez un ou plusieurs sous-réseaux dans le VPC spécifié.
7. Choisissez deux fois Next (Suivant) pour aller à la page Configurer la taille du groupe et les politiques de mise à l'échelle.
8. Sous Taille du groupe, définissez la capacité souhaitée (nombre initial d'instances à lancer immédiatement après la création du groupe Auto Scaling).
9. Dans la section Mise à l'échelle, sous Limites de mise à l'échelle, si votre nouvelle valeur pour la capacité souhaitée est supérieure à la capacité minimale souhaitée et à la capacité maximale souhaitée, la capacité maximale souhaitée est automatiquement augmentée à la nouvelle valeur de capacité souhaitée. Vous pouvez modifier ces limites si nécessaire. Pour plus d'informations, consultez Définissez des limites de mise à l'échelle pour votre groupe Auto Scaling.
10. Choisissez Skip to review (Passer à la révision).
11. Sur la page Vérifier, sélectionnez Créer un groupe Auto Scaling.

Étapes suivantes

Vous pouvez vérifier que le groupe Auto Scaling a été créé correctement en consultant l'historique des activités. Dans l'onglet Activité, sous Historique de l'activité, la colonne Statut indique si votre groupe Auto Scaling a réussi à lancer des instances. Si les instances ne sont pas lancées ou si elles sont lancées, mais sont aussitôt résiliées, consultez les rubriques suivantes pour connaître les causes et les résolutions possibles :

- Dépanner Amazon EC2 Auto Scaling : échecs de lancement d'instance EC2
- Résoudre les problèmes d'Amazon EC2 Auto Scaling : AMI
- Résoudre les problèmes liés aux instances défectueuses dans Amazon EC2 Auto Scaling

Vous pouvez maintenant attacher un équilibreur de charge dans la même région que votre groupe Auto Scaling, si vous le souhaitez. Pour plus d'informations, voir Utiliser Elastic Load Balancing pour répartir le trafic sur les instances dans votre groupe Auto Scaling..

Groupes Auto Scaling combinant plusieurs types d'instances et options d'achat

Vous pouvez lancer et mettre automatiquement à l'échelle une flotte d'instances à la demande et d'instances Spot au sein d'un même groupe Auto Scaling. Outre les remises accordées sur l'utilisation des instances Spot, vous pouvez utiliser des instances réservées ou un Savings Plan afin de bénéficier de réductions sur les tarifs standard des instances à la demande. Ces facteurs vous permettent de réaliser des économies optimales sur les instances EC2 et vous font bénéficier de la mise à l'échelle et des performances souhaitées pour votre application.

Les instances Spot sont des capacités inutilisées disponibles à des prix très réduits par rapport au prix d'EC2 On-Demand. Les instances Spot constituent un choix économique si vous êtes flexible quant au moment où vos applications s'exécutent et à la possibilité de les interrompre. Ils peuvent être utilisés pour diverses applications flexibles et tolérantes aux pannes. Les exemples incluent les serveurs Web apatrides, les points de terminaison d'API, les applications de mégadonnées et d'analyse, les charges de travail conteneurisées, les pipelines CI/CD, le calcul haute performance et haut débit (HPC/HTC), les charges de travail de rendu et d'autres charges de travail flexibles.

Pour plus d'informations, consultez la section Options d'achat d'instances dans le guide de l'utilisateur Amazon EC2.

Rubriques

Présentation de la configuration

Cette rubrique fournit une vue d'ensemble et les meilleures pratiques pour créer un groupe d'instances mixtes.

Table des matières

Présentation

Pour créer un groupe d'instances mixtes, deux options s'offrent à vous :

- Sélection du type d'instance basée sur les attributs : définissez vos exigences de calcul pour choisir automatiquement vos types d'instances en fonction de leurs attributs d'instance spécifiques.
- Sélection manuelle du type d'instance : choisissez manuellement les types d'instance adaptés à votre charge de travail.

Manual selection

Les étapes suivantes expliquent comment créer un groupe d'instances mixtes en choisissant manuellement des types d'instance :

1. Sélectionnez un modèle de lancement contenant les paramètres pour lancer une instance EC2. Les paramètres des modèles de lancement sont facultatifs, mais Amazon EC2 Auto Scaling ne peut pas lancer une instance si l'identifiant Amazon Machine Image (AMI) est absent du modèle de lancement.
2. Choisissez l'option pour remplacer le modèle de lancement.
3. Choisissez manuellement les types d'instances adaptés à votre charge de travail.

4. Spécifiez les pourcentages des instances à la demande et des instances Spot à lancer.
5. Choisissez les stratégies d'allocation qui déterminent la façon dont Amazon EC2 Auto Scaling satisfait les capacités à la demande et Spot des types d'instances possibles.
6. Choisissez les zones de disponibilité et les sous-réseaux VPC dans lesquels vous souhaitez lancer vos instances.
7. Indiquez la taille initiale du groupe (la capacité souhaitée), ainsi que la taille minimale et maximale du groupe.

Des remplacements sont nécessaires pour remplacer le type d'instance déclaré dans le modèle de lancement et utiliser plusieurs types d'instances intégrés dans la définition de ressources du groupe Auto Scaling. Pour plus d'informations sur les types d'instances disponibles, consultez la section Types d'instances dans le guide de l'utilisateur Amazon EC2.

Vous pouvez également configurer les paramètres facultatifs suivants pour chaque type d'instance :

- `LaunchTemplateSpecification`— Vous pouvez attribuer un modèle de lancement différent à un type d'instance selon vos besoins. Cette option n'est actuellement pas disponible à partir de la console. Pour plus d'informations, consultez Utiliser un modèle de lancement différent pour un type d'instance.
- `WeightedCapacity`— Vous décidez dans quelle mesure l'instance compte pour la capacité souhaitée par rapport au reste des instances de votre groupe. Si vous spécifiez une valeur `WeightedCapacity` pour un type d'instance, vous devez spécifier une valeur `WeightedCapacity` pour tous les types d'instance. Par défaut, chaque instance compte pour un dans la capacité souhaitée. Pour plus d'informations, consultez Configurer un groupe Auto Scaling pour utiliser les poids d'instance.

Attribute-based selection

Pour permettre à Amazon EC2 Auto Scaling de choisir automatiquement vos types d'instances en fonction de leurs attributs d'instance spécifiques, suivez les étapes suivantes pour créer un groupe d'instances mixte en spécifiant vos besoins de calcul :

1. Sélectionnez un modèle de lancement contenant les paramètres pour lancer une instance EC2. Les paramètres des modèles de lancement sont facultatifs, mais Amazon EC2 Auto Scaling ne peut pas lancer une instance si l'identifiant Amazon Machine Image (AMI) est absent du modèle de lancement.

2. Choisissez l’option pour remplacer le modèle de lancement.
3. Spécifiez les attributs d’instance qui correspondent à vos exigences de calcul, telles que les vCPU et la mémoire.
4. Spécifiez les pourcentages des instances à la demande et des instances Spot à lancer.
5. Choisissez les stratégies d'allocation qui déterminent la façon dont Amazon EC2 Auto Scaling satisfait les capacités à la demande et Spot des types d'instances possibles.
6. Choisissez les zones de disponibilité et les sous-réseaux VPC dans lesquels vous souhaitez lancer vos instances.
7. Indiquez la taille initiale du groupe (la capacité souhaitée), ainsi que la taille minimale et maximale du groupe.

Des remplacements sont nécessaires pour remplacer le type d’instance déclaré dans le modèle de lancement et utiliser un ensemble d’attributs d’instance qui décrivent vos exigences de calcul. Pour les attributs pris en charge, consultez InstanceRequirementsle manuel Amazon EC2 Auto Scaling API Reference. Vous pouvez également utiliser un modèle de lancement qui contient déjà la définition des attributs d’instance.

Vous pouvez également configurer le paramètre `LaunchTemplateSpecification` dans la structure de remplacement pour attribuer un modèle de lancement différent à un ensemble d’exigences d’instance selon les besoins. Cette option n’est actuellement pas disponible à partir de la console. Pour plus d'informations, consultez la section LaunchTemplateOverrides dans le manuel Amazon EC2 Auto Scaling API Reference.

Par défaut, vous définissez le nombre d’instances pour qu’il corresponde à la capacité souhaitée de votre groupe Auto Scaling.

Vous pouvez également définir la valeur de la capacité souhaitée comme le nombre de vCPU ou la quantité de mémoire. Pour ce faire, utilisez la propriété `DesiredCapacityType` dans le fonctionnement de l’API `CreateAutoScalingGroup` ou le champ déroulant Type de capacité souhaitée dans la AWS Management Console. Il s’agit d’une alternative utile aux pondérations d’instance.

Flexibilité du type d’instance

Pour améliorer la disponibilité, déployez votre application sur plusieurs types d’instances. Il est recommandé d’utiliser plusieurs types d’instance pour satisfaire les exigences de capacité. Amazon

EC2 Auto Scaling peut ainsi lancer un autre type d'instance si la capacité d'instance est insuffisante dans les zones de disponibilité que vous avez choisies.

Si la capacité des instances Spot est insuffisante, Amazon EC2 Auto Scaling poursuivra ses tentatives de lancement à partir d'autres pools d'instances Spot. (Les pools qu'il utilise sont déterminés par votre choix de types d'instances et de stratégie d'allocation.) Amazon EC2 Auto Scaling vous aide à tirer parti des économies réalisées grâce aux instances Spot en les lançant à la place des instances à la demande.

Nous vous recommandons d'être flexible sur au moins 10 types d'instance pour chaque charge de travail. Lorsque vous choisissez des types d'instance, ne vous limitez pas aux nouveaux types d'instance les plus populaires. Choisir des types d'instance de génération plus ancienne a tendance à entraîner moins d'interruptions Spot, car ils sont moins demandés par les clients à la demande.

Flexibilité des zones de disponibilité

Nous vous recommandons fortement de répartir votre groupe Auto Scaling sur plusieurs zones de disponibilité. Avec plusieurs zones de disponibilité, vous pouvez concevoir des applications qui basculent automatiquement d'une zone à l'autre pour une plus grande résilience.

L'avantage supplémentaire est que vous pouvez accéder à un groupe de capacités Amazon EC2 plus important par rapport aux groupes d'une seule zone de disponibilité. Dans la mesure où la capacité fluctue en toute indépendance pour chaque type d'instance de chaque zone de disponibilité, il est souvent possible d'obtenir davantage de capacité de calcul lorsque l'on fait preuve de souplesse dans le choix à fois des types d'instances et de zones de disponibilité.

Pour plus d'informations sur l'utilisation des zones de disponibilité multiples, consultez Exemple : répartir les instances dans les zones de disponibilité.

prix Spot max

Lorsque vous créez votre groupe Auto Scaling à l'aide du AWS CLI ou d'un SDK, vous pouvez spécifier le `SpotMaxPrice` paramètre. Le paramètre `SpotMaxPrice` détermine le prix maximum que vous êtes prêt à payer pour une heure d'instance Spot.

Lorsque vous indiquez le paramètre `WeightedCapacity` dans vos remplacements (ou `"DesiredCapacityType": "vcpu"` ou `"DesiredCapacityType": "memory-mib"` au niveau du groupe), le prix maximum représente le prix unitaire maximum, et non le prix maximum pour une instance complète.

Nous vous recommandons fortement de ne pas indiquer de prix maximum. Votre application peut ne pas fonctionner si vous ne recevez pas d'Instances Spot, par exemple lorsque votre prix maximum est trop bas. Si vous ne spécifiez pas de prix maximum, la valeur par défaut est le prix à la demande. Vous payez uniquement le prix pour les instances Spot que vous lancez. Vous pouvez toujours bénéficier des remises importantes proposées par les instances Spot. Ces remises sont possibles en raison de la tarification stable des instances Spot qui est disponible grâce au modèle de tarification Spot. Pour plus d'informations, consultez la section Tarification et économies dans le guide de l'utilisateur Amazon EC2.

Rééquilibrage de capacité proactif

Si votre cas d'utilisation le permet, nous vous recommandons un rééquilibrage de la capacité. Le rééquilibrage de capacité vous permet de maintenir la disponibilité de la charge de travail en augmentant de manière proactive votre flotte avec une nouvelle instance Spot avant qu'une instance Spot en cours ne reçoive l'avis d'interruption d'instance Spot de deux minutes.

Lorsque le rééquilibrage de la capacité est activé, Amazon EC2 Auto Scaling tente de remplacer de manière proactive les instances Spot qui ont reçu une recommandation de rééquilibrage. Cela vous permet de rééquilibrer votre charge de travail en de nouvelles instances Spot qui ne présentent pas un risque élevé d'interruption.

Pour plus d'informations, consultez Utiliser le rééquilibrage de la capacité pour gérer les interruptions Spot Amazon EC2.

Comportement de mise à l'échelle.

Lorsque vous créez un groupe d'instances mixtes, il utilise des instances à la demande par défaut. Pour utiliser des instances Spot, vous devez modifier le pourcentage du groupe à lancer en tant qu'instances à la demande. Vous pouvez spécifier n'importe quel nombre compris entre 0 et 100 pour le pourcentage d'instances à la demande.

En option, vous pouvez également désigner un nombre de base d'instances à la demande pour commencer. Si vous procédez de la sorte, Amazon EC2 Auto Scaling attend le lancement des instances Spot jusqu'à ce que la capacité de base soit atteinte lorsque le groupe monte en puissance. Tout dépassement de la capacité de base utilise le pourcentage à la demande pour déterminer le nombre d'instances à la demande et le nombre d'instances ponctuelles à lancer.

Amazon EC2 Auto Scaling convertit le pourcentage en nombre équivalent d'instances. Si le résultat est un nombre fractionnaire, il est arrondi à l'entier supérieur en faveur des instances à la demande.

Le tableau suivant montre le comportement du groupe Auto Scaling à mesure que la taille du groupe augmente et diminue.

Exemple : comportement de mise à l'échelle

Options d’achat	La taille de groupe et le nombre total d’instances en cours d’exécution, toutes options d’achat confondues			
	10	20	30	40
Exemple 1 : base de 10, 50/50 % à la demande/Spot				
Instances à la demande (montant de base)	10	10	10	10
On-Demand instances	0	5	10	15
Spot instances	0	5	10	15
Exemple 2 : base de 0, 0/100 % à la demande/Spot				
Instances à la demande (montant de base)	0	0	0	0
On-Demand instances	0	0	0	0
Spot instances	10	20	30	40

Options d’achat	La taille de groupe et le nombre total d’instances en cours d’exécution, toutes options d’achat confondues			
Exemple 3 : base de 0, 60/40 % à la demande/Spot				
Instances à la demande (montant de base)	0	0	0	0
On-Demand instances	6	12	18	24
Spot instances	4	8	12	16
Exemple 4 : base de 0, 100/0 % à la demande/Spot				
Instances à la demande (montant de base)	0	0	0	0
On-Demand instances	10	20	30	40
Spot instances	0	0	0	0
Exemple 5 : base de 12, 0/100 % à la demande/Spot				
Instances à la demande (montant de base)	10	12	12	12

Options d’achat	La taille de groupe et le nombre total d’instances en cours d’exécution, toutes options d’achat confondues		
On-Demand instances	0	0	0	0
Spot instances	0	8	18	28

Lorsque la taille du groupe augmente, Amazon EC2 Auto Scaling tente d’équilibrer votre capacité uniformément entre les zones de disponibilité que vous avez indiquées. Ensuite, il lance des types d'instance en fonction de la stratégie d'allocation qui est spécifiée.

Lorsque la taille du groupe diminue, Amazon EC2 Auto Scaling identifie d’abord lequel des deux types (Spot ou à la demande) doit être résilié. Il essaye ensuite de résilier les instances de manière équilibrée dans les zones de disponibilité que vous avez indiquées. Cela favorise également la résiliation des instances d’une manière qui correspond le mieux à vos stratégies d’allocation. Pour plus d’informations sur les politiques de mise hors service, consultez la section Configurer les politiques de résiliation pour Amazon EC2 Auto Scaling.

Disponibilité régionale des types d’instances

La disponibilité des types d'instances EC2 varie en fonction de vos Région AWS besoins. Par exemple, les types d’instance de la nouvelle génération peuvent ne pas encore être disponibles dans une région donnée. En raison des variations de disponibilité des instances d’une région à l’autre, vous pouvez rencontrer des problèmes lorsque vous effectuez des demandes par programmation si plusieurs types d’instances dans vos remplacements ne sont pas disponibles dans votre région. L’utilisation de plusieurs types d’instances qui ne sont pas disponibles dans votre région peut entraîner l’échec total de la demande. Pour résoudre le problème, effectuez de nouveau la demande avec différents types d’instance, en vous assurant que chaque type d’instance est disponible dans la région. Pour rechercher les types d’instance offerts par emplacement, utilisez la commande describe-instance-type-offerings. Pour plus d'informations, consultez Trouver un type d'instance Amazon EC2 dans le guide de l'utilisateur Amazon EC2.

Ressources connexes

Pour en savoir plus sur les meilleures pratiques relatives aux instances Spot, consultez la section Meilleures pratiques pour EC2 Spot dans le guide de l'utilisateur Amazon EC2.

Limites

Après avoir ajouté des remplacements à un groupe Auto Scaling à l'aide d'une politique d'instances mixtes, vous pouvez mettre à jour les remplacements avec l'appel d'`UpdateAutoScalingGroup`API, mais pas les supprimer. Pour supprimer complètement les dérogations, vous devez d'abord changer de groupe Auto Scaling afin d'utiliser un modèle de lancement ou une configuration de lancement au lieu d'une politique d'instances mixtes. Vous pouvez ensuite ajouter à nouveau une politique d'instances mixtes sans aucune dérogation.

Stratégies d’allocation

Lorsque vous utilisez plusieurs types d’instance, vous gérez la façon dont Amazon EC2 Auto Scaling satisfait vos capacités à la demande et Spot des types d’instance possibles. Pour ce faire, vous devez définir des stratégies d’allocation.

Pour consulter les meilleures pratiques relatives à un groupe d'instances mixtes, consultezPrésentation de la configuration.

Table des matières

Spot instances

Amazon EC2 Auto Scaling fournit les stratégies d’allocation suivantes pour les instances Spot :

`price-capacity-optimized` (recommandé)

La stratégie d'allocation optimisée en termes de prix et de capacité prend en compte à la fois le prix et la capacité afin de sélectionner les groupes d'instances Spot les moins susceptibles d'être interrompus et dont le prix est le plus bas possible.

Nous vous recommandons cette stratégie lorsque vous débutez. Pour plus d'informations, consultez la section Présentation de la stratégie price-capacity-optimized d'allocation pour les instances Spot EC2 AWS sur le blog.

`capacity-optimized`

Amazon EC2 Auto Scaling sollicite vos instances Spot du pool avec une capacité optimale pour le nombre d'instances qui sont lancées.

Avec les instances Spot, la tarification change lentement au fil du temps en fonction des tendances à long terme en matière d'offre et de demande. Cependant, la capacité fluctue en temps réel. La stratégie `capacity-optimized` lance automatiquement des Instances Spot dans les pools les plus disponibles en examinant les données de capacité en temps réel et en prédisant les instances les plus disponibles. Cela permet de minimiser les interruptions pour les charges de travail de nature à entraîner des coûts plus élevés associés au redémarrage du travail et aux points de contrôle. Pour donner à certains types d'instance une plus grande chance de démarrer en premier, utilisez `capacity-optimized-prioritized`.

`capacity-optimized-prioritized`

Vous définissez l'ordre des types d'instance dans la liste des remplacements de modèle de lancement de la priorité la plus élevée à la plus basse (du premier au dernier de la liste). Amazon EC2 Auto Scaling implémente les priorités de type d'instance sur la base du meilleur effort, mais optimise d'abord la capacité. C'est une bonne option pour les charges de travail pour lesquelles la possibilité de perturbation doit être minimisée, mais la priorité de certains types d'instances est également importante. Notez que si la stratégie d'allocation à la demande est définie sur `prioritized`, la même priorité est appliquée lors de l'exécution de la capacité à la demande.

`lowest-price`

Amazon EC2 Auto Scaling sollicite vos instances Spot en utilisant les groupes de prix les plus bas au sein d'une zone de disponibilité, à travers le nombre N de groupes d'instances Spot que vous spécifiez pour le paramètre Groupes de prix les plus bas. Par exemple, si vous spécifiez quatre types d'instances et quatre zones de disponibilité, votre groupe Auto Scaling a accès à un maximum de 16 pools d'instances Spot. (Quatre dans chaque zone de disponibilité.) Si vous spécifiez deux pools d'instances Spot (N=2) pour la stratégie d'allocation, votre groupe Auto Scaling peut puiser dans les deux pools les moins chers de chaque zone de disponibilité afin de répondre à votre capacité Spot.

Cette stratégie prenant uniquement en compte que le prix des instances et non la capacité disponible, elle peut entraîner des taux d'interruption élevés.

Notez qu'Amazon EC2 Auto Scaling s'efforce de puiser les Instances Spot dans le nombre N de groupes que vous spécifiez. Cependant, si un pool manque de capacité Spot pour répondre à la capacité souhaitée, Amazon EC2 Auto Scaling continue à satisfaire la demande en puisant dans le pool le moins cher suivant. Pour atteindre la capacité souhaitée, vous pouvez recevoir des instances Spot de plus de groupes que votre nombre N spécifié. De même, si la majorité des pools ne disposent d'aucune capacité Spot, la totalité de la capacité souhaitée sera peut-être puisée à partir d'un nombre N de groupes inférieur à celui que vous avez spécifié.

> **Note**
>
> Si vous configurez vos instances Spot pour qu'elles soient lancées avec AMD SEV-SNP activé, des frais d'utilisation horaires supplémentaires vous seront facturés, équivalant à 10 % du taux horaire à la demande du type d'instance sélectionné. Si la stratégie d'allocation utilise le prix comme entrée, Amazon EC2 Auto Scaling n'inclut pas ces frais supplémentaires ; seul le prix Spot est utilisé.

On-Demand instances

Amazon EC2 Auto Scaling fournit les stratégies d'allocation suivantes pour les instances à la demande :

`lowest-price`

Amazon EC2 Auto Scaling déploie automatiquement le type d'instance le moins cher dans chaque zone de disponibilité en fonction du prix actuel des instances à la demande.

Pour garantir que la capacité souhaitée est atteinte, vous pouvez recevoir des instances à la demande de plus d'un type d'instance dans chaque zone de disponibilité. Cela dépend de la capacité que vous demandez.

`prioritized`

Pour satisfaire la capacité à la demande, Amazon EC2 Auto Scaling détermine quel type d'instance utiliser en premier en se fondant sur les types d'instance dans la liste des remplacements du modèle de lancement. Par exemple, vous avez spécifié trois remplacements de modèle de lancement dans l'ordre suivant : `c5.large`, `c4.large` et `c3.large`. Lors du lancement de vos instances à la demande, le groupe Auto Scaling satisfait la capacité à la demande dans l'ordre suivant : `c5.large`, puis `c4.large`, enfin `c3.large`.

Tenez compte des éléments suivants lorsque vous gérez l'ordre de priorité de vos instances à la demande :

- Vous pouvez payer votre utilisation à l'avance et bénéficier de réductions importantes sur les instances à la demande en utilisant des Savings Plans ou des instances réservées. Pour plus d'informations, consultez la page Amazon EC2 pricing (Tarification Amazon EC2).
- Avec les instances réservées, la réduction par rapport à la tarification standard des instances à la demande s'applique si Amazon EC2 Auto Scaling lance les types d'instances

correspondants. Cela signifie que si vous avez des instances réservées inutilisées pour `c4.large`, vous pouvez définir la priorité de vos types d'instance de manière à donner la priorité la plus élevée pour vos instances réservées à un type d'instance `c4.large`. Lorsqu'une instance `c4.large` est lancée, vous recevez la tarification des instances réservées.

- Avec les Savings Plans, la réduction par rapport à la tarification standard des instances à la demande s'applique lorsque vous utilisez Amazon EC2 Instance Savings Plans ou Compute Savings Plans. Avec Savings Plans, vous bénéficiez d'une plus grande flexibilité lors de la hiérarchisation de vos types d'instances. Tant que vous utilisez des types d'instances couverts par votre Savings Plan, vous pouvez les classer dans n'importe quel ordre de priorité. Vous pouvez également modifier occasionnellement l'ordre complet de vos types d'instances, tout en bénéficiant du tarif réduit du Savings Plan. Pour en savoir plus sur les Savings Plans, consultez le Guide de l'utilisateur des Savings Plans.

Comment les stratégies d'allocation fonctionnent avec les pondérations

Lorsque vous spécifiez le `WeightedCapacity` paramètre dans vos overrides (`"DesiredCapacityType": "vcpu"`ou `"DesiredCapacityType": "memory-mib"` au niveau du groupe), les stratégies d'allocation fonctionnent exactement comme elles le font pour les autres groupes Auto Scaling.

La seule différence est que lorsque vous choisissez la `price-capacity-optimized` stratégie `lowest-price` or, vos instances proviennent des pools d'instances dont le prix unitaire est le plus bas dans chaque zone de disponibilité. Pour plus d'informations, consultez Configurer un groupe Auto Scaling pour utiliser les poids d'instance.

Par exemple, imaginons que vous avez un groupe Auto Scaling qui a plusieurs types d'instances avec des quantités variables de vCPU. Vous utilisez `lowest-price` pour vos stratégies d'allocation Spot et à la demande. Si vous choisissez d'attribuer des pondérations basées sur le nombre de vCPU de chaque type d'instance, Amazon EC2 Auto Scaling lance les types d'instance ayant le prix le plus bas selon les valeurs de pondération que vous avez attribuées (par exemple, par vCPU) au moment de l'exécution. S'il s'agit d'une instance Spot, cela signifie le prix Spot le plus bas par vCPU. S'il s'agit d'une Instance à la demande, cela signifie le prix à la demande le plus bas par vCPU.

Créer un groupe d'instances mixtes en utilisant la sélection du type d'instance basée sur des attributs

Au lieu de choisir manuellement les types d'instance pour votre groupe d'instances mixtes, vous pouvez spécifier un ensemble d'attributs d'instance qui décrivent vos besoins en calcul. Lorsque

Amazon EC2 Auto Scaling lance des instances, tous les types d'instance utilisés par le groupe Auto Scaling doivent correspondre à vos attributs d'instance requis. C'est ce qu'on appelle la sélection de type d'instance basée sur des attributs.

Cette approche est idéale pour les charges de travail et les cadres qui peuvent être flexibles quant aux types d'instance qu'ils utilisent, comme les conteneurs, le big data et le CI/CD.

Voici les avantages de la sélection du type d'instance basée sur les attributs :

- Flexibilité optimale pour les instances Spot : Amazon EC2 Auto Scaling peut choisir parmi un large éventail de types d'instances pour le lancement d'instances Spot. Cela répond à la bonne pratique Spot d'être flexible sur les types d'instance, ce qui donne au service Amazon EC2 Spot une meilleure chance de trouver et d'allouer votre quantité requise de capacité de calcul.
- Utilisez facilement les bons types d'instances : compte tenu du grand nombre de types d'instances disponibles, la recherche des types d'instances adaptés à votre charge de travail peut prendre beaucoup de temps. Lorsque vous spécifiez des attributs d’instance, les types d’instance auront automatiquement les attributs requis pour votre charge de travail.
- Utilisation automatique de nouveaux types d'instances : vos groupes Auto Scaling peuvent utiliser des types d'instances de nouvelle génération au fur et à mesure de leur publication. Les types d'instance de nouvelle génération sont automatiquement utilisés lorsqu'ils correspondent à vos besoins et s'alignent sur les stratégies d'allocation que vous choisissez pour votre groupe Auto Scaling.

Rubriques

Fonctionnement de la sélection de type d’instance basée sur des attributs

Avec la sélection du type d'instance basée sur les attributs, au lieu de fournir une liste de types d'instances spécifiques, vous fournissez une liste des attributs d'instance dont vos instances ont besoin, tels que :

- Nombre de vCPU : nombre minimum et maximum de vCPU par instance.
- Mémoire : mémoire minimale et maximale GiBs par instance.
- Stockage local : s'il faut utiliser EBS ou des volumes de stockage d'instance pour le stockage local.
- Performances éclatantes : s'il faut utiliser la famille d'instances T, y compris les types T4g, T3a, T3 et T2.

De nombreuses options sont disponibles pour définir les exigences de votre instance. Pour une description de chaque option et des valeurs par défaut, consultez InstanceRequirementsle manuel Amazon EC2 Auto Scaling API Reference.

Lorsque votre groupe Auto Scaling doit lancer une instance, il recherche les types d'instances qui correspondent aux attributs que vous avez spécifiés et qui sont disponibles dans cette zone de disponibilité. La stratégie d'allocation détermine ensuite le type d'instance correspondant à lancer. Par défaut, la sélection du type d'instance basée sur les attributs comporte une fonctionnalité de protection des prix activée pour empêcher votre groupe Auto Scaling de lancer des types d'instances dépassant vos seuils budgétaires.

Par défaut, vous utilisez le nombre d'instances comme unité de mesure lorsque vous définissez la capacité souhaitée de votre groupe Auto Scaling, ce qui signifie que chaque instance compte pour une unité.

Vous pouvez également définir la valeur de la capacité souhaitée comme le nombre de vCPU ou la quantité de mémoire. Pour ce faire, utilisez le champ déroulant Type de capacité souhaité dans le champ AWS Management Console ou la `DesiredCapacityType` propriété dans l'opération `CreateAutoScalingGroup` ou `UpdateAutoScalingGroup` API. Amazon EC2 Auto Scaling lance ensuite le nombre d'instances nécessaires pour atteindre la capacité de vCPU ou de mémoire souhaitée. Par exemple, si vous utilisez des vCPU comme type de capacité souhaité et que vous utilisez des instances avec 2 vCPU chacune, une capacité souhaitée de 10 vCPU lancera 5 instances. Il s’agit d’une alternative utile aux pondérations d’instance.

Protection des prix

Grâce à la protection des prix, vous pouvez spécifier le prix maximum que vous êtes prêt à payer pour les instances EC2 lancées par votre groupe Auto Scaling. La protection des prix est une fonctionnalité qui empêche votre groupe Auto Scaling d'utiliser des types d'instances que vous jugeriez trop chers, même s'ils correspondent aux attributs que vous avez spécifiés.

La protection des prix est activée par défaut et comporte des seuils de prix distincts pour les instances à la demande et les instances ponctuelles. Lorsqu'Amazon EC2 Auto Scaling doit lancer de nouvelles instances, aucun type d'instance dont le prix est supérieur au seuil pertinent n'est lancé.

Rubriques

Protection des prix à la demande

Pour les instances à la demande, vous définissez le prix maximum à la demande que vous êtes prêt à payer sous forme de pourcentage supérieur au prix à la demande identifié. Le prix à la demande identifié est le prix du type d'instance C, M ou R de génération actuelle le moins cher avec les attributs que vous avez spécifiés.

Si une valeur de protection des prix à la demande n'est pas explicitement définie, un prix à la demande maximum par défaut supérieur de 20 % au prix à la demande identifié sera utilisé.

Protection des prix au comptant

Par défaut, Amazon EC2 Auto Scaling applique automatiquement une protection tarifaire optimale des instances Spot afin de sélectionner de manière cohérente un large éventail de types d'instances. Vous pouvez également définir vous-même la protection des prix manuellement. Toutefois, laisser Amazon EC2 Auto Scaling le faire à votre place peut améliorer les chances que votre capacité Spot soit atteinte.

Vous pouvez définir manuellement la protection des prix à l'aide de l'une des options suivantes. Si vous définissez manuellement la protection des prix, nous vous recommandons d'utiliser la première option.

- Pourcentage d'un prix à la demande identifié : le prix à la demande identifié est le prix du type d'instance C, M ou R de génération actuelle le moins cher avec les attributs que vous avez spécifiés.
- Un pourcentage supérieur au prix spot identifié : le prix spot identifié est le prix du type d'instance C, M ou R de génération actuelle le moins cher avec les attributs que vous avez spécifiés. Nous vous déconseillons d'utiliser cette option car les prix au comptant peuvent fluctuer et, par conséquent, votre seuil de protection contre les prix peut également fluctuer.

Personnalisez la protection des prix

Vous pouvez personnaliser les seuils de protection des prix dans la console Amazon EC2 Auto Scaling ou à l'aide des SDK AWS CLI .

- Dans la console, utilisez les paramètres de protection des prix à la demande et de protection des prix au comptant dans Attributs d'instance supplémentaires.
- Dans la InstanceRequirementsstructure, pour spécifier le seuil de protection des prix des instances à la demande, utilisez la `OnDemandMaxPricePercentageOverLowestPrice` propriété. Pour spécifier le seuil de protection des prix de l'instance Spot, utilisez la propriété `MaxSpotPriceAsPercentageOfOptimalOnDemandPrice` ou la `SpotMaxPricePercentageOverLowestPrice` propriété.

Si vous définissez le type de capacité souhaité (`DesiredCapacityType`) sur vCPU ou Gio de mémoire, la protection tarifaire s'applique en fonction du prix par vCPU ou par mémoire plutôt que du prix par instance.

Vous pouvez également désactiver la protection des prix. Pour n'indiquer aucun seuil de protection des prix, spécifiez un pourcentage élevé, tel que`999999`.

> **Note**
>
> Si aucun type d'instance C, M ou R de génération actuelle ne correspond aux attributs que vous avez spécifiés, la protection des prix reste applicable. Si aucune correspondance n'est trouvée, le prix identifié provient des types d'instances de la génération actuelle les moins chers ou, à défaut, des types d'instances de la génération précédente les moins chers, qui correspondent à vos attributs.

Prérequis

- Créer un modèle de lancement. Pour plus d'informations, consultez Créer un modèle de lancement pour un groupe Auto Scaling.
- Vérifiez que le modèle de lancement ne demande pas déjà des instances Spot.

Création d'un groupe d'instances mixtes avec sélection du type d'instance basée sur les attributs (console)

Utilisez la procédure suivante pour créer un groupe d'instances mixtes à l'aide de la sélection de type d'instance basée sur des attributs. Pour vous aider à suivre les étapes de manière efficace, certaines sections facultatives sont ignorées.

Pour la plupart des charges de travail polyvalentes, il suffit de spécifier le nombre de vCPU et de mémoire dont vous avez besoin. Pour les cas d'utilisation avancés, vous pouvez spécifier des attributs tels que le type de stockage, les interfaces réseau, le fabricant du CPU et le type d'accélérateur.

Pour consulter les meilleures pratiques relatives à un groupe d'instances mixtes, consultezPrésentation de la configuration.

Pour créer un groupe d'instances mixtes

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Dans la barre de navigation en haut de l'écran, choisissez la même région Région AWS que celle utilisée lors de la création du modèle de lancement.
3. Choisissez Create an Auto Scaling group (Créer un groupe Auto Scaling).
4. Dans la page Choisir un modèle de lancement ou une configuration, dans Nom du groupe Auto Scaling, entrez un nom pour le groupe Auto Scaling.
5. Pour choisir votre modèle de lancement, procédez comme suit :

    a. Dans Launch template (Modèle de lancement), choisissez un modèle de lancement existant.

    b. Pour Version du modèle de lancement, indiquez si le groupe Auto Scaling utilise la version par défaut, la version la plus récente ou une version spécifique du modèle de lancement lors de l'évolutivité horizontale.

c. Vérifiez que votre modèle de lancement prend en charge toutes les options que vous envisagez d'utiliser, puis choisissez Next (Suivant).

6. Sur la page Choisir les options de lancement d’instance, procédez comme suit :

a. Pour Instance type requirements (Exigences en matière de type d'instance), sélectionnez Override launch template (Remplacer le modèle de lancement).

> **Note**
>
> Si vous avez choisi un modèle de lancement qui contient déjà un ensemble d’attributs d’instance, tels que des vCPU et de la mémoire, les attributs d’instance sont affichés. Ces attributs sont ajoutés aux propriétés du groupe Auto Scaling, que vous pouvez mettre à jour à tout moment sur la console Amazon EC2 Auto Scaling.

b. Sous Specify instance attributes (Spécifier les attributs d'instance), commencez par saisir vos besoins en vCPU et en mémoire.

- Pour vCPU, saisissez les nombres minimum et maximum de vCPU souhaités. Pour ne spécifier aucune limite, sélectionnez No minimum (Pas de minimum), No maximum (Pas de maximum), ou les deux.
- Pour Memory (GiB) (Mémoire (Gio)), saisissez la quantité minimale et maximale de mémoire souhaitée. Pour ne spécifier aucune limite, sélectionnez No minimum (Pas de minimum), No maximum (Pas de maximum), ou les deux.

c. (Facultatif) Pour Additional instance attributes (Attributs d’instance supplémentaires), vous pouvez éventuellement spécifier un ou plusieurs attributs pour exprimer vos exigences de calcul plus en détail. Chaque attribut supplémentaire ajoute des contraintes supplémentaires à votre demande.

d. Développez Aperçu des types d'instances correspondants pour afficher les types d'instance dotés des attributs que vous avez spécifiés.

e. Dans Options d’achat d’instance, pour Distribution des instances, spécifiez les pourcentages du groupe à lancer en tant qu’instances à la demande et en tant qu’instances Spot. Si votre application est sans état, tolérante aux pannes et peut gérer l’interruption d’une instance, vous pouvez spécifier un pourcentage plus élevé d’Instances Spot.

f. (Facultatif) Lorsque vous spécifiez un pourcentage pour les instances Spot, sélectionnez Inclure la capacité de base à la demande, puis spécifiez la quantité minimale de la capacité initiale du groupe Auto Scaling qui doit être remplie par des instances à la demande. Tout

ce qui dépasse la capacité de base utilise les paramètres de distribution des instances pour déterminer le nombre d'instances à la demande et d'instances Spot à lancer.

g. Sous Allocation strategies (Stratégies d'allocation), le Lowest price (Prix le plus bas) est automatiquement sélectionné pour la On-Demand allocation strategy (Stratégie d'allocation à la demande) et ne peut pas être modifié.

h. Pour Spot allocation strategy (Stratégie d'allocation d'instances Spot), choisissez une stratégie d'allocation. Price capacity optimized (Capacité de prix optimisée) est sélectionné par défaut. Lowest price (Tarif le plus bas) est masqué par défaut et n'apparaît que lorsque vous choisissez Show all strategies (Afficher toutes les stratégies). Si vous choisissez Tarif le plus bas, saisissez le nombre de groupes de prix les plus bas pour diversifier les offres pour les Groupes de prix les plus bas.

i. Dans Rééquilibrage de la capacité, choisissez d'activer ou de désactiver le Rééquilibrage de la capacité. Utiliser le rééquilibrage de la capacité pour répondre automatiquement quand vos instances Spot sont sur le point de se résilier à cause d'une interruption des instances Spot. Pour plus d'informations, consultez Utiliser le rééquilibrage de la capacité pour gérer les interruptions Spot Amazon EC2.

j. Sous Network (Réseau), pour VPC, choisissez un VPC. Le groupe Auto Scaling doit être créé dans le même VPC que le groupe de sécurité que vous avez spécifié dans votre modèle de lancement.

k. Pour Availability Zones and subnets (Zones de disponibilité et sous-réseaux), choisissez un ou plusieurs sous-réseaux dans le VPC spécifié. Utilisez les sous-réseaux dans plusieurs zones de disponibilité pour une haute disponibilité. Pour plus d'informations, consultez Considérations à prendre en compte lors du choix des sous-réseaux VPC.

l. Appuyez sur Suivant, Suivant.

7. Pour l'étape Configure group size and scaling policies (Configurer la taille du groupe et les politiques de mise à l'échelle), procédez comme suit :

a. Si vous voulez que la capacité souhaitée soit mesurée en unités autres que des instances, choisissez l'option appropriée pour Taille du groupe et Type de capacité souhaité. Units (Unités), vCPUs (vCPU) et Memory GiB (Gio de mémoire) sont pris en charge. Par défaut, Amazon EC2 Auto Scaling spécifie Units (Unités), ce qui se traduit par le nombre d'instances.

b. Définissez la capacité souhaitée en fonction de la taille initiale de votre groupe Auto Scaling.

c. Dans la section Mise à l'échelle, sous Limites de mise à l'échelle, si votre nouvelle valeur pour la capacité souhaitée est supérieure à la capacité minimale souhaitée et à la capacité

maximale souhaitée, la capacité maximale souhaitée est automatiquement augmentée à la nouvelle valeur de capacité souhaitée. Vous pouvez modifier ces limites si nécessaire. Pour plus d'informations, consultez Définissez des limites de mise à l'échelle pour votre groupe Auto Scaling.

8. Choisissez Skip to review (Passer à la révision).
9. Sur la page Vérifier, sélectionnez Créer un groupe Auto Scaling.

Création d'un groupe d'instances mixtes avec sélection du type d'instance basée sur les attributs ()AWS CLI

Pour créer un groupe d'instances mixtes avec la ligne de commande

Utilisez l'une des commandes suivantes :

- create-auto-scaling-group (AWS CLI)
- AutoScalingGroupe New-AS (1)AWS Tools for Windows PowerShell

Exemple de configuration

Pour créer un groupe Auto Scaling avec une sélection de type d'instance basée sur des attributs en utilisant la AWS CLI, vous pouvez utiliser la commande suivante create-auto-scaling-group.

Les attributs d'instance suivants sont spécifiés :

- `VCpuCount` : les types d'instances doivent avoir un minimum de quatre vCPU et un maximum de huit vCPU.
- `MemoryMiB` : les types d'instance doivent avoir un minimum de 16 384 Mio de mémoire.
- `CpuManufacturers` : les types d'instance doivent avoir un processeur fabriqué par Intel.

JSON

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

Voici un exemple de fichier `config.json`.

```
{
    "AutoScalingGroupName": "my-asg",
    "DesiredCapacityType": "units",
```

```
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Default"
            },
            "Overrides": [{
                "InstanceRequirements": {
                    "VCpuCount": {"Min": 4, "Max": 8},
                    "MemoryMiB": {"Min": 16384},
                    "CpuManufacturers": ["intel"]
                }
            }]
        },
        "InstancesDistribution": {
            "OnDemandPercentageAboveBaseCapacity": 50,
            "SpotAllocationStrategy": "price-capacity-optimized"
        }
    },
    "MinSize": 0,
    "MaxSize": 100,
    "DesiredCapacity": 4,
    "DesiredCapacityType": "units",
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

Pour définir la valeur de la capacité souhaitée comme le nombre de vCPU ou la quantité de mémoire, spécifiez `"DesiredCapacityType": "vcpu"` ou `"DesiredCapacityType": "memory-mib"` dans le fichier. Le type de capacité souhaitée par défaut est `units`, qui définit la valeur de la capacité souhaitée comme le nombre d'instances.

YAML

Utilisez la commande create-auto-scaling-group suivante pour créer le groupe Auto Scaling. Cela fait référence à un fichier YAML comme seul paramètre de votre groupe Auto Scaling.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

Voici un exemple de fichier `config.yaml`.

```
---
AutoScalingGroupName: my-asg
```

```
DesiredCapacityType: units
MixedInstancesPolicy:
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceRequirements:
        VCpuCount:
          Min: 2
          Max: 4
        MemoryMiB:
          Min: 2048
        CpuManufacturers:
        - intel
  InstancesDistribution:
    OnDemandPercentageAboveBaseCapacity: 50
    SpotAllocationStrategy: price-capacity-optimized
MinSize: 0
MaxSize: 100
DesiredCapacity: 4
DesiredCapacityType: units
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Pour définir la valeur de la capacité souhaitée comme le nombre de vCPU ou la quantité de mémoire, spécifiez `DesiredCapacityType: vcpu` ou `DesiredCapacityType: memory-mib` dans le fichier. Le type de capacité souhaitée par défaut est `units`, qui définit la valeur de la capacité souhaitée comme le nombre d'instances.

Prévisualisez vos types d’instance

Vous pouvez prévisualiser les types d'instance qui correspondent à vos besoins de calcul sans les lancer et ajuster vos besoins si nécessaire. Lors de la création de votre groupe Auto Scaling dans la console Amazon EC2 Auto Scaling, une prévisualisation des types d'instance apparaît dans la section Preview matching instance types (Prévisualisation des types d'instance correspondants) sur la page Choose instance launch options (Choisir des options de lancement d'instance).

Vous pouvez également prévisualiser les types d'instances en effectuant un appel d'GetInstanceTypesFromInstanceRequirementsAPI Amazon EC2 à l'aide du AWS CLI ou d'un SDK. Passez les paramètres `InstanceRequirements` dans la demande dans le format exact que vous utiliseriez pour créer ou mettre à jour un groupe Auto Scaling. Pour plus d'informations, consultez

la section Aperçu des types d'instances avec des attributs spécifiés dans le guide de l'utilisateur Amazon EC2.

Ressources connexes

Pour en savoir plus sur la sélection du type d'instance basée sur les attributs, consultez la section Sélection du type d'instance basée sur les attributs pour EC2 Auto Scaling et EC2 Fleet sur le blog. AWS

Vous pouvez déclarer une sélection de type d'instance basée sur les attributs lorsque vous créez un groupe Auto Scaling avec AWS CloudFormation. Pour plus d'informations, consultez l'exemple d'extrait dans la section Extraits de modèle de mise à l'échelle automatique du Guide de l'utilisateur AWS CloudFormation .

Créer un groupe d'instances mixtes en choisissant manuellement les types d'instances

Cette rubrique explique comment lancer plusieurs types d'instances dans un seul groupe Auto Scaling en choisissant manuellement vos types d'instances.

Si vous préférez utiliser des attributs d'instance comme critères de sélection des types d'instance, consultez Créer un groupe d'instances mixtes en utilisant la sélection du type d'instance basée sur des attributs.

Table des matières

Prérequis

- Créer un modèle de lancement. Pour plus d'informations, consultez Créer un modèle de lancement pour un groupe Auto Scaling.
- Vérifiez que le modèle de lancement ne demande pas déjà des instances Spot.

Créer un groupe d’instances mixtes (console)

Utilisez la procédure suivante pour créer un groupe d’instances mixtes en choisissant manuellement les types d’instance que votre groupe peut lancer. Pour vous aider à suivre les étapes de manière efficace, certaines sections facultatives sont ignorées.

Pour consulter les meilleures pratiques relatives à un groupe d'instances mixtes, consultezPrésentation de la configuration.

Pour créer un groupe d’instances mixtes

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Dans la barre de navigation en haut de l'écran, choisissez la même région Région AWS que celle utilisée lors de la création du modèle de lancement.
3. Choisissez Create an Auto Scaling group (Créer un groupe Auto Scaling).
4. Dans la page Choisir un modèle de lancement ou une configuration, dans Nom du groupe Auto Scaling, entrez un nom pour le groupe Auto Scaling.
5. Pour choisir votre modèle de lancement, procédez comme suit :

    a. Dans Launch template (Modèle de lancement), choisissez un modèle de lancement existant.

    b. Pour Version du modèle de lancement, indiquez si le groupe Auto Scaling utilise la version par défaut, la version la plus récente ou une version spécifique du modèle de lancement lors de l'évolutivité horizontale.

    c. Vérifiez que votre modèle de lancement prend en charge toutes les options que vous envisagez d'utiliser, puis choisissez Next (Suivant).
6. Sur la page Choisir les options de lancement d’instance, procédez comme suit :

    a. Pour les Exigences relatives au type d'instance, choisissez Override launch template (Remplacer le modèle de lancement), puis Manually add instance types (Ajouter manuellement les types d'instance).

    b. Choisissez vos types d'instance. Vous pouvez utiliser nos recommandations comme point de départ. Family and generation flexible (Famille et génération flexibles) est sélectionnée par défaut.

- (Facultatif) Pour modifier l'ordre des types d'instances, utilisez les flèches. Si vous choisissez une stratégie d'allocation qui prend en charge la priorisation, l'ordre des types d'instance définit leur priorité de lancement.
- Pour supprimer un type d'instance, choisissez X.
- (Facultatif) Pour les cases de la colonne Poids, attribuez une pondération relative à chaque type d'instance. Pour ce faire, entrez le nombre d'unités qu'une instance de ce type compte par rapport à la capacité souhaitée du groupe. Cela peut s'avérer notamment utile si les types d'instance offrent des capacités différentes de vCPU, de mémoire, de stockage ou de bande passante du réseau. Pour plus d'informations, consultez Configurer un groupe Auto Scaling pour utiliser les poids d'instance.

  Notez que si vous choisissez d'utiliser les recommandations de Taille flexible, tous les types d'instance qui font partie de cette section ont automatiquement une valeur de pondération. Si vous ne souhaitez pas spécifier de pondération, décochez les cases de la colonne Weight (Poids) pour tous les types d'instances.

c. Dans Instance purchase options (Options d'achat d'instance), pour Instance distribution (Distribution des instances), spécifiez les pourcentages du groupe à lancer en tant qu'instances à la demande et en tant qu'instances Spot, respectivement. Si votre application est sans état, tolérante aux pannes et peut gérer l'interruption d'une instance, vous pouvez spécifier un pourcentage plus élevé d'Instances Spot.

d. (Facultatif) Lorsque vous spécifiez un pourcentage pour les instances Spot, sélectionnez Inclure la capacité de base à la demande, puis spécifiez la quantité minimale de la capacité initiale du groupe Auto Scaling qui doit être remplie par des instances à la demande. Tout ce qui dépasse la capacité de base utilise les paramètres de distribution des instances pour déterminer le nombre d'instances à la demande et d'instances Spot à lancer.

e. Sous Allocation strategies (Stratégies d'allocation), pour On-Demand allocation strategy (Stratégie d'allocation à la demande), choisissez une stratégie d'allocation. Lorsque vous choisissez manuellement vos types d'instances, Prioritized (Priorisé) est sélectionné par défaut.

f. Pour Spot allocation strategy (Stratégie d'allocation d'instances Spot), choisissez une stratégie d'allocation. Price capacity optimized (Capacité de prix optimisée) est sélectionné par défaut. Lowest price (Tarif le plus bas) est masqué par défaut et n'apparaît que lorsque vous choisissez Show all strategies (Afficher toutes les stratégies).

- Si vous choisissez Tarif le plus bas, saisissez le nombre de groupes de prix les plus bas pour diversifier les offres pour les Groupes de prix les plus bas.
- Si vous choisissez Capacité optimisée, vous pouvez éventuellement cocher la case Prioriser les types d'instances pour permettre à Amazon EC2 Auto Scaling de choisir le type d'instance à lancer en premier en fonction de l'ordre dans lequel vos types d'instances sont répertoriés.

g. Dans Rééquilibrage de la capacité, choisissez d'activer ou de désactiver le Rééquilibrage de la capacité. Utiliser le rééquilibrage de la capacité pour répondre automatiquement quand vos instances Spot sont sur le point de se résilier à cause d'une interruption des instances Spot. Pour plus d'informations, consultez Utiliser le rééquilibrage de la capacité pour gérer les interruptions Spot Amazon EC2.

h. Sous Network (Réseau), pour VPC, choisissez un VPC. Le groupe Auto Scaling doit être créé dans le même VPC que le groupe de sécurité que vous avez spécifié dans votre modèle de lancement.

i. Pour Availability Zones and subnets (Zones de disponibilité et sous-réseaux), choisissez un ou plusieurs sous-réseaux dans le VPC spécifié. Utilisez les sous-réseaux dans plusieurs zones de disponibilité pour une haute disponibilité. Pour plus d'informations, consultez Considérations à prendre en compte lors du choix des sous-réseaux VPC.

j. Appuyez sur Suivant, Suivant.

7. Pour l'étape Configure group size and scaling policies (Configurer la taille du groupe et les politiques de mise à l'échelle), procédez comme suit :

   a. Dans Taille du groupe, pour la Capacité souhaitée, entrez le nombre initial d'instances à lancer.

      Par défaut, la capacité souhaitée est exprimée en nombre d'instances. Si vous avez attribué des pondérations à vos types d'instances, vous devez convertir cette valeur en la même unité de mesure que celle que vous avez utilisée pour attribuer des poids, par exemple le nombre de vCPU.

   b. Dans la section Mise à l'échelle, sous Limites de mise à l'échelle, si votre nouvelle valeur pour la capacité souhaitée est supérieure à la capacité minimale souhaitée et à la capacité maximale souhaitée, la capacité maximale souhaitée est automatiquement augmentée à la nouvelle valeur de capacité souhaitée. Vous pouvez modifier ces limites si nécessaire. Pour plus d'informations, consultez Définissez des limites de mise à l'échelle pour votre groupe Auto Scaling.

8. Choisissez Skip to review (Passer à la révision).
9. Sur la page Vérifier, sélectionnez Créer un groupe Auto Scaling.

Créer un groupe d'instances mixtes (AWS CLI)

Pour créer un groupe d'instances mixtes avec la ligne de commande

Utilisez l'une des commandes suivantes :

- create-auto-scaling-group (AWS CLI)
- AutoScalingGroupe New-AS (1)AWS Tools for Windows PowerShell

Exemples de configuration

Les exemples de configuration suivants montrent comment créer des instances mixtes à l'aide des différentes stratégies d'allocation Spot.

> **Note**
>
> Ces exemples montrent comment utiliser un fichier de configuration au format JSON ou YAML. Si vous utilisez AWS CLI la version 1, vous devez spécifier un fichier de configuration au format JSON. Si vous utilisez AWS CLI la version 2, vous pouvez spécifier un fichier de configuration au format YAML ou JSON.

Exemples

- Exemple 1 : lancer des instances Spot à l'aide de la stratégie d'allocation capacity-optimized
- Exemple 2 : lancer des instances Spot à l'aide de la stratégie d'allocation capacity-optimized-prioritized
- Exemple 3 : lancer des instances Spot à l'aide de la stratégie d'allocation lowest-price diversifiée sur deux pools
- Exemple 4 : Lancer instances Spot à l'aide de la stratégie d'allocation price-capacity-optimized

Exemple 1 : lancer des instances Spot à l'aide de la stratégie d'allocation `capacity-optimized`

La commande create-auto-scaling-group suivante crée un groupe Auto Scaling qui spécifie les éléments suivants :

- Pourcentage du groupe à lancer en tant qu'instances à la demande (`0`) et nombre de base d'instances à la demande (`1`)
- Types d'instance à lancer par ordre de priorité (`c5.large`, `c5a.large`, `m5.large`, `m5a.large`, `c4.large`, `m4.large`, `c3.large`, `m3.large`)
- Les sous-réseaux dans lesquels lancer les instances (`subnet-5ea0c127`, `subnet-6194ea3b`, `subnet-c934b782`) Chacun d'eux correspond à une zone de disponibilité différente.
- Modèle de lancement (`my-launch-template`) et version du modèle de lancement (`$Default`)

Lorsqu'Amazon EC2 Auto Scaling tente de satisfaire votre capacité à la demande, il lance d'abord le type d'instance `c5.large`. Les instances Spot proviennent du pool d'instances Spot optimal de chaque zone de disponibilité en fonction de la capacité d'instances Spot.

JSON

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

Le fichier `config.json` contient le contenu suivant.

```
{
    "AutoScalingGroupName": "my-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Default"
            },
            "Overrides": [
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                },
                {
                    "InstanceType": "m5.large"
                },
                {
                    "InstanceType": "m5a.large"
                },
                {
```

```
                    "InstanceType": "c4.large"
                },
                {
                    "InstanceType": "m4.large"
                },
                {
                    "InstanceType": "c3.large"
                },
                {
                    "InstanceType": "m3.large"
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandBaseCapacity": 1,
            "OnDemandPercentageAboveBaseCapacity": 0,
            "SpotAllocationStrategy": "capacity-optimized"
        }
    },
    "MinSize": 1,
    "MaxSize": 5,
    "DesiredCapacity": 3,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

YAML

Utilisez la commande create-auto-scaling-group suivante pour créer le groupe Auto Scaling. Cela fait référence à un fichier YAML comme seul paramètre de votre groupe Auto Scaling.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

Le fichier `config.yaml` contient le contenu suivant.

```
---
AutoScalingGroupName: my-asg
MixedInstancesPolicy:
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
```

```
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
  InstancesDistribution:
    OnDemandBaseCapacity: 1
    OnDemandPercentageAboveBaseCapacity: 0
    SpotAllocationStrategy: capacity-optimized
MinSize: 1
MaxSize: 5
DesiredCapacity: 3
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Exemple 2 : lancer des instances Spot à l'aide de la stratégie d'allocation **`capacity-optimized-prioritized`**

La commande create-auto-scaling-group suivante crée un groupe Auto Scaling qui spécifie les éléments suivants :

- Pourcentage du groupe à lancer en tant qu'instances à la demande (`0`) et nombre de base d'instances à la demande (1)
- Types d'instance à lancer par ordre de priorité (`c5.large`, `c5a.large`, `m5.large`, `m5a.large`, `c4.large`, `m4.large`, `c3.large`, `m3.large`)
- Les sous-réseaux dans lesquels lancer les instances (`subnet-5ea0c127`, `subnet-6194ea3b`, `subnet-c934b782`) Chacun d'eux correspond à une zone de disponibilité différente.
- Modèle de lancement (`my-launch-template`) et version du modèle de lancement (`$Latest`)

Lorsqu'Amazon EC2 Auto Scaling tente de satisfaire votre capacité à la demande, il lance d'abord le type d'instance `c5.large`. Lorsqu'Amazon EC2 Auto Scaling tente de satisfaire votre capacité Spot, il implémente au mieux les priorités relatives aux types d'instances sur la base du meilleur effort. Cependant, il optimise d'abord la capacité.

JSON

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

Le fichier `config.json` contient le contenu suivant.

```
{
    "AutoScalingGroupName": "my-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "Overrides": [
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                },
                {
                    "InstanceType": "m5.large"
                },
                {
                    "InstanceType": "m5a.large"
                },
                {
                    "InstanceType": "c4.large"
                },
                {
                    "InstanceType": "m4.large"
                },
                {
                    "InstanceType": "c3.large"
                },
                {
                    "InstanceType": "m3.large"
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandBaseCapacity": 1,
            "OnDemandPercentageAboveBaseCapacity": 0,
            "SpotAllocationStrategy": "capacity-optimized-prioritized"
        }
    },
    "MinSize": 1,
```

```
    "MaxSize": 5,
    "DesiredCapacity": 3,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

YAML

Utilisez la commande create-auto-scaling-group suivante pour créer le groupe Auto Scaling. Cela fait référence à un fichier YAML comme seul paramètre de votre groupe Auto Scaling.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

Le fichier `config.yaml` contient le contenu suivant.

```
---
AutoScalingGroupName: my-asg
MixedInstancesPolicy:
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
  InstancesDistribution:
    OnDemandBaseCapacity: 1
    OnDemandPercentageAboveBaseCapacity: 0
    SpotAllocationStrategy: capacity-optimized-prioritized
MinSize: 1
MaxSize: 5
DesiredCapacity: 3
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Exemple 3 : lancer des instances Spot à l'aide de la stratégie d'allocation **`lowest-price`** diversifiée sur deux pools

La commande create-auto-scaling-group suivante crée un groupe Auto Scaling qui spécifie les éléments suivants :

- Pourcentage du groupe à lancer en tant qu'instances à la demande (`50`). (Cela ne spécifie pas de nombre de base d'instances à la demande pour commencer.)
- Types d'instance à lancer par ordre de priorité (`c5.large`, `c5a.large`, `m5.large`, `m5a.large`, `c4.large`, `m4.large`, `c3.large`, `m3.large`)
- Les sous-réseaux dans lesquels lancer les instances (`subnet-5ea0c127`, `subnet-6194ea3b`, `subnet-c934b782`) Chacun d'eux correspond à une zone de disponibilité différente.
- Modèle de lancement (`my-launch-template`) et version du modèle de lancement (`$Latest`)

Lorsqu'Amazon EC2 Auto Scaling tente de satisfaire votre capacité à la demande, il lance d'abord le type d'instance `c5.large`. Pour votre capacité Spot, Amazon EC2 Auto Scaling tente de lancer les instances Spot uniformément sur les deux pools les moins chers de chaque zone de disponibilité.

JSON

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

Le fichier `config.json` contient le contenu suivant.

```
{
    "AutoScalingGroupName": "my-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "Overrides": [
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                },
```

```
                {
                    "InstanceType": "m5.large"
                },
                {
                    "InstanceType": "m5a.large"
                },
                {
                    "InstanceType": "c4.large"
                },
                {
                    "InstanceType": "m4.large"
                },
                {
                    "InstanceType": "c3.large"
                },
                {
                    "InstanceType": "m3.large"
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandPercentageAboveBaseCapacity": 50,
            "SpotAllocationStrategy": "lowest-price",
            "SpotInstancePools": 2
        }
    },
    "MinSize": 1,
    "MaxSize": 5,
    "DesiredCapacity": 3,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

YAML

Utilisez la commande create-auto-scaling-group suivante pour créer le groupe Auto Scaling. Cela fait référence à un fichier YAML comme seul paramètre de votre groupe Auto Scaling.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

Le fichier `config.yaml` contient le contenu suivant.

```
---
```

```
AutoScalingGroupName: my-asg
MixedInstancesPolicy:
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
  InstancesDistribution:
    OnDemandPercentageAboveBaseCapacity: 50
    SpotAllocationStrategy: lowest-price
    SpotInstancePools: 2
MinSize: 1
MaxSize: 5
DesiredCapacity: 3
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Exemple 4 : Lancer instances Spot à l'aide de la stratégie d'allocation **`price-capacity-optimized`**

La commande create-auto-scaling-group suivante crée un groupe Auto Scaling qui spécifie les éléments suivants :

- Pourcentage du groupe à lancer en tant qu'instances à la demande (`30`). (Cela ne spécifie pas de nombre de base d'instances à la demande pour commencer.)
- Types d'instance à lancer par ordre de priorité (`c5.large`, `c5a.large`, `m5.large`, `m5a.large`, `c4.large`, `m4.large`, `c3.large`, `m3.large`)
- Les sous-réseaux dans lesquels lancer les instances (`subnet-5ea0c127`, `subnet-6194ea3b`, `subnet-c934b782`) Chacun d'eux correspond à une zone de disponibilité différente.
- Modèle de lancement (`my-launch-template`) et version du modèle de lancement (`$Latest`)

Lorsqu'Amazon EC2 Auto Scaling tente de satisfaire votre capacité à la demande, il lance d'abord le type d'instance `c5.large`. Pour votre capacité Spot, Amazon EC2 Auto Scaling tente de lancer les

instances Spot à partir des groupes d'instances Spot au prix le plus bas possible, mais également avec une capacité optimale pour le nombre d'instances que vous lancez.

JSON

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

Le fichier `config.json` contient le contenu suivant.

```
{
    "AutoScalingGroupName": "my-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "Overrides": [
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                },
                {
                    "InstanceType": "m5.large"
                },
                {
                    "InstanceType": "m5a.large"
                },
                {
                    "InstanceType": "c4.large"
                },
                {
                    "InstanceType": "m4.large"
                },
                {
                    "InstanceType": "c3.large"
                },
                {
                    "InstanceType": "m3.large"
                }
            ]
```

```
        },
        "InstancesDistribution": {
            "OnDemandPercentageAboveBaseCapacity": 30,
            "SpotAllocationStrategy": "price-capacity-optimized"
        }
    },
    "MinSize": 1,
    "MaxSize": 5,
    "DesiredCapacity": 3,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

YAML

Utilisez la commande create-auto-scaling-group suivante pour créer le groupe Auto Scaling. Cela fait référence à un fichier YAML comme seul paramètre de votre groupe Auto Scaling.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

Le fichier `config.yaml` contient le contenu suivant.

```
---
AutoScalingGroupName: my-asg
MixedInstancesPolicy:
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
  InstancesDistribution:
    OnDemandPercentageAboveBaseCapacity: 30
    SpotAllocationStrategy: price-capacity-optimized
MinSize: 1
MaxSize: 5
DesiredCapacity: 3
```

```
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Configurer un groupe Auto Scaling pour utiliser les poids d'instance

Lorsque vous utilisez plusieurs types d'instances, vous pouvez spécifier le nombre d'unités à associer à chaque type d'instance, puis spécifier la capacité de votre groupe avec la même unité de mesure. Cette option de spécification de capacité est connue sous le nom de poids.

Par exemple, supposons que vous exécutez une application gourmande en calcul qui fonctionne mieux avec au moins 8 vCPU et 15 Go de RAM. Si vous utilisez `c5.2xlarge` comme unité de base, l'un des types d'instance EC2 suivants répondrait aux besoins de votre application.

Exemple de types d'instances

Type d’instance	vCPU	Mémoire (Gio)
`c5.2xlarge`	8	16
`c5.4xlarge`	16	32
`c5.12xlarge`	48	96
`c5.18xlarge`	72	144
`c5.24xlarge`	96	192

Par défaut, tous les types d’instances ont le même poids, quelle que soit leur taille. En d’autres termes, quelle que soit la taille du type d’instance lancé par Amazon EC2 Auto Scaling, chaque instance est comptabilisée de la même façon dans la capacité souhaitée du groupe Auto Scaling.

Cependant, avec les pondérations, vous attribuez une valeur numérique qui indique le nombre d'unités à associer à chaque type d'instance. Par exemple, si les instances sont de tailles différentes, une instance `c5.2xlarge` peut avoir un poids de 2, et une `c5.4xlarge` (qui est deux fois plus grande) peut avoir un poids de 4, et ainsi de suite. Ensuite, lorsqu’Amazon EC2 Auto Scaling redimensionne le groupe, ces pondérations se traduisent par le nombre d’unités que chaque instance compte pour votre capacité souhaitée.

Les pondérations ne changent pas les types d’instance qu’Amazon EC2 Auto Scaling choisit de lancer ; ce choix revient aux stratégies d’allocation. Pour plus d’informations, consultez Stratégies d’allocation.

> **Important**
>
> Pour configurer un groupe Auto Scaling afin qu'il atteigne la capacité souhaitée en fonction du nombre de vCPU ou de la quantité de mémoire de chaque type d'instance, nous vous recommandons d'utiliser la sélection du type d'instance basée sur les attributs. La définition du `DesiredCapacityType` paramètre indique automatiquement le nombre d'unités à associer à chaque type d'instance en fonction de la valeur que vous avez définie pour ce paramètre. Pour plus d'informations, consultez Créer un groupe d'instances mixtes en utilisant la sélection du type d'instance basée sur des attributs.

Table des matières

Considérations

Cette section aborde les principales considérations relatives à la mise en œuvre efficace des pondérations.

- Choisissez quelques types d'instances qui répondent aux besoins de performance de votre application. Déterminez le poids que chaque type d'instance doit prendre en compte dans la capacité souhaitée de votre groupe Auto Scaling en fonction de ses capacités. Ces pondérations s'appliquent aux instances actuelles et futures.
- Évitez les grandes plages de poids. Par exemple, ne spécifiez pas un poids de 1 pour un type d'instance lorsque le type d'instance le plus important suivant a un poids de 200. La différence entre les pondérations des plus petites et des plus grandes ne doit pas non plus être extrême. Les différences de poids extrêmes peuvent avoir un impact négatif sur l'optimisation des coûts et des performances.
- Spécifiez la capacité souhaitée du groupe en unités et non en instances. Par exemple, si vous utilisez des pondérations basées sur le processeur virtuel, définissez le nombre de cœurs souhaité ainsi que le minimum et le maximum.
- Définissez vos pondérations et la capacité souhaitée de sorte que cette dernière soit au moins deux à trois fois supérieure à votre pondération la plus importante.

Lorsque vous mettez à jour des groupes existants, tenez compte des points suivants :

- Lorsque vous ajoutez des pondérations à un groupe existant, incluez des pondérations pour tous les types d'instances actuellement utilisés.
- Lorsque vous ajoutez ou modifiez des poids, Amazon EC2 Auto Scaling lance ou arrête des instances pour atteindre la capacité souhaitée en fonction des nouvelles valeurs de pondération.
- Si vous supprimez un type d'instance, les instances de ce type en cours d'exécution conservent leur dernier poids, même si elles ne sont plus définies.

Comportements de poids des instances

Lorsque vous utilisez des pondérations d'instance, Amazon EC2 Auto Scaling se comporte de la manière suivante :

- La capacité actuelle sera soit à la capacité désirée, soit au-dessus. La capacité actuelle peut dépasser la capacité souhaitée si les instances lancées dépassent les unités de capacité souhaitées restantes. Supposons que vous spécifiez deux types d'instance `c5.2xlarge` et `c5.12xlarge`, et que vous affectez des pondérations d'instance de 2 pour `c5.2xlarge` et de 12 pour `c5.12xlarge`. S'il reste cinq unités pour satisfaire la capacité souhaitée, et qu'Amazon EC2 Auto Scaling approvisionne une instance `c5.12xlarge`, la capacité souhaitée est dépassée de sept unités.
- Lors du lancement d'instances, Amazon EC2 Auto Scaling donne la priorité à la distribution de la capacité entre les zones de disponibilité et au respect des stratégies d'allocation plutôt qu'au dépassement de la capacité souhaitée.
- Amazon EC2 Auto Scaling peut dépasser la limite de capacité maximale afin de maintenir l'équilibre entre les zones de disponibilité, en utilisant vos stratégies d'allocation préférées. La limite stricte imposée par Amazon EC2 Auto Scaling est la capacité souhaitée plus votre poids le plus élevé.

Configurer un groupe Auto Scaling pour utiliser des pondérations

Vous pouvez configurer un groupe Auto Scaling afin d'utiliser des pondérations, comme indiqué dans les exemples AWS CLI suivants. Pour des instructions sur l'utilisation de la console, consultez Créer un groupe d'instances mixtes en choisissant manuellement les types d'instances.

Pour configurer un nouveau groupe Auto Scaling afin d'utiliser des pondérations (AWS CLI)

Utilisez la commande create-auto-scaling-group. Par exemple, la commande suivante crée un nouveau groupe Auto Scaling et attribue des pondérations en spécifiant ce qui suit :

- Pourcentage du groupe à lancer en tant qu'instances à la demande (`0`)
- Stratégie d'allocation des instances Spot dans chaque zone de disponibilité (`capacity-optimized`)
- Types d'instance à lancer par ordre de priorité (`m4.16xlarge`, `m5.24xlarge`)
- Pondérations d'instance correspondant à la différence de taille relative (vCPU) entre les types d'instance (16, 24)
- Sous-réseaux dans lesquels lancer les instances (`subnet-5ea0c127`, `subnet-6194ea3b`, `subnet-c934b782`) chacun correspondant à une zone de disponibilité différente
- Modèle de lancement (`my-launch-template`) et version du modèle de lancement (`$Latest`)

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

Le fichier `config.json` contient le contenu suivant.

```
{
    "AutoScalingGroupName": "my-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "Overrides": [
                {
                    "InstanceType": "m4.16xlarge",
                    "WeightedCapacity": "16"
                },
                {
                    "InstanceType": "m5.24xlarge",
                    "WeightedCapacity": "24"
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandPercentageAboveBaseCapacity": 0,
            "SpotAllocationStrategy": "capacity-optimized"
```

```
        }
    },
    "MinSize": 160,
    "MaxSize": 720,
    "DesiredCapacity": 480,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
    "Tags": []
}
```

Pour configurer un groupe Auto Scaling existant afin d'utiliser des pondérations (AWS CLI)

Utilisez la commande update-auto-scaling-group. Par exemple, la commande suivante attribue des pondérations aux types d'instances d'un groupe Auto Scaling existant en spécifiant ce qui suit :

- Types d'instance à lancer par ordre de priorité (`c5.18xlarge`, `c5.24xlarge`, `c5.2xlarge`, `c5.4xlarge`)
- Pondérations d'instance correspondant à la différence de taille relative (vCPU) entre les types d'instance (18, 24, 2, 4)
- Le nouveau, augmentation de la capacité désirée, qui est plus important que le pondération la plus importante

```
aws autoscaling update-auto-scaling-group --cli-input-json file://~/config.json
```

Le fichier `config.json` contient le contenu suivant.

```
{
    "AutoScalingGroupName": "my-existing-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "Overrides": [
                {
                    "InstanceType": "c5.18xlarge",
                    "WeightedCapacity": "18"
                },
                {
                    "InstanceType": "c5.24xlarge",
                    "WeightedCapacity": "24"
                },
                {
                    "InstanceType": "c5.2xlarge",
```

```
                    "WeightedCapacity": "2"
                },
                {
                    "InstanceType": "c5.4xlarge",
                    "WeightedCapacity": "4"
                }
            ]
        }
    },
    "MinSize": 0,
    "MaxSize": 100,
    "DesiredCapacity": 100
}
```

Pour vérifier les pondérations à l'aide de la ligne de commande

Utilisez l'une des commandes suivantes :

- describe-auto-scaling-groups (AWS CLI)
- AutoScalingGroupe Get-AS ()AWS Tools for Windows PowerShell

Exemple de prix Spot par heure d'unité

Le tableau suivant compare le prix horaire des instances Spot dans différentes zones de disponibilité de la région USA Est (Virginie du Nord) avec le prix des instances à la demande dans la même région. Les prix affichés sont des exemples de prix et non des prix en réels. Ce sont vos coûts par heure d'instance.

Exemple : prix Spot par heure d'instance

Type d'instance	us-east-1a	us-east-1b	us-east-1c	Tarification à la demande
c5.2xlarge	$0.180	$0.191	$0.170	$0.34
c5.4xlarge	$0.341	$0.361	$0.318	$0.68
c5.12xlarge	$0.779	$0.777	$0.777	$2.04

Type d’instance	us-east-1a	us-east-1b	us-east-1c	Tarification à la demande
c5.18xlarge	$1.207	$1.475	$1.357	$3.06
c5.24xlarge	$1.555	$1.555	$1.555	$4.08

Avec les pondérations d’instance, vous pouvez évaluer vos coûts en fonction de ce que vous utilisez par heure d’unité. Vous pouvez déterminer le prix par heure d'unité en divisant le prix pour un type d'instance par le nombre d'unités qu'il représente. Pour les instances à la demande, le prix par heure d'unité est le même lors du déploiement d'un type d'instance que lors du déploiement d'une taille différente du même type d'instance. Par contre, le prix Spot par heure d'unité varie en fonction du groupe d'instances Spot.

L’exemple suivant montre comment le calcul du prix Spot par unité d’heure fonctionne avec les pondérations d’instance. Pour faciliter le calcul, supposons que vous souhaitez lancer des instances Spot uniquement dans la région `us-east-1a`. Le prix par heure unitaire est indiqué dans le tableau suivant.

Exemple : prix Spot par heure d'unité

Type d’instance	us-east-1a	Pondération de l’instance	Prix par heure d’unité
c5.2xlarge	$0.180	2	$0.090
c5.4xlarge	$0.341	4	$0.085
c5.12xlarge	$0.779	12	$0.065
c5.18xlarge	$1.207	18	$0.067
c5.24xlarge	$1.555	24	$0.065

Utiliser un modèle de lancement différent pour un type d’instance

Outre l’utilisation de plusieurs types d’instance, vous pouvez également vous servir de plusieurs modèles de lancement.

Par exemple, supposons que vous configuriez un groupe Auto Scaling pour les applications à forte intensité de calcul et que vous souhaitiez inclure une combinaison de types d’instances C5, C5a et C6g. Cependant, les instances C6g sont équipées d'un processeur AWS Graviton basé sur l'architecture Arm 64 bits, tandis que les instances C5 et C5a fonctionnent sur des processeurs Intel x86 64 bits. Les AMI des instances C5 et C5a fonctionnent sur chacune d’elles, mais pas sur les instances C6g. Pour résoudre ce problème, utilisez un modèle de lancement différent pour les instances C6g. Vous pouvez toujours utiliser le même modèle de lancement pour les instances C5 et C5a.

Cette section décrit les procédures d'utilisation du pour AWS CLI effectuer des tâches liées à l'utilisation de plusieurs modèles de lancement. Actuellement, cette fonction n'est disponible que si vous utilisez l'interface AWS CLI ou un kit SDK. Elle n'est pas disponible à partir de la console.

Table des matières

Configurer un groupe Auto Scaling afin d’utiliser plusieurs modèles de lancement

Vous pouvez configurer un groupe Auto Scaling afin d’utiliser plusieurs modèles de lancement, comme indiqué dans les exemples suivants.

Pour configurer un nouveau groupe Auto Scaling afin d’utiliser plusieurs modèles de lancement (AWS CLI)

Utilisez la commande create-auto-scaling-group. Par exemple, la commande suivante crée un nouveau groupe Auto Scaling. Celui-ci spécifie les types d'instances `c5.large`, `c5a.large` et `c6g.large`, et définit un nouveau modèle de lancement pour le type d'instance `c6g.large` afin qu'une AMI appropriée soit utilisée pour lancer les instances Arm. Amazon EC2 Auto Scaling s'appuie sur l'ordre des types d'instances afin de déterminer le type d'instance à utiliser en premier pour satisfaire la capacité à la demande.

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

Le fichier `config.json` contient le contenu suivant.

```
{
  "AutoScalingGroupName":"my-asg",
  "MixedInstancesPolicy":{
    "LaunchTemplate":{
      "LaunchTemplateSpecification":{
        "LaunchTemplateName":"my-launch-template-for-x86",
        "Version":"$Latest"
      },
      "Overrides":[
        {
          "InstanceType":"c6g.large",
          "LaunchTemplateSpecification": {
            "LaunchTemplateName": "my-launch-template-for-arm",
            "Version": "$Latest"
          }
        },
        {
          "InstanceType":"c5.large"
        },
        {
          "InstanceType":"c5a.large"
        }
      ]
    },
    "InstancesDistribution":{
      "OnDemandBaseCapacity": 1,
      "OnDemandPercentageAboveBaseCapacity": 50,
      "SpotAllocationStrategy": "capacity-optimized"
    }
  },
  "MinSize":1,
  "MaxSize":5,
  "DesiredCapacity":3,
  "VPCZoneIdentifier":"subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
  "Tags":[ ]
}
```

Pour configurer un groupe Auto Scaling existant afin d'utiliser plusieurs modèles de lancement (AWS CLI)

Utilisez la commande update-auto-scaling-group. Par exemple, la commande suivante attribue le modèle de lancement intitulé *my-launch-template-for-arm* au type d'instance *c6g.large* du groupe Auto Scaling intitulé *my-asg*.

```
aws autoscaling update-auto-scaling-group --cli-input-json file://~/config.json
```

Le fichier `config.json` contient le contenu suivant.

```
{
  "AutoScalingGroupName":"my-asg",
  "MixedInstancesPolicy":{
    "LaunchTemplate":{
      "Overrides":[
        {
          "InstanceType":"c6g.large",
          "LaunchTemplateSpecification": {
            "LaunchTemplateName": "my-launch-template-for-arm",
            "Version": "$Latest"
          }
        },
        {
          "InstanceType":"c5.large"
        },
        {
          "InstanceType":"c5a.large"
        }
      ]
    }
  }
}
```

Pour vérifier les modèles de lancement d'un groupe Auto Scaling

Utilisez l'une des commandes suivantes :

- describe-auto-scaling-groups (AWS CLI)
- AutoScalingGroupe Get-AS ()AWS Tools for Windows PowerShell

Ressources connexes

Vous trouverez un exemple de spécification de plusieurs modèles de lancement à l'aide de la sélection du type d'instance basée sur les attributs dans un AWS CloudFormation modèle sur AWS re:Post.

Créer des groupes Auto Scaling à l’aide de configurations de lancement

> **Important**
>
> Vous ne pouvez pas appeler `CreateLaunchConfiguration` avec les nouveaux types d’instances Amazon EC2 publiés après le 31 décembre 2022. De plus, les nouveaux comptes créés le 1er juin 2023 ou après cette date n’auront pas la possibilité de créer de nouvelles configurations de lancement via la console. À l'avenir, les nouveaux comptes ne pourront pas créer de nouvelles configurations de lancement à l'aide de la console, de l'API, de la CLI et CloudFormation. Effectuez la migration vers des modèles de lancement pour vous assurer de ne pas avoir à créer de nouvelles configurations de lancement maintenant ou à l'avenir. Pour plus d’informations sur la migration de vos groupes Auto Scaling vers les modèles de lancement, consultez la section Migrez vos groupes Auto Scaling pour lancer des modèles.

Si vous avez créé une configuration de lancement ou une instance EC2, vous pouvez créer un groupe Auto Scaling qui utilise la configuration de lancement comme modèle de configuration pour ses instances EC2. La configuration de lancement spécifie plusieurs informations, notamment l’identifiant d’AMI, le type d’instance, la paire de clés, les groupes de sécurité et le mappage de périphérique de stockage en mode bloc pour les instances. Pour plus d’informations sur la création des configurations de lancement, consultez la section Créez une configuration de lancement.

Vous devez disposer des autorisations nécessaires pour créer un groupe Auto Scaling. Vous devez également disposer des autorisations nécessaires pour créer un rôle lié à un service qu’Amazon EC2 Auto Scaling utilise pour effectuer les actions en votre nom, s’il n’existe pas encore. Pour obtenir des exemples de politiques IAM qu’un administrateur peut utiliser comme référence pour vous accorder des autorisations, consultez Exemples de politiques basées sur l’identité.

Table des matières

- Créer un groupe Auto Scaling à l'aide des paramètres d'une instance existante

Créer un groupe Auto Scaling à l'aide d'une configuration du lancement

> **Important**
>
> Nous fournissons des informations sur les configurations de lancement pour les clients qui n'ont pas encore migré des configurations de lancement vers les modèles de lancement. Pour plus d'informations sur la migration de vos groupes Auto Scaling vers les modèles de lancement, consultez la section Migrez vos groupes Auto Scaling pour lancer des modèles.

Lorsque vous créez un groupe Auto Scaling, vous devez indiquer les informations nécessaires pour configurer les instances Amazon EC2, les zones de disponibilité et les sous-réseaux VPC pour les instances, la capacité souhaitée et les limites de capacité minimale et maximale.

La procédure suivante montre comment créer un groupe Auto Scaling avec une configuration du lancement. Vous ne pouvez pas modifier une configuration du lancement après l'avoir créée, mais vous pouvez remplacer la configuration du lancement d'un groupe Auto Scaling. Pour plus d'informations, consultez Modifier la configuration du lancement pour un groupe Auto Scaling.

Prérequis

- Vous devez avoir créé une configuration du lancement. Pour plus d'informations, consultez Créez une configuration de lancement.

Pour créer un groupe Auto Scaling avec une configuration du lancement (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Dans la barre de navigation en haut de l'écran, choisissez la même Région AWS que celle que vous avez utilisée lors de la création de la configuration de lancement.
3. Choisissez Créer un groupe Auto Scaling.
4. Dans la page Choisir un modèle de lancement ou une configuration, dans Nom du groupe Auto Scaling, entrez un nom pour le groupe Auto Scaling.
5. Pour choisir une configuration du lancement, procédez comme suit :

a. Pour Launch template (Modèle de lancement), choisissez Switch to launch configuration (Basculer vers la configuration du lancement).

b. Pour Launch configuration (Configuration de lancement), choisissez une configuration du lancement existante.

c. Vérifiez que votre modèle de lancement prend en charge toutes les options que vous envisagez d'utiliser, puis choisissez Next (Suivant).

6. Sur la page Configurer les options de lancement de l'instance, sous Network (Réseau), pour VPC, choisissez un VPC. Le groupe Auto Scaling doit être créé dans le même VPC que le groupe de sécurité que vous avez spécifié dans votre configuration du lancement.

7. Pour Availability Zones and subnets (Zones de disponibilité et sous-réseaux), choisissez un ou plusieurs sous-réseaux dans le VPC spécifié. Utilisez les sous-réseaux dans plusieurs zones de disponibilité pour une haute disponibilité. Pour plus d'informations, consultez Considérations à prendre en compte lors du choix des sous-réseaux VPC.

8. Choisissez Suivant.

   Vous pouvez également accepter le reste des valeurs par défaut, puis choisir Skip to review (Passer à la révision).

9. (Facultatif) Sur la page Configure advanced options (Configurer les option avancées), configurez les options suivantes, puis choisissez Next (Suivant) :

   a. Sous Paramètres supplémentaires, Surveillance, indiquez si vous souhaitez activer la collecte des métriques de CloudWatch groupe. Ces métriques fournissent des mesures qui peuvent être des indicateurs d'un problème potentiel, comme le nombre d'instances en cours de résiliation ou le nombre d'instances en attente. Pour plus d'informations, consultez Surveillez CloudWatch les métriques de vos groupes et instances Auto Scaling.

   b. Pour Activer le préchauffage de l'instance par défaut, sélectionnez cette option et choisissez le temps de préchauffage de votre application. Si vous créez un groupe Auto Scaling doté d'une politique de dimensionnement, la fonctionnalité de préchauffage de l'instance par défaut améliore les CloudWatch métriques Amazon utilisées pour le dimensionnement dynamique. Pour plus d'informations, consultez Définir la préparation par défaut d'instance d'un groupe Auto Scaling.

10. (Facultatif) Sur la page Configure group size and scaling policies (Configurer les politiques de taille de groupe et de mise à l'échelle), configurez les options suivantes, puis choisissez Next (Suivant) :

a. Dans Taille du groupe, pour la Capacité souhaitée, entrez le nombre initial d'instances à lancer.

b. Dans la section Mise à l'échelle, sous Limites de mise à l'échelle, si votre nouvelle valeur pour la capacité souhaitée est supérieure à la capacité minimale souhaitée et à la capacité maximale souhaitée, la capacité maximale souhaitée est automatiquement augmentée à la nouvelle valeur de capacité souhaitée. Vous pouvez modifier ces limites si nécessaire. Pour plus d'informations, consultez Définissez des limites de mise à l'échelle pour votre groupe Auto Scaling.

c. Pour le dimensionnement automatique, indiquez si vous souhaitez créer une politique de dimensionnement de suivi des cibles. Vous pouvez également élaborer cette politique après avoir créé votre groupe Auto Scaling.

   Si vous choisissez la politique de dimensionnement de suivi des cibles, suivez les instructions dans Création d'une politique de suivi des cibles et d'échelonnement pour créer la politique.

d. Pour la politique de maintenance des instances, indiquez si vous souhaitez créer une politique de maintenance des instances. Vous pouvez également élaborer cette politique après avoir créé votre groupe Auto Scaling. Pour créer une politique, suivez les instructions fournies dans Définir une politique de maintenance des instances.

e. Sous Instance scale-in protection (Protection contre la diminution en charge des instances), choisissez si vous souhaitez activer la protection contre la diminution de la taille d'instance. Pour plus d'informations, consultez Utiliser la protection de la taille d'instance.

11. (Facultatif) Pour recevoir des notifications, dans Add notification (Ajouter une notification), configurez la notification, puis choisissez Next (Suivant). Pour plus d'informations, consultez Options de notification Amazon SNS pour Amazon EC2 Auto Scaling.
12. (Facultatif) Pour ajouter des balises, choisissez Add tag (Ajouter une balise), fournissez une clé de balise et une valeur pour chaque balise, puis choisissez Next (Suivant). Pour plus d'informations, consultez Baliser des groupes et des instances Auto Scaling.
13. Sur la page Review, sélectionnez Create Auto Scaling group (Créer un groupe Auto Scaling).

Pour créer un groupe Auto Scaling avec la ligne de commande

Vous pouvez utiliser l'une des commandes suivantes :

- create-auto-scaling-group (AWS CLI)

- Nouveautés-AS AutoScalingGroup ()AWS Tools for Windows PowerShell

Créer un groupe Auto Scaling à l'aide des paramètres d'une instance existante

> **Important**
>
> Nous fournissons des informations sur les configurations de lancement pour les clients qui n'ont pas encore migré des configurations de lancement vers les modèles de lancement. Pour plus d'informations sur la migration de vos groupes Auto Scaling vers les modèles de lancement, consultez la section Migrez vos groupes Auto Scaling pour lancer des modèles.

Si c'est la première fois que vous créez un groupe Auto Scaling, nous vous recommandons d'utiliser la console pour créer un modèle de lancement à partir d'une instance EC2 existante. Ensuite, utilisez le modèle de lancement pour créer un groupe Auto Scaling. Pour cette procédure, veuillez consulter Créer un groupe Auto Scaling avec l'Amazon EC2 Launch Wizard.

La procédure suivante montre comment créer un groupe Auto Scaling en spécifiant une instance existante à utiliser comme base pour le lancement d'autres instances. Plusieurs paramètres sont nécessaires pour créer une instance EC2, tels que l'ID Amazon Machine Image (AMI), le type d'instance, la paire de clés et le groupe de sécurité. Toutes ces informations sont également utilisées par Amazon EC2 Auto Scaling pour lancer des instances en votre nom lorsqu'il y a un besoin de mise à l'échelle. Ces informations sont stockées soit dans un modèle de lancement, soit dans une configuration du lancement.

Lorsque vous utilisez une instance existante, Amazon EC2 Auto Scaling crée un groupe Auto Scaling qui lance les instances en fonction d'une configuration du lancement qui est créée en même temps. La nouvelle configuration du lancement porte le même nom que le groupe Auto Scaling, et elle inclut certains détails de configuration de l'instance identifiée.

Les détails de configuration suivants sont copiés de l'instance identifiée dans la configuration du lancement :

- ID d'AMI
- Type d'instance
- Paire de clés

- Groupes de sécurité
- Type d'adresse IP (publique ou privée)
- Profil d'instance IAM, le cas échéant
- Surveillance (vrai ou faux)
- Optimisé pour EBS (vrai ou faux)
- Paramètre de location, en cas de lancement sur un VPC (partagé ou dédié)
- ID du noyau et ID du disque RAM, le cas échéant
- Données utilisateur, le cas échéant
- Prix Spot (maximum)

Le sous-réseau VPC et la zone de disponibilité sont copiés depuis l'instance identifiée vers la propre définition de ressource du groupe Auto Scaling.

Si l'instance identifiée se trouve dans un groupe de placement, le nouveau groupe Auto Scaling lance des instances dans le même groupe de placement que l'instance identifiée. Comme les paramètres de configuration du lancement ne permettent pas de spécifier un groupe de placement, le groupe de placement est copié dans l'attribut `PlacementGroup` du nouveau groupe Auto Scaling.

Les détails de configuration suivants ne sont pas copiés de votre instance identifiée :

- Stockage : les périphériques de bloc (volumes EBS et volumes de stockage d'instances) ne sont pas copiés à partir de l'instance identifiée. Au lieu de cela, le mappage de périphériques de stockage en mode bloc créé dans le cadre de la création de l'AMI détermine quels périphériques sont utilisés.
- Nombre d'interfaces réseau : les interfaces réseau ne sont pas copiées à partir de votre instance identifiée. Au lieu de cela, Amazon EC2 Auto Scaling utilise ses paramètres par défaut pour créer une interface réseau, qui est l'interface réseau principale (eth0).
- Options de métadonnées d'instance : les paramètres de métadonnées accessibles, de version des métadonnées et de limite de saut de réponse aux jetons ne sont pas copiés à partir de l'instance identifiée. Au lieu de cela, Amazon EC2 Auto Scaling utilise ses paramètres par défaut. Pour plus d'informations, consultez Configurer les options de métadonnées d'instance.
- Équilibreurs de charge : si l'instance identifiée est enregistrée avec un ou plusieurs équilibreurs de charge, les informations sur l'équilibreur de charge ne sont pas copiées sur l'équilibreur de charge ou l'attribut de groupe cible du nouveau groupe Auto Scaling.

- Identifications : si l'instance identifiée possède des identifications, ces dernières ne sont pas copiées dans l'attribut `Tags` du nouveau groupe Auto Scaling.

Prérequis

L'instance EC2 doit répondre aux critères suivants :

- L'instance ne fait pas partie d'un autre groupe Auto Scaling.
- L'instance a pour statut `running`.
- L'AMI qui a été utilisée pour lancer l'instance doit toujours exister.

Créer un groupe Auto Scaling à partir d'une instance EC2 (console)

Pour créer un groupe Auto Scaling à partir d'une instance EC2

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Dans le panneau de navigation, sous Instances, choisissez Instances, puis sélectionnez une instance.
3. Choisissez Actions, Instance Settings (Paramètres de l'instance), puis Attach to Auto Scaling Group (Attacher à un groupe Auto Scaling).
4. Sur la page Attach to Auto Scaling Group (Attacher à un groupe Auto Scaling), sélectionnez New Auto Scaling group (Nouveau groupe Auto Scaling), saisissez un nom de groupe, puis choisissez Attach (Attacher).

   Une fois attachée, une instance est considérée comme faisant partie du groupe Auto Scaling. Le nouveau groupe Auto Scaling est créé avec une nouvelle configuration de lancement et le même nom spécifié pour le groupe Auto Scaling. Le groupe Auto Scaling a une capacité souhaitée et une taille maximale de 1.
5. (Facultatif) Pour modifier les paramètres du groupe Auto Scaling, dans le panneau de navigation, sous Auto Scaling, choisissez Auto Scaling Groups (Groupes Auto Scaling). Cochez la case en regard du nouveau groupe , cliquez sur le bouton Edit (Modifier) situé au-dessus de la liste des groupes, modifiez les paramètres si nécessaire, puis choisissez Update (Mettre à jour).

Créer un groupe Auto Scaling à partir d'une instance EC2 (AWS CLI)

La procédure suivante expliquer comment utiliser une commande CLI pour créer un groupe Auto Scaling à partir d'une instance EC2.

Cette procédure n'ajoute pas l'instance au groupe Auto Scaling. Pour que l'instance soit attachée, vous devez exécuter attachez des instances une fois votre groupe Auto Scaling créé.

Avant de commencer, recherchez l'ID de l'instance EC2 avec la console Amazon EC2 ou la commande describe-instances.

Pour utiliser l'instance actuelle comme modèle

- Utilisez la commande create-auto-scaling-group suivante pour créer un groupe Auto Scaling `my-asg-from-instance` à partir de l'instance EC2 `i-0e69cc3f05f825f4f`.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg-from-instance \
  --instance-id i-0e69cc3f05f825f4f --min-size 1 --max-size 2 --desired-capacity 2
```

Pour vérifier que votre groupe Auto Scaling possède des instances lancées

- Utilisez la commande describe-auto-scaling-groups suivante pour vérifier que le groupe Auto Scaling a bien été créé.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg-from-instance
```

L'exemple de réponse suivant montre que la capacité souhaitée du groupe est 2, que le groupe possède 2 instances en cours d'exécution, et que la configuration du lancement est également nommée `my-asg-from-instance`.

```
{
  "AutoScalingGroups":[
    {
      "AutoScalingGroupName":"my-asg-from-instance",
      "AutoScalingGroupARN":"arn",
      "LaunchConfigurationName":"my-asg-from-instance",
      "MinSize":1,
      "MaxSize":2,
```

```
      "DesiredCapacity":2,
      "DefaultCooldown":300,
      "AvailabilityZones":[
        "us-west-2a"
      ],
      "LoadBalancerNames":[],
      "TargetGroupARNs":[],
      "HealthCheckType":"EC2",
      "HealthCheckGracePeriod":0,
      "Instances":[
        {
          "InstanceId":"i-06905f55584de02da",
          "InstanceType":"t2.micro",
          "AvailabilityZone":"us-west-2a",
          "LifecycleState":"InService",
          "HealthStatus":"Healthy",
          "LaunchConfigurationName":"my-asg-from-instance",
          "ProtectedFromScaleIn":false
        },
        {
          "InstanceId":"i-087b42219468eacde",
          "InstanceType":"t2.micro",
          "AvailabilityZone":"us-west-2a",
          "LifecycleState":"InService",
          "HealthStatus":"Healthy",
          "LaunchConfigurationName":"my-asg-from-instance",
          "ProtectedFromScaleIn":false
        }
      ],
      "CreatedTime":"2020-10-28T02:39:22.152Z",
      "SuspendedProcesses":[ ],
      "VPCZoneIdentifier":"subnet-6bea5f06",
      "EnabledMetrics":[ ],
      "Tags":[ ],
      "TerminationPolicies":[
        "Default"
      ],
      "NewInstancesProtectedFromScaleIn":false,
      "ServiceLinkedRoleARN":"arn",
      "TrafficSources":[]
    }
  ]
}
```

Pour afficher la configuration du lancement

- Utilisez la commande describe-launch-configurations suivante pour afficher les détails de la configuration du lancement.

```
aws autoscaling describe-launch-configurations --launch-configuration-names my-asg-from-instance
```

Voici un exemple de sortie :

```
{
  "LaunchConfigurations":[
    {
      "LaunchConfigurationName":"my-asg-from-instance",
      "LaunchConfigurationARN":"arn",
      "ImageId":"ami-0528a5175983e7f28",
      "KeyName":"my-key-pair-uswest2",
      "SecurityGroups":[
        "sg-05eaec502fcdadc2e"
      ],
      "ClassicLinkVPCSecurityGroups":[ ],
      "UserData":"",
      "InstanceType":"t2.micro",
      "KernelId":"",
      "RamdiskId":"",
      "BlockDeviceMappings":[ ],
      "InstanceMonitoring":{
        "Enabled":true
      },
      "CreatedTime":"2020-10-28T02:39:22.321Z",
      "EbsOptimized":false,
      "AssociatePublicIpAddress":true
    }
  ]
}
```

Pour résilier l'instance

- Vous pouvez résilier l'instance si vous n'en n'avez plus besoin. La commande terminate-instances suivant résilie l'instance`i-0e69cc3f05f825f4f`.

```
aws ec2 terminate-instances --instance-ids i-0e69cc3f05f825f4f
```

Après avoir résilié une instance Amazon EC2, vous ne pouvez pas la redémarrer. Une fois résiliée, ses données sont perdues et le volume ne peut être attaché à aucune instance. Pour en savoir plus sur la résiliation d'instances, consultez la section Résiliation d'une instance dans le guide de l'utilisateur Amazon EC2.

Mettre à jour un groupe Auto Scaling

Vous pouvez mettre à jour la plupart des informations de votre groupe Auto Scaling. Vous ne pouvez pas mettre à jour le nom d'un groupe Auto Scaling ni le modifier Région AWS.

Pour mettre à jour un groupe Auto Scaling (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Choisissez votre groupe Auto Scaling pour afficher les informations le concernant, avec les onglets Détails, Activité, Dimensionnement automatique, Gestion des instances, Surveillance et Actualisation des instances.
3. Choisissez les onglets correspondant aux zones de configuration qui vous intéressent et mettez à jour les paramètres selon vos besoins. Pour chaque paramètre que vous modifiez, sélectionnez Mettre à jour pour enregistrer les modifications apportées à la configuration du groupe Auto Scaling.

   - Onglet Détails

     Voici les paramètres généraux de votre groupe Auto Scaling. Vous pouvez les modifier et les gérer de la même manière que lors de la création d'un groupe Auto Scaling.

     La section Configurations avancées contient certaines options qui ne sont pas disponibles lors de la création du groupe, telles que les politiques de résiliation, la stabilisation, les processus d'interruption et la durée de vie maximale de l'instance. Vous pouvez également afficher, mais pas modifier, le groupe de placement et le rôle lié à un service du groupe Auto Scaling.

     Si le groupe est associé à des ressources Elastic Load Balancing, consultez Ajouter et supprimer des zones de disponibilité avant de modifier les zones de disponibilité. Certaines restrictions relatives à l'équilibreur de charge peuvent vous empêcher d'appliquer les

modifications apportées aux zones de disponibilité de votre groupe aux zones de disponibilité de votre équilibreur de charge.

- Onglet Activité
  - Notifications d'activité — Notifications Amazon SNS
- Onglet Mise à l'échelle automatique
  - Politiques de dimensionnement dynamiques — Politiques de dimensionnement dynamiques
  - Politiques de dimensionnement prédictif — Politiques de dimensionnement prédictif
  - Actions planifiées — Actions planifiées
- Onglet Gestion des instances
  - Crochets Lifecycle — Crochets Lifecycle
  - Piscine chaude — Piscines chaudes
- Onglet Surveillance
  - Il n'y a qu'une seule option dans cet onglet, qui vous permet d'activer ou de désactiver la collecte de métriques de CloudWatch groupe.

Pour mettre à jour un groupe Auto Scaling avec la ligne de commande

Vous pouvez utiliser l'une des commandes suivantes :

- update-auto-scaling-group (AWS CLI)
- Mettre à jour en tant que AutoScaling groupe ()AWS Tools for Windows PowerShell

Mise à jour des instances Auto Scaling

Si vous associez un nouveau modèle de lancement ou une configuration de lancement à un groupe Auto Scaling, toutes les nouvelles instances recevront la configuration mise à jour. Les instances existantes continuent à s'exécuter avec leur configuration initiale. Pour appliquer vos modifications aux instances existantes, vous disposez des options suivantes :

- Lancer une actualisation d'instance pour remplacer les anciennes instances. Pour plus d'informations, consultez Utiliser une actualisation d'instance pour mettre à jour les instances d'un groupe Auto Scaling.
- Attendez que les activités de mise à l'échelle remplacent progressivement des instances plus anciennes par des instances plus récentes, en fonction de vos stratégies de résiliation.

- Résiliez-les manuellement afin qu'elles soient remplacées par votre groupe Auto Scaling.

> **Note**
>
> Vous pouvez modifier les attributs d'instance suivants en les indiquant dans le modèle de lancement ou dans la configuration de lancement :
>
> - Amazon Machine Image (AMI)
> - périphériques de stockage en mode bloc
> - paire de clés
> - type d'instance
> - groupes de sécurité
> - données utilisateur
> - surveillance
> - Profil d'instance IAM
> - location de placement
> - kernel
> - ramdisk
> - si l'instance possède une adresse IP publique

Baliser des groupes et des instances Auto Scaling

Une balise est une étiquette d'attribut personnalisée que vous attribuez ou attribuez à une AWS ressource. AWS Chaque balise se compose de deux parties :

- Une clé d'identification (par exemple, `costcenter`, `environment` ou `project`)
- Un champ facultatif appelé valeur d'identification (par exemple, `111122223333` ou `production`)

Les balises vous permettent d'effectuer les actions suivantes :

- Suivez vos AWS coûts. Vous activez ces balises sur le AWS Billing and Cost Management tableau de bord. AWS utilise les balises pour classer vos coûts et vous fournir un rapport mensuel de

répartition des coûts. Pour plus d'informations, veuillez consulter Utilisation des étiquettes de répartition des coûts dans le AWS Billing Guide de l'utilisateur.

- Contrôlez l'accès aux groupes Auto Scaling basé sur des balises. Vous pouvez utiliser des conditions dans vos politiques IAM pour contrôler l'accès aux groupes Auto Scaling en fonction des balises de ce groupe Auto Scaling. Pour plus d'informations, consultez Balises pour la sécurité.
- Filtrez et recherchez des groupes Auto Scaling en fonction des balises que vous insérez. Pour plus d'informations, consultez Utilisation d'identifications pour filtrer les groupes Auto Scaling.
- Identifiez et organisez vos AWS ressources. Beaucoup Services AWS prennent en charge le balisage. Vous pouvez donc attribuer le même tag aux ressources provenant de différents services pour indiquer que les ressources sont liées.

Vous pouvez baliser les groupes Auto Scaling nouveaux ou existants. Vous pouvez également propager les balises d'un groupe Auto Scaling aux instances EC2 qu'il lance.

Les balises ne sont pas propagées aux volumes Amazon EBS. Pour ajouter des balises aux volumes Amazon EBS, spécifiez les balises dans un modèle de lancement. Pour plus d'informations, consultez Créer un modèle de lancement pour un groupe Auto Scaling.

Vous pouvez créer et gérer des balises via AWS Management Console le ou AWS CLI les SDK.

Table des matières

Limites d'utilisation et de dénomination des balises

Les restrictions de base suivantes s'appliquent aux balises :

- Le nombre maximum d'identifications par ressource est de 50.

- Le nombre maximum de balises susceptibles d'être ajoutées ou supprimées avec un seul appel est de 25.
- La longueur de clé maximale est de 128 caractères Unicode.
- La longueur de valeur maximale est de 256 caractères Unicode.
- Les clés et valeurs d'étiquette sont sensibles à la casse. La bonne pratique consiste à choisir une politique pour mettre des balises en majuscule et mettre en œuvre cette politique de manière cohérente sur tous les types de ressources.
- N'utilisez pas le `aws:` préfixe dans les noms ou les valeurs de vos balises, car il est réservé à AWS l'usage. Vous ne pouvez pas modifier ou supprimer les noms ou valeurs des balises avec ce préfixe, et elles ne sont pas comptées dans vos balises par quota de groupe.

Cycle de vie de balisage d'instance EC2

Si vous avez choisi de propager des balises vers vos instances EC2, celles-ci sont gérées comme suit :

- Lorsqu'un groupe Auto Scaling lance des instances, il ajoute les balises aux instances lors de la création de ressources plutôt qu'après la création de la ressource.
- Le groupe Auto Scaling ajoute automatiquement une identification aux instances avec une clé `aws:autoscaling:groupName` et une valeur du nom du groupe Auto Scaling.
- Si vous spécifiez des balises d'instance dans votre modèle de lancement et que vous avez choisi de propager les balises de votre groupe à ses instances, toutes les balises sont fusionnées. Si la même clé d'identification est spécifiée pour une identification dans votre modèle de lancement et une identification dans votre groupe Auto Scaling, la valeur de l'identification du groupe est prioritaire.
- Lorsque vous attachez des instances existantes, le groupe Auto Scaling ajoute les balises aux instances en remplaçant les anciennes balises par la même clé de balise. Il ajoute également une identification avec une clé de `aws:autoscaling:groupName` et une valeur du nom du groupe Auto Scaling.
- Quand vous détachez une instance d'un groupe Auto Scaling, celui-ci ne supprime que la balise `aws:autoscaling:groupName`.

Baliser vos groupes Auto Scaling

Lorsque vous ajoutez une balise au groupe Auto Scaling, vous pouvez spécifier si elle doit être ajoutée aux instances lancées dans le groupe Auto Scaling. Si vous modifiez une balise, la version mise à jour de la balise est ajoutée aux instances lancées dans le groupe Auto Scaling après le changement. Si vous créez ou modifiez une balise pour un groupe Auto Scaling, ces changements ne sont pas appliqués aux instances déjà en cours d'exécution dans le groupe Auto Scaling.

Table des matières

Ajouter ou modifier des balises (console)

Pour étiqueter un groupe Auto Scaling lors de la création

Lorsque vous utilisez la console Amazon EC2 pour créer un groupe Auto Scaling, vous pouvez spécifier les clés et les valeurs des balises sur la page Add tags (Ajouter des balises) de l'assistant Create Auto Scaling group (Créer un groupe Auto Scaling). Pour propager une balise aux instances lancées dans le groupe Auto Scaling, assurez-vous que vous conservez l'option Tag new instances (Baliser les nouvelles instances) pour cette balise sélectionnée. Sinon, vous pouvez la désactiver.

Pour ajouter ou modifier des balises pour un groupe Auto Scaling existant

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page Auto Scaling groups (Groupes Auto Scaling).
3. Sous l'onglet Details (Détails) choisissez Tags (Balises), Edit (Modifier).
4. Pour modifier des identifications existantes, modifiez Key (Clé) et Value (Valeur).
5. Pour ajouter une nouvelle identification, choisissez Add tag (Ajouter une identification) et modifiez Key (Clé) et Value (Valeur). Vous pouvez continuer de sélectionner l'option Tag New Instances (Baliser les nouvelles instances) pour ajouter automatiquement la balise aux instances lancées dans le groupe Auto Scaling, sinon annulez sa sélection.
6. Lorsque vous avez fini d'ajouter des balises, choisissez Update (Mettre à jour).

Ajouter ou modifier des balises (AWS CLI)

Les exemples suivants montrent comment utiliser le AWS CLI pour ajouter des balises lorsque vous créez des groupes Auto Scaling et pour ajouter ou modifier des balises pour des groupes Auto Scaling existants.

Pour baliser un groupe Auto Scaling lors de la création

Utilisez la commande create-auto-scaling-group pour créer un groupe Auto Scaling et ajouter une balise, par exemple,**environment=production**, dans le groupe Auto Scaling. La balise est également ajoutée aux instances lancées dans le groupe Auto Scaling.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-configuration-name my-launch-config --min-size 1 --max-size 3 \
  --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
  --tags Key=environment,Value=production,PropagateAtLaunch=true
```

Pour créer ou modifier des balises pour un groupe Auto Scaling existant

Utilisez la commande create-or-update-tags pour créer ou modifier une balise. Par exemple, la commande suivante ajoute les balises **Name=my-asg** et **costcenter=cc123**. Les balises sont également ajoutées aux instances lancées dans le groupe Auto Scaling après cette modification. Si une balise avec cette clé existe déjà, la balise existante est remplacée. La console Amazon EC2 associe le nom d'affichage de chaque instance avec le nom spécifié pour la clé Name (sensible à la casse).

```
aws autoscaling create-or-update-tags \
  --tags ResourceId=my-asg,ResourceType=auto-scaling-group,Key=Name,Value=my-
asg,PropagateAtLaunch=true \
  ResourceId=my-asg,ResourceType=auto-scaling-
group,Key=costcenter,Value=cc123,PropagateAtLaunch=true
```

Décrire les balises d'un groupe Auto Scaling (AWS CLI)

Si vous souhaitez afficher les balises qui sont appliquées à un groupe Auto Scaling spécifique, vous pouvez utiliser l'une des commandes suivantes :

- describe-tags — Vous indiquez le nom de votre groupe Auto Scaling pour afficher la liste des balises du groupe spécifié.

```
aws autoscaling describe-tags --filters Name=auto-scaling-group,Values=my-asg
```

Voici un exemple de réponse.

```
{
    "Tags": [
        {
            "ResourceType": "auto-scaling-group",
            "ResourceId": "my-asg",
            "PropagateAtLaunch": true,
            "Value": "production",
            "Key": "environment"
        }
    ]
}
```

- describe-auto-scaling-groups — Vous indiquez le nom de votre groupe Auto Scaling pour afficher les attributs du groupe spécifié, y compris les balises.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Voici un exemple de réponse.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            "LaunchTemplate": {
                "LaunchTemplateId": "lt-0b97f1e282EXAMPLE",
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "MinSize": 1,
            "MaxSize": 5,
            "DesiredCapacity": 1,
            ...
            "Tags": [
                {
```

```
                    "ResourceType": "auto-scaling-group",
                    "ResourceId": "my-asg",
                    "PropagateAtLaunch": true,
                    "Value": "production",
                    "Key": "environment"
                }
            ],
            ...
        }
    ]
}
```

Supprimer des balises

Vous pouvez supprimer une balise associée au groupe Auto Scaling à tout moment.

Table des matières

Supprimer des balises (console)

Pour supprimer une balise

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Activez la case à cocher en regard d'un groupe existant.

   Un volet fractionné s'ouvre en bas de la page Auto Scaling groups (Groupes Auto Scaling).
3. Sous l'onglet Details (Détails) choisissez Tags (Balises), Edit (Modifier).
4. Choisissez Remove (Supprimer) en regard de la balise.
5. Choisissez Mettre à jour.

Supprimer des balises (AWS CLI)

Utilisez la commande delete-tags pour supprimer une balise. Par exemple, la commande suivante supprime une balise avec une clé **`environment`**.

```
aws autoscaling delete-tags --tags "ResourceId=my-asg,ResourceType=auto-scaling-
group,Key=environment"
```

Vous devez préciser la clé de balise, mais pas la valeur. Si vous spécifiez une valeur et qu'elle est incorrecte, la balise n'est pas supprimée.

Balises pour la sécurité

Utilisez des balises pour vérifier que le demandeur (tel qu'un utilisateur ou un rôle IAM) dispose des autorisations de créer, modifier ou supprimer des groupes Auto Scaling spécifiques. Fournissez des informations de balise dans l'élément de condition d'une politique IAM à l'aide des clés de condition suivantes :

- Utilisez `autoscaling:ResourceTag/`*`tag-key`*`: `*`tag-value`* pour accorder (ou refuser) aux utilisateurs des actions sur des groupes Auto Scaling avec des balises spécifiques.
- Utilisez `aws:RequestTag/`*`tag-key`*`: `*`tag-value`* pour exiger qu'une balise spécifique soit présente (ou non) dans une demande.
- Utilisez `aws:TagKeys [`*`tag-key`*`, ...]` pour exiger que des clés de balise spécifiques soient présentes (ou non) dans une demande.

Par exemple, vous pouvez refuser l'accès à tous les groupes Auto Scaling qui incluent une balise avec la clé **`environment`** et la valeur **`production`**, comme illustré dans l'exemple suivant.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Deny",
            "Action": [
                "autoscaling:CreateAutoScalingGroup",
                "autoscaling:UpdateAutoScalingGroup",
                "autoscaling:DeleteAutoScalingGroup"
            ],
            "Resource": "*",
            "Condition": {
                "StringEquals": {"autoscaling:ResourceTag/environment": "production"}
            }
        }
    ]
```

```
}
```

Pour plus d'informations sur l'utilisation des clés de condition afin de contrôler l'accès aux groupes Auto Scaling, consultez Fonctionnement d'Amazon EC2 Auto Scaling avec IAM.

Contrôler l'accès aux balises

Utilisez des balises pour vérifier que le demandeur (tel qu'un utilisateur ou un rôle IAM) dispose des autorisations d'ajouter, modifier ou supprimer des balises pour des groupes Auto Scaling.

L'exemple de politique IAM suivant donne l'autorisation principale de supprimer uniquement la balise avec la clé **`temporary`** des groupes Auto Scaling.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "autoscaling:DeleteTags",
            "Resource": "*",
            "Condition": {
                "ForAllValues:StringEquals": { "aws:TagKeys": ["temporary"] }
            }
        }
    ]
}
```

Pour plus d'exemples de politiques IAM qui appliquent des contraintes sur les balises spécifiées pour les groupes Auto Scaling, consultez Contrôler les clés de balise et les valeurs de balise pouvant être utilisées.

> **Note**
>
> Même si vous disposez d'une politique qui empêche vos utilisateurs d'exécuter une opération d'étiquetage (ou d'annulation de désétiquetage) sur un groupe Auto Scaling, cela ne les empêche pas de modifier manuellement les identifications sur les instances après les avoir lancées. Pour des exemples qui contrôlent l'accès aux balises sur les instances EC2, consultez Exemple : ressources de balisage dans le guide de l'utilisateur Amazon EC2.

Utilisation d'identifications pour filtrer les groupes Auto Scaling

Les exemples suivants vous montrent comment utiliser des filtres avec la commande describe-auto-scaling-groups pour décrire des groupes Auto Scaling avec des identifications spécifiques. Le filtrage par balises est limité au SDK AWS CLI ou à un SDK et n'est pas disponible depuis la console.

Considérations relatives au filtrage

- Vous pouvez spécifier plusieurs filtres et plusieurs valeurs de filtre dans une seule requête.
- Vous ne pouvez pas utiliser des caractères génériques avec les valeurs de filtre.
- Les valeurs de filtre sont sensibles à la casse.

Exemple : décrire les groupes Auto Scaling avec une paire clé-valeur d'identification spécifique

La commande suivante montre comment filtrer les résultats pour n'afficher que les groupes Auto Scaling dont la paire clé/valeur d'identification est **`environment=production`**.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-key,Values=environment Name=tag-value,Values=production
```

Voici un exemple de réponse.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            "LaunchTemplate": {
                "LaunchTemplateId": "lt-0b97f1e282EXAMPLE",
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "MinSize": 1,
            "MaxSize": 5,
            "DesiredCapacity": 1,
            ...
            "Tags": [
                {
                    "ResourceType": "auto-scaling-group",
                    "ResourceId": "my-asg",
                    "PropagateAtLaunch": true,
```

```
                    "Value": "production",
                    "Key": "environment"
                }
            ],
            ...
        },

    ... additional groups ...

    ]
}
```

Vous pouvez également spécifier des identifiants à l'aide d'un filtre `tag:<key>`. Par exemple, la commande suivante montre comment filtrer les résultats pour n'afficher que les groupes Auto Scaling avec une paire clé et valeur d'identification de **`environment=production`**. Ce filtre est formaté comme suit : `Name=tag:<key>,Values=<value>`, avec **`<key>`** et **`<value>`** représentant une paire clé et valeur d'identification.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag:environment,Values=production
```

Vous pouvez également filtrer la AWS CLI sortie à l'aide de l'`--query`option. L'exemple suivant montre comment limiter la AWS CLI sortie de la commande précédente au nom du groupe, à la taille minimale, à la taille maximale et aux attributs de capacité souhaités uniquement.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag:environment,Values=production \
  --query "AutoScalingGroups[].{AutoScalingGroupName: AutoScalingGroupName, MinSize:
 MinSize, MaxSize: MaxSize, DesiredCapacity: DesiredCapacity}"
```

Voici un exemple de réponse.

```
[
    {
        "AutoScalingGroupName": "my-asg",
        "MinSize": 0,
        "MaxSize": 10,
        "DesiredCapacity": 1
    },

    ... additional groups ...
```

```
]
```

Pour plus d'informations sur le filtrage, consultez la section Filtrage AWS CLI de la sortie dans le guide de AWS Command Line Interface l'utilisateur.

Exemple : décrire les groupes Auto Scaling dont les identifications correspondent à la clé d'identification spécifiée

La commande suivante montre comment filtrer les résultats pour afficher uniquement les groupes Auto Scaling avec l'identification **`environment`**, quelle que soit la valeur de l'identification.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-key,Values=environment
```

Exemple : décrire les groupes d'Auto Scaling dont les identifications correspondent à l'ensemble des clés d'identification spécifiées

La commande suivante montre comment filtrer les résultats pour afficher uniquement les groupes Auto Scaling avec des identifications pour **`environment`** et **`project`**, quelle que soit la valeur de l'identification.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-key,Values=environment Name=tag-key,Values=project
```

Exemple : décrire les groupes Auto Scaling dont les identifications correspondent à au moins une des clés d'identification spécifiées

La commande suivante montre comment filtrer les résultats pour afficher uniquement les groupes Auto Scaling avec des identifications pour **`environment`** ou **`project`**, quelle que soit la valeur de l'identification.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-key,Values=environment,project
```

Exemple : décrire les groupes Auto Scaling avec la valeur d'identification spécifiée

La commande suivante montre comment filtrer les résultats pour afficher uniquement les groupes Auto Scaling dont la valeur de l'identification est **`production`**, quelle que soit la clé de l'identification.

```
aws autoscaling describe-auto-scaling-groups \
```

```
  --filters Name=tag-value,Values=production
```

Exemple : décrire les groupes Auto Scaling avec l'ensemble des valeurs d'identification spécifiées

La commande suivante montre comment filtrer les résultats pour afficher uniquement les groupes Auto Scaling dont l'identification est **`production`** et **`development`**, quelle que soit la clé d'identification.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-value,Values=production Name=tag-value,Values=development
```

Exemple : décrire les groupes Auto Scaling dont les identifications correspondent à au moins une des valeurs d'identification spécifiées

La commande suivante montre comment filtrer les résultats pour afficher uniquement les groupes Auto Scaling dont la valeur de l'identification est **`production`** ou **`development`**, quelle que soit la clé de l'identification.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-value,Values=production,development
```

Exemple : décrire les groupes d'Auto Scaling dont les identifications correspondent à plusieurs clés et valeurs d'identification

Vous pouvez également combiner des filtres pour créer une logique AND et OR personnalisée pour effectuer un filtrage plus complexe.

La commande suivante montre comment filtrer les résultats pour afficher uniquement les groupes Auto Scaling avec un ensemble spécifique d'identifications. Une clé d'identification est **`environment`** AND la valeur de l'identification est (**`production`** OR **`development`**) AND l'autre clé d'identification est **`costcenter`** AND la valeur de l'identification est **`cc123`**.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag:environment,Values=production,development
 Name=tag:costcenter,Values=cc123
```

Politiques de maintenance des instances

Vous pouvez configurer une politique de maintenance des instances pour votre groupe Auto Scaling afin de répondre à des exigences de capacité spécifiques lors d'événements entraînant le

remplacement d'instances, tels que l'actualisation d'une instance ou le processus de surveillance de l'état.

Supposons que vous possédez un groupe Auto Scaling qui contient un petit nombre d'instances. Vous souhaitez éviter les perturbations potentielles liées à la résiliation puis au remplacement d'une instance lorsque les surveillances de l'état indiquent qu'une instance est défectueuse. Avec une politique de maintenance des instances, vous pouvez vous assurer qu'Amazon EC2 Auto Scaling lance d'abord une nouvelle instance, puis attend qu'elle soit complètement prête avant de résilier l'instance défectueuse.

Une politique de maintenance des instances vous aide également à minimiser les perturbations potentielles dans les cas où plusieurs instances sont remplacées en même temps. Vous définissez les paramètres de pourcentage minimal et maximal d'intégrité de la politique, et votre groupe Auto Scaling ne peut augmenter ou diminuer la capacité dans cette plage minimale-maximale que lors du remplacement d'instances. Une plage étendue augmente le nombre d'instances qui peuvent être remplacées en même temps.

Table des matières

Présentation des politiques de maintenance des instances

Cette rubrique fournit une vue d'ensemble des options disponibles et décrit les éléments à prendre en compte lorsque vous créez une politique de maintenance d'instance.

Table des matières

Présentation

Lorsque vous créez une politique de maintenance des instances pour votre groupe Auto Scaling, la politique affecte les événements Amazon EC2 Auto Scaling qui entraînent le remplacement des instances. Cela se traduit par des comportements de remplacement plus cohérents au sein du même groupe Auto Scaling. Cela vous permet également d'optimiser la disponibilité ou le coût de votre groupe en fonction de vos besoins.

Les options de configuration suivantes sont disponibles dans la console :

- Lancer avant toute résiliation : une nouvelle instance doit d'abord être mise en service avant qu'une instance existante puisse être résiliée. Cette approche est un bon choix pour les applications qui privilégient la disponibilité plutôt que les économies de coûts.
- Résilier et lancer : les nouvelles instances sont mises en service en même temps que les instances existantes sont résiliées. Cette approche est un bon choix pour les applications qui privilégient les économies de coûts plutôt que la disponibilité. C'est également un bon choix pour les applications qui ne doivent pas libérer plus de capacité que ce qui est actuellement disponible, même lors du remplacement d'instances.
- Politique personnalisée : cette option vous permet de configurer votre politique avec une plage minimale et maximale personnalisée pour la quantité de capacité que vous souhaitez mettre à disposition lors du remplacement d'instances. Cette approche peut vous aider à trouver le juste équilibre entre le coût et la disponibilité.

Par défaut, un groupe Auto Scaling n'a pas de politique de maintenance des instances, ce qui l'oblige à répondre aux événements de maintenance des instances avec les comportements par défaut. Les comportements par défaut sont décrits dans le tableau suivant.

Comportements par défaut des événements de maintenance des instances

Événement	Description	Comportement par défaut
Échec de la surveillance de l'état	Cela se produit automatiq uement lorsque les instances échouent à leurs surveilla nces de l'état. Amazon EC2 Auto Scaling remplace les instances qui échouent à leurs surveillances de l'état. Pour	Résilier et lancer.

Événement	Description	Comportement par défaut
	comprendre les causes des échecs liés aux surveillances de l'état, consultez Surveillance de l'état des instances dans un groupe Auto Scaling.	
Actualisation d'instance	Se produit lorsque vous actualisez une instance. En fonction de votre configuration, une actualisation d'instance remplace les instances une par une, plusieurs à la fois ou toutes à la fois. Pour plus d'informations, consultez Utiliser une actualisation d'instance pour mettre à jour les instances d'un groupe Auto Scaling.	Résilier et lancer.
Durée de vie maximale de l'instance	Cela se produit automatiquement lorsque les instances atteignent la durée de vie maximale que vous indiquez pour votre groupe Auto Scaling. Amazon EC2 Auto Scaling remplace les instances qui atteignent leur durée de vie maximale. Pour plus d'informations, consultez Remplacer des instances Auto Scaling en fonction de la durée de vie maximale de l'instance.	Résilier et lancer.

Événement	Description	Comportement par défaut
Rééquilibrage	Cela se produit automatiquement si des changements sous-jacents entraînent un déséquilibre du groupe. Amazon EC2 Auto Scaling rééquilibre le groupe dans les situations suivantes : • Une zone de disponibilité dont la capacité était auparavant insuffisante se rétablit, ou vous ajoutez ou supprimez une zone de disponibilité du groupe. Lorsque cela se produit, votre groupe Auto Scaling essaye de s'équilibrer uniformément entre les zones de disponibilité. Pour plus d'informations, consultez Activités de rééquilibrage. • Vous activez le rééquilibrage de la capacité sur votre groupe Auto Scaling, qui essaye de lancer de nouvelles instances Spot avant que les instances existantes ne soient interrompues à mesure que la disponibilité des instances Spot change. Pour plus d'informations, consultez Utiliser le rééquilibrage de la capacité pour gérer les	Lancer avant toute résiliation. Amazon EC2 Auto Scaling peut dépasser les limites de taille de votre groupe jusqu'à 10 % de sa capacité maximale. Toutefois, si vous utilisez le rééquilibrage de la capacité, il ne peut dépasser ces limites que de 10 % de la capacité souhaitée.

Événement	Description	Comportement par défaut
	interruptions Spot Amazon EC2. • Vous mettez à jour votre groupe Auto Scaling, qui remplace progressivement les instances en fonction des nouvelles options d'achat que vous avez choisies lors de la mise à jour d'une politique d'instanc es mixtes. Pour plus d'informations, consultez Mettre à jour un groupe Auto Scaling.	

Amazon EC2 Auto Scaling continuera à être résilié et lancé par défaut dans les situations suivantes. Par conséquent, lorsque l'une de ces situations se produit, la capacité de votre groupe peut être inférieure au seuil inférieur de votre politique de maintenance des instances.

- Lorsqu'une instance est résiliée de façon inattendue, par exemple en raison d'une action humaine. Amazon EC2 Auto Scaling remplace les instances qui ne fonctionnent plus. Pour plus d'informations, consultez Surveillance de l'état Amazon EC2.
- Lorsqu'Amazon EC2 redémarre, arrête ou retire une instance dans le cadre d'un événement planifié avant qu'Amazon EC2 Auto Scaling ne puisse lancer l'instance de remplacement. Pour plus d'informations sur ces événements, consultez la section Événements planifiés pour vos instances dans le guide de l'utilisateur Amazon EC2.
- Lorsque le service Amazon EC2 Spot initie une interruption d'instance Spot et qu'une instance Spot est ensuite résiliée de force.

Avec les instances Spot, si vous avez activé le rééquilibrage de la capacité sur votre groupe Auto Scaling, l'instance possède peut-être déjà une instance en attente provenant d'un autre groupe Spot que nous avons lancé avant de démarrer l'interruption Spot. Pour de plus amples informations sur le

fonctionnement du rééquilibrage de la capacité, consultez Utiliser le rééquilibrage de la capacité pour gérer les interruptions Spot Amazon EC2.

Cependant, étant donné que la disponibilité des instances Spot n'est pas garantie et qu'elles peuvent être résiliées moyennant un préavis d'interruption de deux minutes, le seuil inférieur de votre politique de maintenance des instances peut être dépassé si les instances sont interrompues avant le lancement de vos nouvelles instances.

Concepts de base

Avant de commencer, familiarisez-vous avec les principaux concepts et termes suivants :

Capacité souhaitée

La Capacité souhaitée correspond à la capacité initiale du groupe Auto Scaling à sa création. Il s'agit également de la capacité que le groupe essaye de maintenir lorsqu'aucune condition de dimensionnement n'est attachée au groupe.

Politique de maintenance des instances

Une politique de maintenance des instances contrôle si une instance est d'abord mise en service avant qu'une instance existante ne soit résiliée en cas d'événements de maintenance de l'instance. Elle détermine également jusqu'où votre groupe Auto Scaling peut aller en dessous et au-dessus de la capacité souhaitée pour remplacer plusieurs instances en même temps.

Pourcentage maximal d'intégrité

Le pourcentage maximal d'intégrité est le pourcentage de la capacité souhaitée que votre groupe Auto Scaling peut atteindre lors du remplacement d'instances. Il représente le pourcentage maximal du groupe qui peut être en service et en bon état, ou en attente, pour assurer votre charge de travail. Dans la console, vous pouvez définir le pourcentage maximal d'intégrité lorsque vous utilisez l'option Lancer avant toute résiliation ou l'option Politique personnalisée. Les valeurs valides sont comprises entre 100 et 200 %.

Pourcentage minimal d'intégrité

Le pourcentage minimal d'intégrité est le pourcentage de la capacité souhaitée pour rester en service, en bon état et prête à être utilisée pour assurer votre charge de travail lors du remplacement d'instances. Une instance est considérée comme saine et prête à être utilisée une fois qu'elle a effectué avec succès son premier contrôle de santé et que le temps de préchauffage spécifié est écoulé. Dans la console, vous pouvez définir le pourcentage minimal d'intégrité

lorsque vous utilisez l'option Résilier et lancer ou l'option Politique personnalisée. Les valeurs valides sont comprises entre 0 et 100 %.

> **Note**
>
> Pour remplacer les instances plus rapidement, vous pouvez définir un faible pourcentage minimal d'intégrité. Toutefois, s'il n'y a pas suffisamment d'instances saines en cours d'exécution, cela peut réduire la disponibilité. Nous vous recommandons de sélectionner une valeur raisonnable afin de maintenir la disponibilité dans les situations où plusieurs instances doivent être remplacées.

Préparation d'instance

Si vos instances ont besoin de temps pour s'initialiser une fois qu'elles sont entrées dans l'état `InService`, activez la préparation d'instance par défaut pour votre groupe Auto Scaling. Grâce à la préparation d'instance par défaut, vous pouvez empêcher que les instances ne soient prises en compte dans le calcul du pourcentage minimal d'intégrité avant qu'elles ne soient prêtes. Cela garantit qu'Amazon EC2 Auto Scaling prend en compte le temps nécessaire pour disposer d'une capacité suffisante pour assurer la charge de travail avant de résilier les instances existantes.

L'avantage supplémentaire est que vous pouvez améliorer les CloudWatch métriques Amazon utilisées pour le dimensionnement dynamique lorsque vous activez le préchauffage de l'instance par défaut. Si votre groupe Auto Scaling dispose de politiques de dimensionnement, lorsqu'il évolue, il utilise la même période de préchauffage par défaut pour éviter que les instances ne soient prises en compte dans CloudWatch les métriques avant la fin de leur initialisation.

Pour plus d'informations, consultez Définir la préparation par défaut d'instance d'un groupe Auto Scaling.

Période de grâce de surveillance de l'état

Amazon EC2 Auto Scaling détermine si une instance est saine en fonction du statut des surveillances de l'état que votre groupe Auto Scaling utilise. Pour plus d'informations, consultez Surveillance de l'état des instances dans un groupe Auto Scaling.

Pour vous assurer que ces surveillances de l'état commencent le plus rapidement possible, ne définissez pas une période de grâce de la surveillance de l'état du groupe trop élevée, mais suffisamment élevée pour que vos surveillances de l'état Elastic Load Balancing déterminent si une

cible est disponible pour traiter les demandes. Pour plus d'informations, consultez Définir la période de grâce de la surveillance de l'état pour un groupe Auto Scaling.

Mettre à l'échelle votre groupe Auto Scaling

Une politique de maintenance des instances s'applique uniquement aux événements de maintenance d'instance et n'empêche pas le redimensionnement manuel ou automatique du groupe.

Lorsque des politiques de mise à l'échelle ou des actions planifiées sont associées à votre groupe Auto Scaling, elles peuvent s'exécuter en parallèle pendant que les événements de maintenance des instances se produisent. Dans ce cas, elles peuvent augmenter ou diminuer la capacité souhaitée du groupe, mais uniquement dans les limites de dimensionnement que vous avez définies. Pour plus d'informations sur ces limites, consultez Définissez des limites de mise à l'échelle pour votre groupe Auto Scaling.

Exemples de scénarios

Dans un scénario typique, votre politique de maintenance des instances et la capacité souhaitée peuvent ressembler à ce qui suit :

- Pourcentage minimal d'intégrité = 90 %
- Pourcentage maximal d'intégrité = 120 %
- Capacité souhaitée = 100

Lors d'un événement de maintenance d'instance, votre groupe Auto Scaling peut compter entre 90 et 120 instances. Après l'événement, le groupe compte à nouveau 100 instances.

Lorsque vous utilisez une politique de maintenance des instances avec un groupe Auto Scaling doté d'un groupe chaud, les pourcentages minimal et maximal d'intégrité sont appliqués séparément au groupe Auto Scaling et au groupe chaud.

Supposons qu'il s'agisse de votre configuration :

- Pourcentage minimal d'intégrité = 90 %
- Pourcentage maximal d'intégrité = 120 %
- Capacité souhaitée = 100
- Taille d'un groupe chaud = 10

Si vous lancez une actualisation d'instance pour recycler les instances du groupe, Amazon EC2 Auto Scaling remplace d'abord les instances du groupe Auto Scaling, puis les instances du groupe chaud. Alors qu'Amazon EC2 Auto Scaling travaille toujours au remplacement des instances du groupe Auto Scaling, le groupe peut compter entre 90 et 120 instances. Une fois la préparation du groupe terminée, Amazon EC2 Auto Scaling peut remplacer les instances du groupe chaud. Pendant ce temps, le groupe chaud peut compter entre 9 et 12 instances.

Définir une politique de maintenance des instances pour votre groupe Auto Scaling

Vous pouvez créer une politique de maintenance des instances au moment de la création d'un groupe Auto Scaling. Vous pouvez également la créer pour les groupes existants.

En définissant une politique de maintenance des instances pour votre groupe Auto Scaling, vous n'avez plus à indiquer de valeurs pour les paramètres de pourcentage minimal et maximal d'intégrité de la fonction d'actualisation des instances, sauf si vous souhaitez remplacer la politique de maintenance des instances.

Dans la console, Amazon EC2 Auto Scaling fournit des options pour vous aider à démarrer.

Table des matières

Définir une politique de maintenance des instances

Pour définir une politique de maintenance des instances pour un groupe Auto Scaling, appliquez l'une des méthodes suivantes :

Console

Pour définir une politique de maintenance des instances pour un nouveau groupe (console)

1. Suivez les instructions de la rubrique Créer un groupe Auto Scaling avec un modèle de lancement et terminez chaque étape de la procédure, jusqu'à l'étape 11.
2. Sur la page Configurer la taille du groupe et les politiques de mise à l'échelle, pour la capacité souhaitée, saisissez le nombre initial d'instances à lancer.

3. Dans la section Mise à l'échelle, sous Limites de mise à l'échelle, si votre nouvelle valeur pour la capacité souhaitée est supérieure à la capacité minimale souhaitée et à la capacité maximale souhaitée, la capacité maximale souhaitée est automatiquement augmentée à la nouvelle valeur de capacité souhaitée. Vous pouvez modifier ces limites si nécessaire.
4. Pour le dimensionnement automatique, indiquez si vous souhaitez créer une politique de dimensionnement de suivi des cibles. Vous pouvez également élaborer cette politique après avoir créé votre groupe Auto Scaling.

   Si vous choisissez la politique de dimensionnement de suivi des cibles, suivez les instructions dans Création d'une politique de suivi des cibles et d'échelonnement pour créer la politique.
5. Dans la section Politique de maintenance des instances, choisissez l'une des options disponibles :

   - Lancer avant toute résiliation : une nouvelle instance doit d'abord être mise en service avant qu'une instance existante puisse être résiliée. Il s'agit d'un bon choix pour les applications qui privilégient la disponibilité plutôt que les économies de coûts.
   - Résilier et lancer : les nouvelles instances sont mises en service en même temps que les instances existantes sont résiliées. Il s'agit d'un bon choix pour les applications qui privilégient les économies de coûts plutôt que la disponibilité. C'est également un bon choix pour les applications qui ne doivent pas libérer plus de capacité que ce qui est actuellement disponible.
   - Politique personnalisée : cette option vous permet de configurer votre politique avec une plage minimale et maximale personnalisée pour la quantité de capacité que vous souhaitez mettre à disposition lors du remplacement d'instances. Cela peut vous aider à trouver le juste équilibre entre le coût et la disponibilité.
6. Pour Définir un pourcentage d'intégrité, saisissez des valeurs pour l'un ou les deux champs suivants. Les champs activés varient en fonction de l'option que vous avez choisie à l'étape précédente.

   - Min : définit le pourcentage minimal d'intégrité requis pour procéder au remplacement des instances.
   - Max : définit le pourcentage maximal d'intégrité possible lors du remplacement d'instances.
7. Développez la section Afficher la capacité pendant les remplacements en fonction de la capacité souhaitée pour confirmer comment les valeurs Min et Max s'appliquent à votre groupe. Les valeurs exactes utilisées dépendent de la valeur de capacité souhaitée, qui changera si le groupe est mis à l'échelle.

8. Poursuivez en effectuant les étapes de la section Créer un groupe Auto Scaling avec un modèle de lancement.

AWS CLI

Pour définir une politique de maintenance des instances pour un nouveau groupe (AWS CLI)

Ajoutez l'option `--instance-maintenance-policy` à la commande create-auto-scaling-group. L'exemple suivant définit une politique de maintenance des instances pour un nouveau groupe Auto Scaling intitulé *my-asg*.

```
aws autoscaling create-auto-scaling-group \
  --launch-template LaunchTemplateName=my-launch-template,Version='1' \
  --auto-scaling-group-name my-asg \
  --min-size 1 \
  --max-size 10 \
  --desired-capacity 5 \
  --default-instance-warmup 20 \
  --instance-maintenance-policy '{
      "MinHealthyPercentage": 90,
      "MaxHealthyPercentage": 120
    }' \
  --vpc-zone-identifier "subnet-5e6example,subnet-613example,subnet-c93example"
```

Console

Pour définir une politique de maintenance des instances pour un groupe existant (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Dans la barre de navigation située en haut de l'écran, choisissez l' Région AWS dans laquelle vous avez créé votre groupe Auto Scaling.
3. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
4. Dans l'onglet Détails, choisissez Politique de maintenance des instances, puis Modifier.
5. Pour définir une politique de maintenance des instances pour le groupe, choisissez l'une des options disponibles :

- Lancer avant toute résiliation : une nouvelle instance doit d'abord être mise en service avant qu'une instance existante puisse être résiliée. Il s'agit d'un bon choix pour les applications qui privilégient la disponibilité plutôt que les économies de coûts.
- Résilier et lancer : les nouvelles instances sont mises en service en même temps que les instances existantes sont résiliées. Il s'agit d'un bon choix pour les applications qui privilégient les économies de coûts plutôt que la disponibilité. C'est également un bon choix pour les applications qui ne doivent pas libérer plus de capacité que ce qui est actuellement disponible.
- Politique personnalisée : cette option vous permet de configurer votre politique avec une plage minimale et maximale personnalisée pour la quantité de capacité que vous souhaitez mettre à disposition lors du remplacement d'instances. Cela peut vous aider à trouver le juste équilibre entre le coût et la disponibilité.

6. Pour Définir un pourcentage d'intégrité, saisissez des valeurs pour l'un ou les deux champs suivants. Les champs activés varient en fonction de l'option que vous avez choisie à l'étape précédente.

   - Min : définit le pourcentage minimal d'intégrité requis pour procéder au remplacement des instances.
   - Max : définit le pourcentage maximal d'intégrité possible lors du remplacement d'instances.

7. Développez la section Afficher la capacité pendant les remplacements en fonction de la capacité souhaitée pour confirmer comment les valeurs Min et Max s'appliquent à votre groupe. Les valeurs exactes utilisées dépendent de la valeur de capacité souhaitée, qui changera si le groupe est mis à l'échelle.
8. Choisissez Mettre à jour.

AWS CLI

Pour définir une politique de maintenance des instances pour un groupe existant (AWS CLI)

Ajoutez l'option `--instance-maintenance-policy` à la commande update-auto-scaling-group. L'exemple suivant définit une politique de maintenance des instances pour le groupe Auto Scaling indiqué.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --instance-maintenance-policy '{
      "MinHealthyPercentage": 90,
```

```
    "MaxHealthyPercentage": 120
  }'
```

Supprimer une politique de maintenance des instances

Si vous souhaitez arrêter d'utiliser une politique de maintenance des instances dans votre groupe Auto Scaling, vous pouvez la supprimer.

Console

Pour supprimer une politique de maintenance des instances (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Dans la barre de navigation située en haut de l'écran, choisissez l' Région AWS dans laquelle vous avez créé votre groupe Auto Scaling.
3. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
4. Dans l'onglet Détails, choisissez Politique de maintenance des instances, puis Modifier.
5. Choisissez Aucune politique de maintenance des instances.
6. Choisissez Mettre à jour.

AWS CLI

Pour supprimer une politique de maintenance des instances (AWS CLI)

Ajoutez l'option `--instance-maintenance-policy` à la commande update-auto-scaling-group. L'exemple suivant supprime la politique de maintenance des instances du groupe Auto Scaling indiqué.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --instance-maintenance-policy '{
      "MinHealthyPercentage": -1,
      "MaxHealthyPercentage": -1
    }'
```

Hooks de cycle de vie Amazon EC2 Auto Scaling

Amazon EC2 Auto Scaling vous permet d'ajouter des hooks de cycle de vie à vos groupes Auto Scaling. Ces hooks vous permettent de créer des solutions qui sont informées des événements du cycle de vie des instances Auto Scaling, puis d'exécuter une action personnalisée sur les instances lorsque l'événement du cycle de vie correspondant se produit. Un hook de cycle de vie fournit à l'action de cycle de vie un délai (défini par défaut sur une heure) pour attendre que l'action se termine avant que l'instance ne passe à l'état suivant.

Exemple d'utilisation de hooks de cycle de vie avec des instances Auto Scaling :

- Lorsqu'un événement de montée en puissance se produit, l'instance nouvellement lancée termine sa séquence de démarrage et passe à un état d'attente. Pendant que l'instance est en attente, elle exécute un script pour télécharger et installer les packages logiciels nécessaires à votre application afin d'être entièrement prête pour commencer à recevoir du trafic. Au terme de l'installation du logiciel, le script envoie la commande complete-lifecycle-action pour poursuivre le processus.
- Lorsqu'un événement de scale-in se produit, un hook du cycle de vie suspend l'instance avant qu'elle ne soit résiliée et vous envoie une notification via Amazon. EventBridge Lorsque l'instance est en état d'attente, vous pouvez appeler une AWS Lambda fonction ou vous connecter à l'instance pour télécharger des journaux ou d'autres données avant que l'instance ne soit complètement arrêtée.

Les hooks de cycle de vie sont souvent utilisés pour déterminer à quel moment les instances sont enregistrées auprès d'Elastic Load Balancing. En ajoutant un hook de cycle de vie de lancement à votre groupe Auto Scaling, vous pouvez vérifier que vos scripts d'amorçage se sont déroulés avec succès et que les applications présentes sur les instances sont prêtes à accepter le trafic avant d'être enregistrées auprès de l'équilibreur de charge à la fin du hook de cycle de vie.

Table des matières

- Effectuer une action de cycle de vie
- Tutoriel : configurer les données utilisateur pour récupérer l'état du cycle de vie cible via des métadonnées d'instance
- Didacticiel : configurer un hook de cycle de vie qui appelle une fonction Lambda

Disponibilité des hooks de cycle de vie

Le tableau suivant répertorie les hooks de cycle de vie disponibles en fonction des différents scénarios.

Événement	Lancement ou résiliation d'instance[1]	Durée de vie maximale de l'instance : instances de remplacement	Actualisation d'instance : instances de remplacement	Rééquilibrage de la capacité : instances de remplacement	Groupes d'instances pré-initialisées : instances entrant et sortant du groupe d'instances pré-initialisées
Lancement d'une instance	✓	✓	✓	✓	✓
Résiliation d'une instance	✓	✓	✓	✓	✓

[1] S’applique à tous les lancements et toutes les résiliations, qu’ils/elles soient initié(e)s automatiquement ou manuellement, par exemple lorsque vous appelez les opérations `SetDesiredCapacity` ou `TerminateInstanceInAutoScalingGroup`. Ne s'applique pas lorsque vous attachez ou détachez des instances, placez des instances en mode veille ou sortez des instances du mode veille, ou supprimez le groupe avec l'option Forcer la suppression.

Considérations et restrictions relatives aux hooks de cycle de vie

Lorsque vous utilisez des hooks de cycle de vie, tenez compte des remarques et limitations suivantes :

- Amazon EC2 Auto Scaling fournit son propre cycle de vie pour faciliter la gestion des groupes Auto Scaling. Ce cycle de vie se distingue de celui des autres instances EC2. Pour plus d'informations, consultez Cycle de vie d'une instance Amazon EC2 Auto Scaling. Les instances d'un groupe d'instances pré-initialisées ont également leur propre cycle de vie, tel que décrit dans Transitions de l'état du cycle de vie pour les instances dans un groupe d'instances pré-initialisées.
- Vous pouvez utiliser des hooks de cycle de vie avec des instances Spot, mais un hook de cycle de vie n'empêche pas la résiliation d'une instance lorsque la capacité requise n'est plus disponible, ce qui peut arriver à tout moment avec un préavis d'interruption de deux minutes. Pour plus d'informations, consultez la section Interruptions des instances Spot dans le guide de l'utilisateur Amazon EC2. Cependant, vous pouvez activer le rééquilibrage des capacités pour remplacer de manière proactive les instances Spot qui ont reçu une recommandation de rééquilibrage du service Amazon EC2 Spot, un signal envoyé lorsqu'une instance Spot présente un risque élevé d'interruption. Pour plus d'informations, consultez Utiliser le rééquilibrage de la capacité pour gérer les interruptions Spot Amazon EC2.
- Les instances peuvent rester en état d'attente pour une durée limitée. Le délai d'expiration par défaut pour un hook de cycle de vie est d'une heure (délai de pulsation). Il existe également un délai d'attente global spécifiant la durée maximale de conservation d'une instance en état d'attente. Le délai d'attente global est de 48 heures ou de 100 fois le délai de pulsation, la valeur la plus faible étant retenue.
- À la fin du hook de cycle de vie, le résultat est abandonner ou continuer. Si une instance est en cours de lancement, Continuer indique que les actions ont abouti et qu'Amazon EC2 Auto Scaling peut mettre l'instance en service. Sinon, ABANDONNER indique que les actions personnalisées ont échoué et que nous pouvons résilier et remplacer l'instance. Si une instance est en cours de résiliation, les paramètres Abandonner et Continuer permettent tous les deux de résilier l'instance. Cependant, le paramètre Abandonner met fin aux actions restantes, notamment les autres hooks de cycle de vie, tandis que le paramètre Continuer permet aux autres hooks de cycle de vie d'aller jusqu'à leur terme.
- Amazon EC2 Auto Scaling limite la vitesse de lancement des instances si les hooks de cycle de vie échouent systématiquement. Assurez-vous donc de tester et de corriger toute erreur permanente dans vos actions de cycle de vie.

- La création et la mise à jour de hooks de cycle de vie à l'aide du AWS CLI AWS CloudFormation, ou d'un SDK fournissent des options non disponibles lors de la création d'un hook de cycle de vie à partir du AWS Management Console. Par exemple, le champ permettant de spécifier l'ARN d'une rubrique SNS ou d'une file d'attente SQS n'apparaît pas dans la console, car Amazon EC2 Auto Scaling envoie déjà des événements à Amazon. EventBridge Ces événements peuvent être filtrés et redirigés vers AWS des services tels que Lambda, Amazon SNS et Amazon SQS selon les besoins.
- Vous pouvez ajouter plusieurs hooks de cycle de vie à un groupe Auto Scaling lors de sa création, en appelant l'CreateAutoScalingGroupAPI à l'aide du AWS CLI AWS CloudFormation, ou d'un SDK. Toutefois, chaque hook doit avoir la même cible de notification et le même rôle IAM, s'il est spécifié. Pour créer des hooks de cycle de vie avec différentes cibles de notification et différents rôles, créez les hooks de cycle de vie un par un dans des appels séparés à l'API PutLifecycleHook.
- Si vous ajoutez un hook de cycle de vie pour le lancement d'une instance, la période de grâce de la surveillance de l'état commence dès que l'instance atteint l'état `InService`. Pour plus d'informations, consultez Définir la période de grâce de la surveillance de l'état pour un groupe Auto Scaling.

Considérations relatives à la mise à l'échelle

- Les politiques de dimensionnement dynamique s'adaptent en fonction des données CloudWatch métriques, telles que le processeur et les E/S réseau, qui sont agrégées sur plusieurs instances. Lors d'une montée en puissance, Amazon EC2 Auto Scaling ne comptabilise pas immédiatement une nouvelle instance dans les métriques d'instance agrégées du groupe Auto Scaling. Il attend que l'instance atteigne l'état `InService` et que la préparation d'instance soit terminée. Pour plus d'informations, consultez Considérations sur les performances de la mise à l'échelle dans la rubrique de préparation d'instance par défaut.
- À grande échelle, les métriques d'instance agrégées peuvent ne pas refléter instantanément la suppression d'une instance en cours de résiliation. Celle-ci cesse d'être comptabilisée dans les métriques d'instance agrégées du groupe peu après le début du workflow de résiliation d'Amazon EC2 Auto Scaling.
- Dans la plupart des cas où des hooks de cycle de vie sont appelés, les activités de mise à l'échelle dues à des politiques de mise à l'échelle simple sont suspendues jusqu'à ce que les actions du cycle de vie soient terminées et que le temps de stabilisation ait expiré. Si vous définissez un intervalle long pour le temps de stabilisation, la reprise de la mise à l'échelle prendra plus de temps. Pour plus d'informations, consultez Les hooks de cycle de vie peuvent entraîner des retards supplémentaires. dans la rubrique de stabilisation. En général, nous vous déconseillons d'utiliser

des politiques de mise à l’échelle simples si vous pouvez plutôt utiliser des politiques de mise à l’échelle d’étape ou de suivi des cibles.

Ressources connexes

Pour une vidéo de présentation, voir AWS re:Invent 2018 : Capacity Management Made Easy with Amazon EC2 Auto Scaling on. YouTube

Nous fournissons quelques extraits de modèles JSON et YAML que vous pouvez utiliser pour comprendre comment déclarer les hooks du cycle de vie dans vos AWS CloudFormation modèles de pile. Pour plus d'informations, consultez la AWS::AutoScaling::LifecycleHookréférence dans le guide de AWS CloudFormation l'utilisateur.

Vous pouvez également consulter notre GitHubréférentiel pour télécharger des exemples de modèles et de scripts de données utilisateur pour les hooks du cycle de vie.

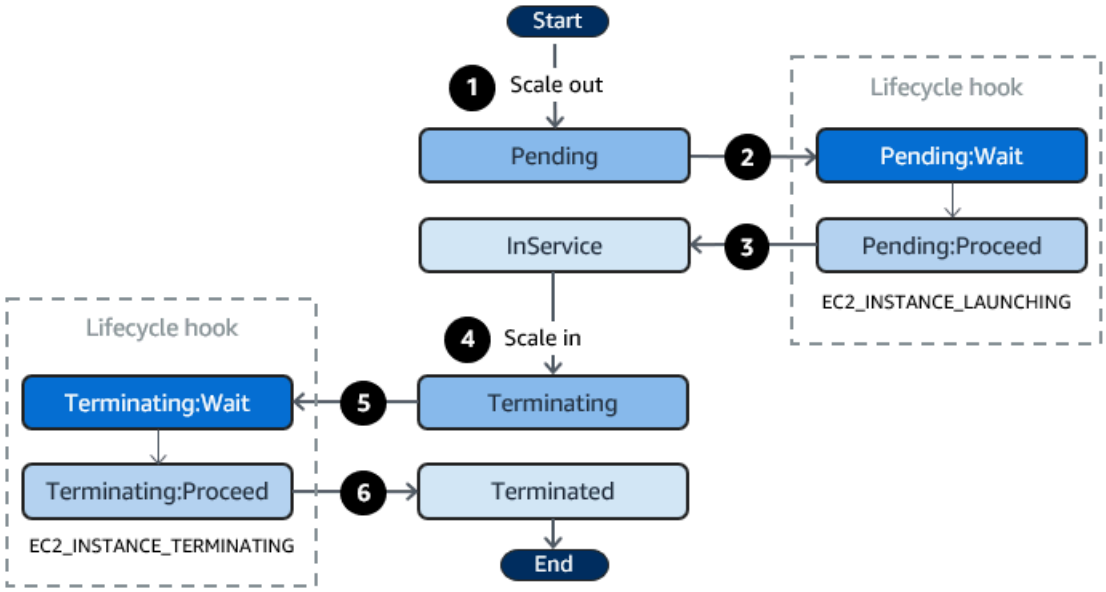
Pour des exemples d’utilisation des hooks de cycle de vie, consultez les articles de blog suivants.

- Création d’un système de sauvegarde pour les instances dimensionnées à l’aide de l’exécution de commande Lambda et Amazon EC2
- Exécutez le code avant de résilier une instance EC2 AutoScaling.

Fonctionnement des hooks de cycle de vie

Une instance Amazon EC2 passe par différents états depuis son lancement jusqu'à sa résiliation. Vous pouvez créer des actions personnalisées correspondant aux réactions de votre groupe Auto Scaling lorsqu’une instance passe à un état d’attente à cause d’un hook de cycle de vie.

L'illustration suivante montre les transitions entre les états des instances d'Auto Scaling lorsque vous utilisez des hooks de cycle de vie pour effectuer une mise à l'échelle externe et une mise à l'échelle interne.



Comme représenté dans le schéma précédent :

1. Le groupe Auto Scaling répond à un événement de montée en puissance et entame le processus de lancement d'une instance.
2. Le hook de cycle de vie met l'instance en attente (état `Pending:Wait`), puis exécute une action personnalisée.

   L'instance reste dans un état d'attente jusqu'à ce que vous terminiez l'action du cycle de vie ou que le délai d'expiration se termine. Par défaut, l'instance reste en attente pendant une heure, puis le groupe Auto Scaling poursuit le processus de lancement (`Pending:Proceed`). Si vous avez besoin de plus de temps, vous pouvez redémarrer le délai d'attente en enregistrant une pulsation. Si vous terminez l'action du cycle de vie alors que l'action personnalisée est terminée et que le délai d'attente n'a pas encore expiré, le groupe Auto Scaling poursuit le processus de lancement.
3. L'instance passe à l'état `InService` et la période de grâce de surveillance de l'état commence. Cependant, avant que l'instance n'affiche l'état `InService`, si le groupe Auto Scaling est associé à un équilibreur de charge Elastic Load Balancing, l'instance est enregistrée auprès de l'équilibreur de charge et celui-ci commence à vérifier son état. Au terme de la période de grâce de surveillance de l'état, Amazon EC2 Auto Scaling commence à vérifier l'état de l'instance.
4. Le groupe Auto Scaling répond à un événement de mise à l'échelle horizontale et entame le processus de résiliation de l'instance. Si le groupe Auto Scaling est utilisé avec Elastic Load Balancing, l'instance en cours de résiliation est d'abord désenregistrée de l'équilibreur de charge. Si Connection Draining est activé pour l'équilibreur de charge, l'instance cesse d'accepter de

nouvelles connexions et attend que les connexions existantes soient drainées avant de finaliser le processus de désenregistrement.

5. Le hook de cycle de vie met l'instance en attente (état `Terminating:Wait`), puis exécute une action personnalisée.

   L'instance reste en attente jusqu'à ce que vous ayez finalisé l'action de cycle de vie, ou jusqu'à ce que le délai d'attente (défini par défaut sur une heure) soit écoulé. Une fois l'exécution du hook de cycle de vie finalisée ou le délai d'attente écoulé, l'instance passe à l'état suivant (`Terminating:Proceed`).

6. L'instance est résiliée.

> **Important**
>
> Les instances d'un groupe d'instances pré-initialisées ont également leur propre cycle de vie avec des états d'attente correspondants, tel que décrit dans Transitions de l'état du cycle de vie pour les instances dans un groupe d'instances pré-initialisées.

Vous préparer à ajouter un hook de cycle de vie à un groupe Auto Scaling

Avant d'ajouter un hook de cycle de vie à votre groupe Auto Scaling, assurez-vous que votre cible de notification ou votre script de données utilisateur est correctement configuré.

- Pour exécuter un script de données utilisateur afin d'effectuer des actions personnalisées sur vos instances lors de leur lancement, vous n'avez pas besoin de configurer de cible de notification. Cependant, vous devez déjà avoir créé le modèle de lancement ou la configuration de lancement qui spécifie votre script de données utilisateur et les avoir associés à votre groupe Auto Scaling. Pour plus d'informations sur les scripts de données utilisateur, consultez la section Exécuter des commandes sur votre instance Linux au lancement dans le guide de l'utilisateur Amazon EC2.
- Pour signaler à Amazon EC2 Auto Scaling que l'action du cycle de vie est terminée, vous devez ajouter l'appel CompleteLifecycled'API Action au script, et vous devez créer manuellement un rôle IAM avec une politique qui autorise les instances Auto Scaling à appeler cette API. Votre modèle de lancement ou votre configuration de lancement doit spécifier ce rôle à l'aide d'un profil d'instance IAM qui est attaché à vos instances Amazon EC2 lors du lancement. Pour plus d'informations, consultez Effectuer une action de cycle de vie et Rôle IAM pour les applications qui s'exécutent sur des instances Amazon EC2.

- Pour utiliser un service tel que Lambda pour effectuer une action personnalisée, vous devez déjà avoir créé une EventBridge règle et spécifié une fonction Lambda comme cible. Pour plus d'informations, consultez Configurer une cible de notification pour les notifications de cycle de vie.
- Pour permettre à Lambda de signaler à Amazon EC2 Auto Scaling une fois l'action du cycle de vie terminée, vous devez CompleteLifecycleajouter l'appel d'API Action au code de fonction. Vous devez également avoir attaché une politique IAM au rôle d'exécution de la fonction pour accorder à Lambda l'autorisation de terminer les actions de cycle de vie. Pour plus d'informations, consultez Didacticiel : configurer un hook de cycle de vie qui appelle une fonction Lambda.
- Pour utiliser un service tel qu'Amazon SNS ou Amazon SQS pour effectuer une action personnalisée, vous devez déjà avoir créé la rubrique SNS ou la file d'attente SQS et disposer de son Amazon Resource Name (ARN). Vous devez également avoir déjà créé le rôle IAM qui donne à Amazon EC2 Auto Scaling l'accès à votre rubrique SNS ou cible SQS et avoir prêt son ARN. Pour plus d'informations, consultez Configurer une cible de notification pour les notifications de cycle de vie.

> **Note**
>
> Par défaut, lorsque vous ajoutez un hook de cycle de vie dans la console, Amazon EC2 Auto Scaling envoie des notifications d'événements liés au cycle de vie à Amazon. EventBridge L'utilisation EventBridge d'un script de données utilisateur est une bonne pratique recommandée. Pour créer un hook de cycle de vie qui envoie des notifications directement à Amazon SNS ou Amazon SQS, utilisez AWS CLI le AWS CloudFormation, ou un SDK pour ajouter le hook de cycle de vie.

Configurer une cible de notification pour les notifications de cycle de vie

Vous pouvez ajouter des hooks de cycle de vie à un groupe Auto Scaling pour exécuter des actions personnalisées lorsqu'une instance entre en état d'attente. Vous pouvez choisir un service cible pour effectuer ces actions en fonction de votre approche de développement préférée.

La première approche utilise Amazon EventBridge pour appeler une fonction Lambda qui exécute l'action souhaitée. La deuxième approche consiste à créer une rubrique Amazon Simple Notification Service (Amazon SNS) dans laquelle les notifications sont publiées. Les clients peuvent s'abonner à la rubrique SNS et recevoir les messages publiés à l'aide d'un protocole pris en charge. La dernière approche consiste à utiliser Amazon Simple Queue Service (Amazon SQS), un système

de messagerie utilisé par les applications distribuées pour échanger des messages via un modèle d'interrogation.

À titre de bonne pratique, nous vous recommandons d'utiliser EventBridge. Les notifications envoyées à Amazon SNS et Amazon SQS contiennent les mêmes informations que celles auxquelles Amazon EC2 Auto Scaling envoie. EventBridge Auparavant EventBridge, la pratique standard consistait à envoyer une notification à SNS ou SQS et à intégrer un autre service à SNS ou SQS pour effectuer des actions programmatiques. Aujourd'hui, vous EventBridge offre davantage d'options pour les services que vous pouvez cibler et facilite la gestion des événements à l'aide d'une architecture sans serveur.

Les procédures suivantes expliquent comment configurer votre cible de notification.

N'oubliez pas que si vous avez un script de données utilisateur dans votre modèle de lancement ou configuration de lancement qui configure vos instances lors de leur lancement, vous n'avez pas besoin de notification pour effectuer des actions personnalisées sur vos instances.

Table des matières

> **Important**
>
> La EventBridge règle, la fonction Lambda, la rubrique Amazon SNS et la file d'attente Amazon SQS que vous utilisez avec les hooks du cycle de vie doivent toujours se trouver dans la même région que celle dans laquelle vous avez créé votre groupe Auto Scaling.

Acheminez les notifications vers Lambda à l'aide de EventBridge

Vous pouvez configurer une EventBridge règle pour appeler une fonction Lambda lorsqu'une instance entre dans un état d'attente. Amazon EC2 Auto Scaling envoie une notification d'événement EventBridge concernant le cycle de vie de l'instance en cours de lancement ou d'arrêt, ainsi qu'un jeton que vous pouvez utiliser pour contrôler l'action du cycle de vie. Pour obtenir des exemples de ces événements, consultez Référence de l'événement Amazon EC2 Auto Scaling.

> **Note**
>
> Lorsque vous utilisez la règle AWS Management Console pour créer un événement, la console ajoute automatiquement les autorisations IAM nécessaires pour EventBridge autoriser l'appel de votre fonction Lambda. Si vous créez une règle d'événement à l'aide de l' AWS CLI, vous devez explicitement accorder cette autorisation.
> Pour plus d'informations sur la création de règles d'événements dans la EventBridge console, consultez la section Création de EventBridge règles Amazon qui réagissent aux événements dans le guide de EventBridge l'utilisateur Amazon.
> – ou –
> Pour obtenir un tutoriel d'introduction destiné aux utilisateurs de la console, consultez la rubrique Didacticiel : configurer un hook de cycle de vie qui appelle une fonction Lambda. Ce didacticiel explique comment créer une fonction Lambda simple qui écoute les événements de lancement et les enregistre dans un CloudWatch journal des journaux.

Pour créer une EventBridge règle qui invoque une fonction Lambda

1. Créez une fonction Lambda à l'aide de la console Lambda et notez son Amazon Resource Name (ARN). Par exemple, `arn:aws:lambda:`*`region`*`:`*`123456789012`*`:function:`*`my-function`*. Vous avez besoin de l'ARN pour créer une EventBridge cible. Pour plus d'informations, consultez Prise en main de Lambda dans le Guide du développeur AWS Lambda .
2. Pour créer une règle qui correspond aux événements de lancement d'une instance, utilisez la commande put-rule suivante.

   ```
   aws events put-rule --name my-rule --event-pattern file://pattern.json --state
    ENABLED
   ```

   L'exemple suivant illustre le fichier `pattern.json` correspondant à une action de cycle de vie de lancement d'instance. Remplacez le texte en *`italique`* par le nom de votre groupe Auto Scaling.

   ```
   {
     "source": [ "aws.autoscaling" ],
     "detail-type": [ "EC2 Instance-launch Lifecycle Action" ],
     "detail": {
         "AutoScalingGroupName": [ "my-asg" ]
      }
   ```

```
}
```

Si la commande s'exécute correctement, elle EventBridge répond avec l'ARN de la règle. Notez cet ARN. Vous devrez le saisir au cours de l'étape 4.

Pour créer une règle qui correspond à d'autres événements, modifiez le modèle d'événement. Pour plus d'informations, consultez EventBridge À utiliser pour gérer les événements Auto Scaling.

3. Pour spécifier la fonction Lambda à utiliser comme cible pour la règle, utilisez la commande put-targets suivante :

```
aws events put-targets --rule my-rule --targets
 Id=1,Arn=arn:aws:lambda:region:123456789012:function:my-function
```

Dans la commande précédente, *my-rule* est le nom que vous avez spécifié pour la règle lors de l'étape 2, et la valeur du paramètre `Arn` est l'ARN de la fonction que vous avez créé à l'étape 1.

4. Pour ajouter des autorisations permettant à la règle d'invoquer votre fonction Lambda cible, utilisez la commande add-permission Lambda suivante. Cette commande fait confiance au principal de EventBridge service (`events.amazonaws.com`) et étend les autorisations à la règle spécifiée.

```
aws lambda add-permission --function-name my-function --statement-id my-unique-id \
  --action 'lambda:InvokeFunction' --principal events.amazonaws.com --source-arn
 arn:aws:events:region:123456789012:rule/my-rule
```

Dans la commande précédente :

- *my-function* est le nom de la fonction Lambda que vous souhaitez que la règle utilise comme cible.
- *my-unique-id.* est un identifiant unique que vous définissez pour décrire l'instruction dans la politique de la fonction Lambda.
- `source-arn`est l'ARN de la EventBridge règle.

Si la commande s'exécute correctement, vous recevez une sortie similaire à ce qui suit.

```
{
```

```
  "Statement": "{\"Sid\":\"my-unique-id\",
    \"Effect\":\"Allow\",
    \"Principal\":{\"Service\":\"events.amazonaws.com\"},
    \"Action\":\"lambda:InvokeFunction\",
    \"Resource\":\"arn:aws:lambda:us-west-2:123456789012:function:my-function\",
    \"Condition\":
      {\"ArnLike\":
        {\"AWS:SourceArn\":
         \"arn:aws:events:us-west-2:123456789012:rule/my-rule\"}}}"
}
```

La valeur `Statement` est une version de la chaîne JSON correspondant à l'instruction ajoutée à la politique de la fonction Lambda.

5. Une fois que vous avez suivi ces instructions, passez à Ajouter des hooks de cycle de vie qui est l'étape suivante.

Recevoir des notifications à l'aide d'Amazon SNS

Vous pouvez utiliser Amazon SNS pour configurer une cible de notification (une rubrique SNS) permettant de recevoir des notifications lorsqu'une action de cycle de vie se produit. Amazon SNS envoie ensuite les notifications aux destinataires abonnés. Aucune notification publiée dans la rubrique n'est envoyée aux destinataires tant que l'abonnement n'est pas confirmé.

Pour configurer des notifications à l'aide d'Amazon SNS

1. Créez une rubrique Amazon SNS à l'aide de la console Amazon SNS ou de la commande create-topic suivante. Assurez-vous que la rubrique se trouve dans la même région que le groupe Auto Scaling que vous utilisez. Pour plus d'informations, consultez Prise en main d'Amazon SNS dans le Guide du développeur Amazon Simple Notification Service.

   ```
   aws sns create-topic --name my-sns-topic
   ```

2. Notez l'Amazon Resource Name (ARN) de la rubrique, par exemple `arn:aws:sns:region:123456789012:my-sns-topic`. Celui-ci est nécessaire pour créer le hook de cycle de vie.
3. Créez un rôle de service IAM pour accorder à Amazon EC2 Auto Scaling l'autorisation d'accès à votre cible de notification Amazon SNS.

   Pour accorder à Amazon EC2 Auto Scaling l'autorisation d'accès à votre rubrique SNS

a. Ouvrez la console IAM à l'adresse https://console.aws.amazon.com/iam/.

b. Dans le panneau de navigation de gauche, sélectionnez Roles (Rôles).

c. Sélectionnez Create role (Créer un rôle).

d. Pour Select trusted entity (Sélectionner une entité de confiance), choisissez service AWS .

e. Pour votre cas d'utilisation, sous Use cases for other AWS services (Cas d'utilisation sez EC2 Auto Scaling puis EC2 Auto Scaling Notification Access (Notification d'accès EC2 Auto Scaling).

f. Cliquez deux fois sur Next (Suivant) pour aller à la page Name, review, and create (Nommer, vérifier et créer).

g. Pour Role Name (Nom du rôle), saisissez un nom pour votre rôle (par exemple **`my-notification-role`**), puis sélectionnez Create Role (Créer un rôle).

h. Sur la page Roles (Rôles), choisissez le rôle que vous venez de créer pour ouvrir la page Summary (Récapitulatif). Notez l'ARN du rôle. Par exemple, `arn:aws:iam::123456789012:role/my-notification-role`. Celui-ci est nécessaire pour créer le hook de cycle de vie.

4. Une fois que vous avez suivi ces instructions, passez à Ajouter des hooks de cycle de vie (AWS CLI) qui est l'étape suivante.

Recevoir des notifications à l'aide d'Amazon SQS

Vous pouvez utiliser Amazon SQS pour configurer une cible de notification permettant de recevoir des messages lorsqu'une action de cycle de vie se produit. Un consommateur de file d'attente doit alors interroger une file d'attente SQS pour agir sur ces notifications.

> **Important**
>
> Les files d'attente FIFO ne sont pas compatibles avec des hooks de cycle de vie.

Pour configurer des notifications à l'aide d'Amazon SQS

1. Créez une file d'attente Amazon SQS à l'aide de la console Amazon SQS. Assurez-vous que la file d'attente se trouve dans la même région que le groupe Auto Scaling que vous utilisez. Pour plus d'informations, consultez Prise en main d'Amazon SQS dans le Guide du développeur Amazon Simple Queue Service.

2. Notez l'ARN de la file d'attente, par exemple `arn:aws:sqs:us-west-2:123456789012:my-sqs-queue`. Celui-ci est nécessaire pour créer le hook de cycle de vie.
3. Créez un rôle de service IAM pour accorder à Amazon EC2 Auto Scaling l'autorisation d'accès à votre cible de notification Amazon SQS.

   Pour accorder à Amazon EC2 Auto Scaling l'autorisation d'accès à votre file d'attente SQS

   a. Ouvrez la console IAM à l'adresse https://console.aws.amazon.com/iam/.
   b. Dans le panneau de navigation de gauche, sélectionnez Roles (Rôles).
   c. Sélectionnez Create role (Créer un rôle).
   d. Pour Select trusted entity (Sélectionner une entité de confiance), choisissez service AWS .
   e. Pour votre cas d'utilisation, sous Use cases for other AWS services (Cas d'utilisation sez EC2 Auto Scaling puis EC2 Auto Scaling Notification Access (Notification d'accès EC2 Auto Scaling).
   f. Cliquez deux fois sur Next (Suivant) pour aller à la page Name, review, and create (Nommer, vérifier et créer).
   g. Pour Role Name (Nom du rôle), saisissez un nom pour votre rôle (par exemple **`my-notification-role`**), puis sélectionnez Create Role (Créer un rôle).
   h. Sur la page Roles (Rôles), choisissez le rôle que vous venez de créer pour ouvrir la page Summary (Récapitulatif). Notez l'ARN du rôle. Par exemple, `arn:aws:iam::123456789012:role/my-notification-role`. Celui-ci est nécessaire pour créer le hook de cycle de vie.
4. Une fois que vous avez suivi ces instructions, passez à Ajouter des hooks de cycle de vie (AWS CLI) qui est l'étape suivante.

Exemple de message de notification pour Amazon SNS et Amazon SQS

Pendant que l'instance est en attente, un message est publié sur la cible de notification Amazon SNS ou Amazon SQS. Le message comprend les informations suivantes :

- `LifecycleActionToken` : jeton de l'action de cycle de vie.
- `AccountId`— La Compte AWS pièce d'identité.
- `AutoScalingGroupName` : nom du groupe Auto Scaling.
- `LifecycleHookName` : nom du hook de cycle de vie.
- `EC2InstanceId` : ID de l'instance EC2.

- `LifecycleTransition` : type du hook de cycle de vie.
- `NotificationMetadata` : métadonnées de notification.

Voici un exemple de message de notification.

```
Service: AWS Auto Scaling
Time: 2021-01-19T00:36:26.533Z
RequestId: 18b2ec17-3e9b-4c15-8024-ff2e8ce8786a
LifecycleActionToken: 71514b9d-6a40-4b26-8523-05e7ee35fa40
AccountId: 123456789012
AutoScalingGroupName: my-asg
LifecycleHookName: my-hook
EC2InstanceId: i-0598c7d356eba48d7
LifecycleTransition: autoscaling:EC2_INSTANCE_LAUNCHING
NotificationMetadata: hook message metadata
```

Exemple de message de notification de test

Lorsque vous ajoutez pour la première fois un hook de cycle de vie, un message de notification de test est publié sur la cible de notification. Voici un exemple de message de notification de test.

```
Service: AWS Auto Scaling
Time: 2021-01-19T00:35:52.359Z
RequestId: 18b2ec17-3e9b-4c15-8024-ff2e8ce8786a
Event: autoscaling:TEST_NOTIFICATION
AccountId: 123456789012
AutoScalingGroupName: my-asg
AutoScalingGroupARN: arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:042cba90-
ad2f-431c-9b4d-6d9055bcc9fb:autoScalingGroupName/my-asg
```

> **Note**
>
> Pour des exemples d'événements proposés par Amazon EC2 Auto Scaling à EventBridge, consultez. Référence de l'événement Amazon EC2 Auto Scaling

Récupérer l'état du cycle de vie cible via des métadonnées d'instance

Chaque instance Auto Scaling que vous lancez connaît plusieurs états de cycle de vie. Pour invoquer des actions personnalisées dans une instance pour agir sur des transitions d'état du cycle de vie spécifiques, vous devez récupérer l'état du cycle de vie cible via des métadonnées de l'instance.

Par exemple, vous pourriez avoir besoin d'un mécanisme pour détecter la résiliation d'une instance à l'intérieur de l'instance afin d'exécuter du code sur l'instance avant qu'elle ne soit résiliée. Vous pouvez le faire en écrivant du code qui interroge l'état du cycle de vie d'une instance directement à partir de celle-ci. Vous pouvez ensuite ajouter un hook de cycle de vie au groupe Auto Scaling pour que l'instance continue de fonctionner jusqu'à ce que votre code informe la commande complete-lifecycle-action de continuer.

Le cycle de vie de l'instance Auto Scaling comporte deux états stables primaires (`InService` et `Terminated`) et deux états stables secondaires (`Detached` et `Standby`). Si vous utilisez un groupe d'instances pré-initialisées, le cycle de vie comporte quatre états stables supplémentaires : `Warmed:Hibernated`, `Warmed:Running`, `Warmed:Stopped` et `Warmed:Terminated`.

Lorsqu'une instance se prépare à passer à l'un des états stables précédents, Amazon EC2 Auto Scaling met à jour la valeur de l'élément de métadonnées d'instance `autoscaling/target-lifecycle-state`. Pour obtenir l'état du cycle de vie cible depuis l'instance, vous devez utiliser le service de métadonnées d'instance pour le récupérer à partir des métadonnées de l'instance.

> **Note**
>
> Les métadonnées de l'instance sont des données relatives à une instance Amazon EC2 que les applications peuvent utiliser pour demander des informations sur l'instance. Le service des métadonnées d'instance est un composant sur instance que le code local utilise pour accéder aux métadonnées d'instance. Le code local peut inclure des scripts de données utilisateur ou des applications exécutées sur l'instance.

Le code local peut accéder aux métadonnées d'instance à partir d'une instance en cours d'exécution à l'aide de l'une des deux méthodes suivantes : Instance Metadata Service Version 1 (IMDSv1) ou Instance Metadata Service Version 2 (IMDSv2). IMDSv2 utilise des requêtes orientées session et atténue plusieurs types de vulnérabilités qui pourraient être utilisées pour essayer d'accéder aux métadonnées d'instance. Pour en savoir plus sur ces deux méthodes, consultez la section Utiliser IMDSv2 dans le guide de l'utilisateur Amazon EC2.

IMDSv2

```
[ec2-user ~]$ TOKEN=`curl -X PUT "http://169.254.169.254/latest/api/token" -H "X-aws-ec2-metadata-token-ttl-seconds: 21600"` \
&& curl -H "X-aws-ec2-metadata-token: $TOKEN" -v http://169.254.169.254/latest/meta-data/autoscaling/target-lifecycle-state
```

IMDSv1

```
[ec2-user ~]$ curl http://169.254.169.254/latest/meta-data/autoscaling/target-lifecycle-state
```

Voici un exemple de sortie.

```
InService
```

L'état du cycle de vie cible est l'état vers lequel l'instance est en transition. L'état actuel du cycle de vie correspond à l'état dans lequel se trouve l'instance. Ils peuvent être identiques une fois l'action du cycle de vie terminée et que l'instance a terminé sa transition vers l'état du cycle de vie cible. Vous ne pouvez pas récupérer l'état actuel du cycle de vie de l'instance à partir des métadonnées de l'instance.

Amazon EC2 Auto Scaling a commencé à générer l'état du cycle de vie cible le 10 mars 2022. Si votre instance passe à l'un des états du cycle de vie cible après cette date, l'élément d'état du cycle de vie cible est présent dans les métadonnées de l'instance. Sinon, il n'est pas présent et vous recevez une erreur HTTP 404.

Pour plus d'informations sur la récupération des métadonnées d'instance, consultez la section Récupérer les métadonnées d'instance dans le guide de l'utilisateur Amazon EC2.

Pour obtenir un tutoriel qui vous montre comment créer un hook de cycle de vie avec une action personnalisée dans un script de données utilisateur utilisant l'état du cycle de vie cible, reportez-vous à la section Tutoriel : configurer les données utilisateur pour récupérer l'état du cycle de vie cible via des métadonnées d'instance.

> **Important**
>
> Pour que vous puissiez invoquer une action personnalisée dès que possible, votre code local doit fréquemment interroger IMDS et réessayer en cas d'erreur.

Ajouter des hooks de cycle de vie

Pour mettre vos instances Auto Scaling en état d'attente et effectuer des actions personnalisées sur celles-ci, vous pouvez ajouter des hooks de cycle de vie à votre groupe Auto Scaling. Les actions personnalisées sont exécutées au lancement des instances ou avant qu'elles ne se terminent. Les instances restent dans un état d'attente jusqu'à ce que vous terminiez l'action du cycle de vie ou que le délai d'expiration se termine.

Après avoir créé un groupe Auto Scaling à partir du AWS Management Console, vous pouvez y ajouter un ou plusieurs hooks de cycle de vie, jusqu'à un total de 50 hooks de cycle de vie. Vous pouvez également utiliser le AWS CLI AWS CloudFormation, ou un SDK pour ajouter des hooks de cycle de vie à un groupe Auto Scaling lors de sa création.

Par défaut, lorsque vous ajoutez un hook de cycle de vie dans la console, Amazon EC2 Auto Scaling envoie des notifications d'événements liés au cycle de vie à Amazon. EventBridge L'utilisation EventBridge d'un script de données utilisateur est une bonne pratique recommandée. Pour créer un hook de cycle de vie qui envoie des notifications directement à Amazon SNS ou Amazon SQS, vous pouvez utiliser la commande put-lifecycle-hook comme illustré dans les exemples de cette rubrique.

Table des matières

Ajouter des hooks de cycle de vie (console)

Procédez comme suit pour ajouter des hooks de cycle de vie à votre groupe Auto Scaling. Pour ajouter des hooks de cycle de vie pour la montée en puissance (lancement d'instances) et la mise à l'échelle horizontale (résiliation d'instances ou renvois dans le groupe chaud), vous devez créer deux hooks distincts.

Avant de commencer, confirmez que vous avez configuré une action personnalisée, selon vos besoins, comme décrit dans Vous préparer à ajouter un hook de cycle de vie à un groupe Auto Scaling.

Pour ajouter un hook de cycle de vie destiné à la montée en puissance

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling. Un volet fractionné s'ouvre en bas de la page.
3. Sous l'onglet Instance management (Gestion des instances) dans Lifecycle hooks (Hooks du cycle de vie), choisissez Create lifecycle hook (Créer un hook de cycle de vie.
4. Pour définir un hook de cycle de vie pour la montée en puissance (lancement d'instances), procédez comme suit :

   a. Pour le Lifecycle Hook Name (Nom du hook de cycle de vie), spécifiez un nom pour le hook de cycle de vie.

   b. Dans le champ Lifecycle transition (Transition du cycle de vie), choisissez Instance launch (Lancement d'instance).

   c. Pour Délai de pulsation, spécifiez la durée (en secondes) pendant laquelle les instances doivent rester en état d'attente lors de l'évolutivité horizontale avant l'expiration du hook. La plage est comprise entre `30` et `7200` secondes. La définition d'une longue période de délai d'attente donne plus de temps à votre action personnalisée pour aboutir. Puis, si vous terminez avant la fin du délai d'expiration, envoyez la commande complete-lifecycle-action pour permettre à l'instance de passer à l'état suivant.

   d. Pour Default result (Résultat par défaut), définissez l'action à entreprendre lorsque le délai d'attente du hook de cycle de vie est écoulé ou qu'un échec inattendu se produit. Vous pouvez choisir d'ABANDONNER ou de CONTINUER.

      - Si vous sélectionnez CONTINUER, le groupe Auto Scaling peut exécuter n'importe quel hook de cycle de vie, puis mettre l'instance en service.
      - Si vous choisissez ABANDONNER, le groupe Auto Scaling interrompt les actions restantes et résilie immédiatement l'instance.

   e. (Facultatif) Dans le champ Métadonnées de notification, spécifiez les informations supplémentaires que vous souhaitez inclure lorsque Amazon EC2 Auto Scaling envoie un message à la cible de notification.

5. Choisissez Créer.

Pour ajouter un hook de cycle de vie destiné à la mise à l’échelle horizontale

1. Choisissez Créer un hook de cycle de vie pour continuer là où vous vous êtes arrêté après la création d’un hook de cycle de vie destiné à la montée en puissance.
2. Pour définir un hook de cycle de vie pour la mise à l’échelle horizontale (instances résiliées ou revenant à un groupe chaud), procédez comme suit :

   a. Pour le Lifecycle Hook Name (Nom du hook de cycle de vie), spécifiez un nom pour le hook de cycle de vie.

   b. Dans le champ Transition du cycle de vie, choisissez Résiliation d'instance.

   c. Pour Délai de pulsation, spécifiez la durée (en secondes) pendant laquelle les instances doivent rester en état d’attente lors de l’évolutivité horizontale avant l’expiration du hook. Nous recommandons un court délai d'attente de deux `30` à `120` secondes, en fonction du temps dont vous avez besoin pour effectuer les tâches finales, telles que l'extraction des journaux EC2. CloudWatch

   d. Dans le champ Default result (Résultat par défaut), spécifiez l'action que le groupe Auto Scaling doit entreprendre lorsque le délai d'attente est écoulé ou qu'un échec inattendu se produit. Les paramètres ABANDON (ABANDONNER) et CONTINUE (CONTINUER) permettent tous les deux de résilier l'instance.

      - Si vous choisissez CONTINUE (CONTINUER), le groupe Auto Scaling peut exécuter toutes les actions restantes, comme les hooks de cycle de vie, avant la résiliation.
      - Si vous choisissez ABANDONNER, le groupe Auto Scaling résilie immédiatement l’instance.

   e. (Facultatif) Dans le champ Métadonnées de notification, spécifiez les informations supplémentaires que vous souhaitez inclure lorsque Amazon EC2 Auto Scaling envoie un message à la cible de notification.

3. Choisissez Créer.

Ajouter des hooks de cycle de vie (AWS CLI)

Pour créer et mettre à jour des hooks de cycle de vie, utilisez la commande put-lifecycle-hook.

Pour exécuter une action sur l'augmentation de la taille des instances, utilisez la commande suivante.

```
aws autoscaling put-lifecycle-hook --lifecycle-hook-name my-launch-hook  \
  --auto-scaling-group-name my-asg \
  --lifecycle-transition autoscaling:EC2_INSTANCE_LAUNCHING
```

Pour exécuter une action sur la diminution de la taille des instances, ajoutez plutôt la commande suivante.

```
aws autoscaling put-lifecycle-hook --lifecycle-hook-name my-termination-hook  \
  --auto-scaling-group-name my-asg \
  --lifecycle-transition autoscaling:EC2_INSTANCE_TERMINATING
```

Pour recevoir des notifications à l'aide d'Amazon SNS ou d'Amazon SQS, ajoutez les options `--notification-target-arn` et `--role-arn`.

L'exemple suivant crée un hook de cycle de vie qui spécifie une rubrique SNS nommée *my-sns-topic* comme cible de notification.

```
aws autoscaling put-lifecycle-hook --lifecycle-hook-name my-termination-hook  \
  --auto-scaling-group-name my-asg \
  --lifecycle-transition autoscaling:EC2_INSTANCE_TERMINATING \
  --notification-target-arn arn:aws:sns:region:123456789012:my-sns-topic \
  --role-arn arn:aws:iam::123456789012:role/my-notification-role
```

La rubrique reçoit une notification test avec la paire clé-valeur suivante.

```
"Event": "autoscaling:TEST_NOTIFICATION"
```

Par défaut, la commande put-lifecycle-hook crée un hook de cycle de vie avec un délai de pulsation de `3600` secondes (1 heure).

Pour modifier le délai de pulsation d'un hook de cycle de vie existant, ajoutez le paramètre `--heartbeat-timeout`, comme illustré dans l'exemple suivant.

```
aws autoscaling put-lifecycle-hook --lifecycle-hook-name my-termination-hook \
  --auto-scaling-group-name my-asg --heartbeat-timeout 120
```

Si une instance est déjà en état d'attente, vous pouvez empêcher l'interruption du hook de cycle de vie en enregistrant une pulsation à l'aide de la commande CLI record-lifecycle-action-heartbeat. Cela prolonge le délai d'attente de la valeur d'attente spécifiée lorsque vous créez le hook de cycle de

vie. Si vous terminez avant la fin du délai d'expiration, envoyez la commande CLI complete-lifecycle-action pour permettre à l'instance de passer à l'état suivant. Pour plus d'informations et d'exemples, consultez Effectuer une action de cycle de vie.

Effectuer une action de cycle de vie

Lorsqu'un groupe Auto Scaling répond à un événement de cycle de vie, il met l'instance en attente et envoie une notification d'événement. Pendant que l'instance est en attente, vous pouvez exécuter une action personnalisée.

Il est utile d'effectuer l'action du cycle de vie avec le résultat `CONTINUE` si vous terminez avant l'expiration du délai imparti. Si vous n'effectuez pas l'action du cycle de vie, le hook de cycle de vie passe au statut que vous avez indiqué pour le résultat par défaut une fois le délai expiré.

Table des matières

Effectuer une action de cycle de vie (manuel)

La procédure suivante s'applique à l'interface de ligne de commande et n'est pas prise en charge dans la console. Les informations qui doivent être remplacées, comme l'ID de l'instance ou le nom d'un groupe Auto Scaling, sont en italique.

Pour exécuter une action de cycle de vie (AWS CLI)

1. Si vous avez besoin de plus de temps pour terminer l'action personnalisée, utilisez la commande record-lifecycle-action-heartbeat pour redémarrer le délai d'attente et conserver l'instance en état d'attente. Par exemple, si la valeur d'attente est d'1 heure, et que vous appelez cette commande après 30 minutes, l'instance reste en état d'attente pendant 1 heure supplémentaire, soit un total de 90 minutes.

   Vous pouvez spécifier le jeton d'action du cycle de vie que vous avez reçu avec la notification, comme indiqué dans la commande suivante.

   ```
   aws autoscaling record-lifecycle-action-heartbeat --lifecycle-hook-name my-launch-hook \
     --auto-scaling-group-name my-asg --lifecycle-action-token bcd2f1b8-9a78-44d3-8a7a-4dd07d7cf635
   ```

Alternativement, vous pouvez spécifier l'ID de l'instance que vous avez reçu avec la notification, comme indiqué dans la commande suivante.

```
aws autoscaling record-lifecycle-action-heartbeat --lifecycle-hook-name my-launch-hook \
  --auto-scaling-group-name my-asg --instance-id i-1a2b3c4d
```

2. Si vous terminez l'action personnalisée avant la fin du délai d'attente, utilisez la commande complete-lifecycle-action pour que le groupe Auto Scaling puisse poursuivre le processus de lancement ou de résiliation de l'instance. Vous pouvez spécifier le jeton de l'action du cycle de vie, comme illustré dans la commande suivante.

```
aws autoscaling complete-lifecycle-action --lifecycle-action-result CONTINUE \
  --lifecycle-hook-name my-launch-hook --auto-scaling-group-name my-asg \
  --lifecycle-action-token bcd2f1b8-9a78-44d3-8a7a-4dd07d7cf635
```

Sinon, vous pouvez spécifier l'ID de l'instance, comme illustré dans la commande suivante.

```
aws autoscaling complete-lifecycle-action --lifecycle-action-result CONTINUE \
  --instance-id i-1a2b3c4d --lifecycle-hook-name my-launch-hook \
  --auto-scaling-group-name my-asg
```

Effectuer une action de cycle de vie (automatique)

Si vous disposez d'un script de données utilisateur qui configure vos instances après leur lancement, vous n'avez pas besoin d'effectuer manuellement les actions du cycle de vie. Vous pouvez ajouter la commande complete-lifecycle-action au script. Le script peut récupérer l'ID d'instance à partir des métadonnées d'instance et signaler à Amazon EC2 Auto Scaling lorsque les scripts d'amorçage se sont terminés correctement.

Si ce n'est pas le cas, mettez à jour le script pour récupérer l'ID de l'instance à partir de ses métadonnées. Pour plus d'informations, consultez la section Récupérer les métadonnées d'une instance dans le guide de l'utilisateur Amazon EC2.

Si vous utilisez Lambda, vous pouvez également configurer un rappel dans le code de votre fonction pour laisser le cycle de vie de l'instance se poursuivre si l'action personnalisée réussit. Pour plus d'informations, consultez Didacticiel : configurer un hook de cycle de vie qui appelle une fonction Lambda.

Tutoriel : configurer les données utilisateur pour récupérer l'état du cycle de vie cible via des métadonnées d'instance

Une méthode courante pour créer des actions personnalisées pour les hooks du cycle de vie consiste à utiliser les notifications qu'Amazon EC2 Auto Scaling envoie à d'autres services, tels qu'Amazon. EventBridge Toutefois, vous pouvez éviter de devoir créer une infrastructure supplémentaire en utilisant plutôt un script de données utilisateur pour déplacer le code qui configure les instances et effectue l'action du cycle de vie dans les instances elles-mêmes.

Le tutoriel suivant vous montre comment commencer à utiliser un script de données utilisateur et des métadonnées d'instance. Vous créez une configuration de groupe Auto Scaling de base avec un script de données utilisateur qui lit l'état du cycle de vie cible des instances de votre groupe et effectue une action de rappel à une phase spécifique du cycle de vie d'une instance pour poursuivre le processus de lancement.

L'illustration suivante résume le flux d'un événement de scale-out lorsque vous utilisez un script de données utilisateur pour effectuer une action personnalisée. Après le lancement d'une instance, le cycle de vie de l'instance est suspendu jusqu'à ce que le cycle de vie soit terminé, soit en expirant, soit en recevant un signal indiquant à Amazon EC2 Auto Scaling de continuer.

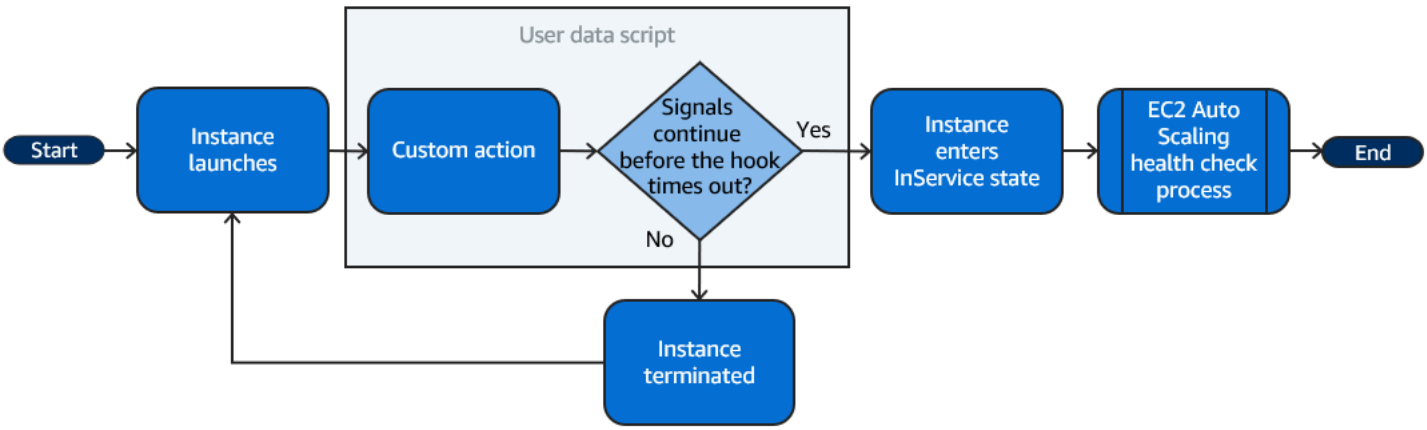


Table des matières

Étape 1 : créer un rôle IAM doté des autorisations nécessaires pour utiliser des actions de cycle de vie

Lorsque vous utilisez le AWS CLI ou un AWS SDK pour envoyer un rappel pour effectuer des actions du cycle de vie, vous devez utiliser un rôle IAM avec des autorisations pour effectuer les actions du cycle de vie.

Pour créer la politique

1. Ouvrez la page Policies (Politiques) de la console IAM et sélectionnez Create policy (Créer une politique).
2. Sélectionnez l'onglet JSON.
3. Dans la case Policy Document (Document de politique), copiez et collez le document de politique suivant dans la case. Remplacez le *sample text* avec votre numéro de compte et le nom du groupe Auto Scaling que vous souhaitez créer (**TestAutoScalingEvent-group**).

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "autoscaling:CompleteLifecycleAction"
      ],
      "Resource":
 "arn:aws:autoscaling:*:123456789012:autoScalingGroup:*:autoScalingGroupName/
TestAutoScalingEvent-group"
    }
  ]
}
```

4. Choisissez Suivant.
5. Pour Policy name (Nom de la politique), saisissez **TestAutoScalingEvent-policy**. Sélectionnez Create policy (Créer une politique).

Lorsque vous avez créé la politique, vous pouvez créer un rôle qui l'utilise.

Pour créer le rôle

1. Dans le panneau de navigation de gauche, sélectionnez Roles (Rôles).
2. Sélectionnez Create role (Créer un rôle).
3. Pour Select trusted entity (Sélectionner une entité de confiance), choisissez service AWS .
4. Pour votre cas d'utilisation, sélectionnez EC2, puis Next (Suivant).
5. Sous Ajouter des autorisations, choisissez la politique que vous avez créée (TestAutoScalingEvent-policy). Ensuite, choisissez Next (Suivant).
6. Sur la page Name, review, and create (Nommer, vérifier et créer), pour Role name (Nom du rôle), saisissez **`TestAutoScalingEvent-role`**, puis choisissez Create role (Créer un rôle).

Étape 2 : créer un modèle de lancement et inclure le rôle IAM et un script de données utilisateur

Créer un modèle de lancement à utiliser avec un groupe Auto Scaling. Incluez le rôle IAM que vous avez créé et l'exemple de script de données utilisateur fourni.

Pour créer un modèle de lancement

1. Ouvrez la page des modèles de lancement de la console Amazon EC2.
2. Choisissez Create launch template (Créer un modèle de lancement).
3. Pour Launch template name (Nom du modèle de lancement), saisissez **`TestAutoScalingEvent-template`**.
4. Sous Guide Auto Scaling, activez la case à cocher.
5. Pour Application and OS Images (Amazon Machine Image) (Amazon machine image [AMI]), choisissez Amazon Linux 2 (HVM), Type de volume SSD, 64 bits (x86) dans la liste Quick Start (Démarrage rapide).
6. Pour Instance type (Type d'instance), choisissez un type d'instance Amazon EC2 (par exemple, « t2.micro »).
7. Pour Advanced details (Détails avancés), développez la section afin d'afficher les champs.
8. Pour le profil d'instance IAM, choisissez le nom du profil d'instance IAM de votre rôle IAM (-role) TestAutoScalingEvent. Un profil d'instance est un conteneur pour un rôle IAM qui permet à Amazon EC2 de transmettre le rôle IAM à une instance lors du lancement de l'instance.

Lorsque vous avez utilisé la console IAM pour créer un rôle IAM, la console a automatiquement créé un profil d'instance portant le même nom que le rôle correspondant.

9. Pour User data (Données utilisateur), copiez et collez l'exemple de script de données utilisateur suivant dans le champ. Remplacez le texte d'exemple `group_name` par le nom du groupe Auto Scaling que vous souhaitez créer et `region` par le nom que Région AWS vous souhaitez que votre groupe Auto Scaling utilise.

```
#!/bin/bash

function get_target_state {
    echo $(curl -s http://169.254.169.254/latest/meta-data/autoscaling/target-
lifecycle-state)
}

function get_instance_id {
    echo $(curl -s http://169.254.169.254/latest/meta-data/instance-id)
}

function complete_lifecycle_action {
    instance_id=$(get_instance_id)
    group_name='TestAutoScalingEvent-group'
    region='us-west-2'

    echo $instance_id
    echo $region
    echo $(aws autoscaling complete-lifecycle-action \
      --lifecycle-hook-name TestAutoScalingEvent-hook \
      --auto-scaling-group-name $group_name \
      --lifecycle-action-result CONTINUE \
      --instance-id $instance_id \
      --region $region)
}

function main {
    while true
    do
        target_state=$(get_target_state)
        if [ \"$target_state\" = \"InService\" ]; then
            # Change hostname
            export new_hostname="${group_name}-$instance_id"
            hostname $new_hostname
```

```
            # Send callback
            complete_lifecycle_action
            break
        fi
        echo $target_state
        sleep 5
    done
}

main
```

Ce simple script de données utilisateur effectue les opérations suivantes :

- Appelle les métadonnées de l'instance pour récupérer l'état du cycle de vie cible et l'ID d'instance à partir des métadonnées de l'instance
- Récupère l'état du cycle de vie cible à plusieurs reprises jusqu'à ce qu'il passe à `InService`
- Modifie le nom d'hôte de l'instance par l'ID d'instance précédé du nom du groupe Auto Scaling, si l'état du cycle de vie cible est `InService`
- Envoie un rappel en appelant la commande CLI complete-lifecycle-action pour signaler à Amazon EC2 Auto Scaling de `CONTINUE` le processus de lancement de l'EC2

10. Choisissez Create launch template (Créer un modèle de lancement).
11. Sur la page de confirmation, choisissez Create Auto Scaling group (Créer un groupe Auto Scaling).

> **Note**
>
> Pour d'autres exemples que vous pouvez utiliser comme référence pour développer votre script de données utilisateur, consultez le GitHub référentiel Amazon EC2 Auto Scaling.

Étape 3 : créer un groupe Auto Scaling

Une fois le modèle de lancement créé, créez un groupe Auto Scaling.

Pour créer un groupe Auto Scaling

1. Dans la page Choose launch template or configuration (Choisir un modèle de lancement ou une configuration), dans Auto Scaling group name (Nom du groupe Auto Scaling), saisissez un nom pour le groupe Auto Scaling (**`TestAutoScalingEvent-group`**).
2. Choisissez Next (Suivant) pour accéder à la page Choose instance launch options (Choisir les options de lancement d'instance).
3. Dans Network (Réseau), choisissez un VPC.
4. Pour Availability Zones and subnets (Zones de disponibilité et sous-réseaux), choisissez un ou plusieurs sous-réseaux dans une ou plusieurs zones de disponibilité.
5. Dans la section Instance type requirements (Exigences relatives au type d'instance), utilisez le paramètre par défaut pour simplifier cette étape. (Ne remplacez pas le modèle de lancement.) Pour ce didacticiel, vous lancerez une seule instance à la demande en utilisant le type d'instance spécifié dans votre modèle de lancement.
6. Choisissez Skip to review (Passez à la révision) en bas de l'écran.
7. Sur la page Review (Vérifier), consultez les paramètres du groupe Auto Scaling, puis choisissez Create Auto Scaling group (Créer un groupe Auto Scaling).

Étape 4 : ajouter un hook de cycle de vie

Ajoutez un hook de cycle de vie pour maintenir l'instance en attente jusqu'à ce que votre action de cycle de vie soit terminée.

Ajouter un hook de cycle de vie

1. Ouvrez la page des groupes Auto Scaling de la console Amazon EC2.
2. Cochez la case située en regard de votre groupe Auto Scaling. Un volet fractionné s'ouvre en bas de la page.
3. Dans le volet inférieur, sous l'onglet Gestion des instances, accédez à Hooks de cycle de vie et choisissez Créer un hook de cycle de vie.
4. Pour définir un hook de cycle de vie pour la montée en puissance (lancement d’instances), procédez comme suit :

    a. Dans le champ Nom du hook de cycle de vie, saisissez **`TestAutoScalingEvent-hook`**.

    b. Dans le champ Lifecycle transition (Transition du cycle de vie), choisissez Instance launch (Lancement d'instance).

c. Pour Heartbeat timeout (Délai de pulsation), saisissez **`300`** pour le nombre de secondes d'attente d'un rappel de votre script de données utilisateur.

d. Dans le champ Default result (Résultat par défaut), choisissez ABANDON (ABANDONNER). Si le hook expire sans recevoir de rappel de votre script de données utilisateur, le groupe Auto Scaling résilie la nouvelle instance.

e. (Facultatif) Conserver Notification metadata (Métadonnées de notification) vide.

5. Choisissez Créer.

Étape 5 : tester et vérifier la fonctionnalité

Pour tester la fonctionnalité, mettez à jour le groupe Auto Scaling en augmentant de 1 la capacité souhaitée du groupe Auto Scaling. Le script de données utilisateur s'exécute et commence à vérifier l'état du cycle de vie cible de l'instance peu après le lancement de l'instance. Le script modifie le nom d'hôte et envoie une action de rappel lorsque l'état du cycle de vie cible est `InService`. Cette opération ne prend généralement que quelques secondes.

Pour augmenter la taille du groupe Auto Scaling

1. Ouvrez la page des groupes Auto Scaling de la console Amazon EC2.
2. Cochez la case située en regard de votre groupe Auto Scaling. Affichez les détails dans un volet inférieur tout en continuant à voir les premières lignes du volet supérieur.
3. Dans le volet inférieur, sous l'onglet Details (Détails), choisissez Group details (Détails du groupe) puis Edit (Modifier).
4. Pour Desired capacity (Capacité désirée), augmentez la valeur actuelle de 1.
5. Choisissez Mettre à jour. Pendant le lancement de l'instance, la colonne Status (Statut) du volet supérieur affiche le statut Mise à jour de la capacité.

Après avoir augmenté la capacité souhaitée, vous pouvez vérifier que votre instance a été lancée avec succès et qu'elle n'est pas résiliée à la description des activités de mise à l'échelle.

Pour afficher l'activité de mise à l'échelle

1. Revenez à la page Groupes Auto Scaling et sélectionnez votre groupe.
2. Dans l'onglet Activité, sous Historique de l'activité, la colonne Statut indique si votre groupe Auto Scaling a réussi à lancer une instance.

3. Si le script de données utilisateur échoue, une fois le délai d'expiration écoulé, vous voyez une activité de mise à l'échelle dont le statut est égal à `Canceled` et un message de statut de `Instance failed to complete user's Lifecycle Action: Lifecycle Action with token e85eb647-4fe0-4909-b341-a6c42EXAMPLE was abandoned: Lifecycle Action Completed with ABANDON Result`.

Étape 6 : Nettoyer

Si vous n'avez plus besoin des ressources que vous avez créées pour ce tutoriel, procédez comme suit pour les supprimer.

Pour supprimer le hook de cycle de vie

1. Ouvrez la page des groupes Auto Scaling de la console Amazon EC2.
2. Cochez la case située en regard de votre groupe Auto Scaling.
3. Sous l'onglet Gestion des instances, accédez à Hooks de cycle de vie et choisissez le hook de cycle de vie (`TestAutoScalingEvent-hook`).
4. Sélectionnez Actions, Delete (Supprimer).
5. Pour confirmer, choisissez de nouveau Delete (Supprimer).

Pour supprimer le modèle de lancement

1. Ouvrez la page des modèles de lancement de la console Amazon EC2.
2. Sélectionnez le modèle de lancement (`TestAutoScalingEvent-template`), puis choisissez Actions, Delete template (Supprimer le modèle).
3. Lorsque vous êtes invité à confirmer l'opération, saisissez **`Delete`** pour confirmer la suppression du modèle de lancement spécifié, puis choisissez Delete (Supprimer).

Si vous avez terminé d'utiliser l'exemple de groupe Auto Scaling, supprimez-le. Vous pouvez également supprimer et la politique d'autorisations et le rôle IAM que vous avez créés.

Pour supprimer le groupe Auto Scaling

1. Ouvrez la page des groupes Auto Scaling de la console Amazon EC2.
2. Cochez la case en regard de votre groupe Auto Scaling (`TestAutoScalingEvent-group`) et choisissez Delete (Supprimer).

3. Lorsque vous êtes invité à confirmer l'opération, saisissez **`delete`** pour confirmer la suppression du groupe Auto Scaling spécifié, puis choisissez Delete (Supprimer).

   Une icône de chargement dans la colonne Name (Nom) indique que le groupe Auto Scaling est en cours de suppression. Quelques minutes sont nécessaires pour résilier les instances et supprimer le groupe.

Suppression du rôle IAM

1. Ouvrez la page Roles (Rôles) de la console IAM.
2. Sélectionnez le rôle de la fonction (`TestAutoScalingEvent-role`).
3. Sélectionnez Delete (Supprimer).
4. Lorsque vous êtes invité à confirmer, saisissez le nom du rôle et choisissez Delete (Supprimer).

Pour supprimer la politique IAM

1. Ouvrez la page Policies (Politiques) de la console IAM.
2. Sélectionnez la politique que vous avez créée (`TestAutoScalingEvent-policy`).
3. Sélectionnez Actions, Supprimer.
4. Lorsque vous êtes invité à confirmer, saisissez le nom de la politique et choisissez Delete (Supprimer).

Ressources connexes

Les rubriques connexes suivantes peuvent être utiles lorsque vous développez du code qui invoque des actions sur les instances en fonction des données disponibles dans les métadonnées de l'instance.

- Récupérer l'état du cycle de vie cible via des métadonnées d'instance. Cette section décrit l'état du cycle de vie pour d'autres cas d'utilisation, tels que la résiliation d'une instance.
- Ajouter des hooks de cycle de vie (console). Cette procédure explique comment ajouter des hooks de cycle de vie pour la montée en puissance (lancement des instances) et la mise à l'échelle horizontale (instances résiliées ou revenant à un groupe chaud).

- Catégories de métadonnées d'instance dans le guide de l'utilisateur Amazon EC2. Cette rubrique répertorie toutes les catégories de métadonnées d'instance que vous pouvez utiliser pour appeler des actions sur les instances EC2.

Pour consulter un didacticiel expliquant comment utiliser Amazon EventBridge pour créer des règles qui invoquent des fonctions Lambda en fonction d'événements survenant dans les instances de votre groupe Auto Scaling, consultez. Didacticiel : configurer un hook de cycle de vie qui appelle une fonction Lambda

Didacticiel : configurer un hook de cycle de vie qui appelle une fonction Lambda

Dans cet exercice, vous allez créer une EventBridge règle Amazon qui inclut un modèle de filtre qui, lorsqu'il est mis en correspondance, invoque une AWS Lambda fonction en tant que cible de la règle. Nous fournissons le modèle de filtre et l'exemple de code de fonction à utiliser.

Si tout est correctement configuré, à la fin de ce didacticiel, la fonction Lambda exécutera une action personnalisée lors du lancement des instances. L'action personnalisée enregistre simplement l'événement dans le flux de CloudWatch log Logs associé à la fonction Lambda.

La fonction Lambda procède également à un rappel pour laisser le cycle de vie de l'instance se poursuivre si cette action est réussie, mais laisse l'instance abandonner le lancement et se résilier si l'action échoue.

L'illustration suivante résume le flux d'un événement de scale-out lorsque vous utilisez une fonction Lambda pour effectuer une action personnalisée. Après le lancement d'une instance, le cycle de vie de l'instance est suspendu jusqu'à ce que le cycle de vie soit terminé, soit en expirant, soit en recevant un signal indiquant à Amazon EC2 Auto Scaling de continuer.

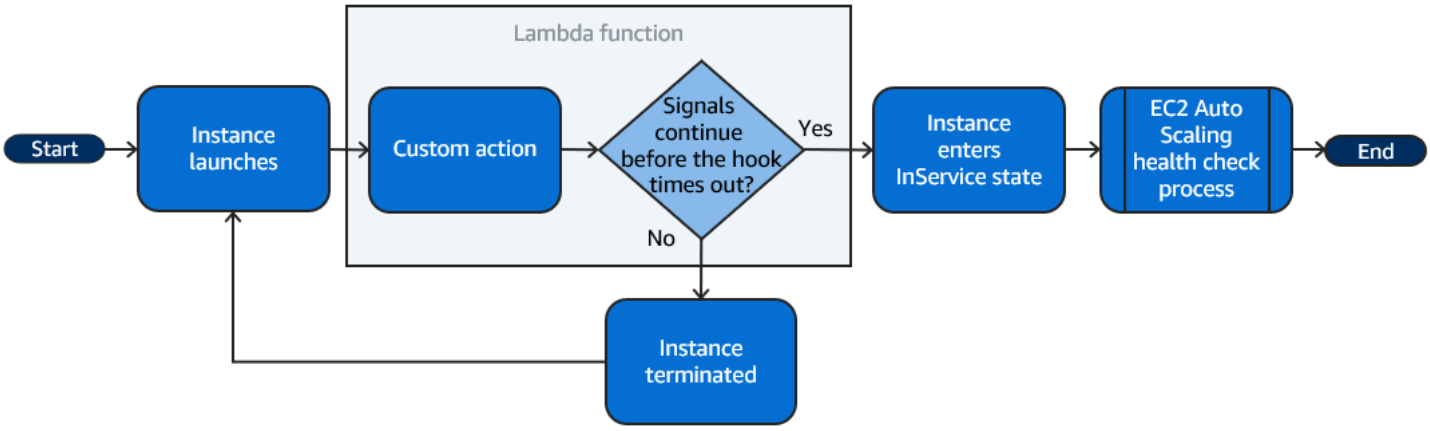


Table des matières

Prérequis

Avant d'entamer ce didacticiel, si vous n'avez pas de groupe Auto Scaling, créez-en un. Pour créer un groupe Auto Scaling, ouvrez la page Groupes Auto Scaling de la console Amazon EC2 et choisissez Créer un groupe Auto Scaling.

Étape 1 : Créer un rôle IAM doté des autorisations nécessaires pour utiliser des actions de cycle de vie

Avant de créer une fonction Lambda, vous devez créer un rôle d'exécution et une politique d'autorisations pour permettre à Lambda d'utiliser des hooks de cycle de vie.

Pour créer la politique

1. Ouvrez la page Policies (Politiques) de la console IAM et sélectionnez Create policy (Créer une politique).
2. Sélectionnez l'onglet JSON.
3. Dans la zone Policy Document (Document relatif à la politique), collez le document suivant, en remplaçant le texte en *italique* par votre numéro de compte et le nom de votre groupe Auto Scaling.

```
{
  "Version": "2012-10-17",
  "Statement": [
```

```
    {
      "Effect": "Allow",
      "Action": [
        "autoscaling:CompleteLifecycleAction"
      ],
      "Resource":
 "arn:aws:autoscaling:*:123456789012:autoScalingGroup:*:autoScalingGroupName/my-
asg"
    }
  ]
}
```

4. Choisissez Suivant.
5. Pour Policy name (Nom de la politique), saisissez **`LogAutoScalingEvent-policy`**. Sélectionnez Create policy (Créer une politique).

Lorsque vous avez créé la politique, vous pouvez créer un rôle qui l'utilise.

Pour créer le rôle

1. Dans le panneau de navigation de gauche, sélectionnez Roles (Rôles).
2. Sélectionnez Create role (Créer un rôle).
3. Pour Select trusted entity (Sélectionner une entité de confiance), choisissez service AWS .
4. Pour votre cas d'utilisation, sélectionnez Lambda, puis Next (Suivant).
5. Sous Ajouter des autorisations, choisissez la politique que vous avez créée (LogAutoScalingEvent-policy) et le nom AWSLambdaBasicExecutionRolede la politique. Ensuite, choisissez Suivant.

> **Note**
>
> La AWSLambdaBasicExecutionRolepolitique dispose des autorisations dont la fonction a besoin pour écrire des CloudWatch journaux dans Logs.

6. Sur la page Name, review, and create (Nommer, vérifier et créer), pour Role name (Nom du rôle), saisissez **`LogAutoScalingEvent-role`**, puis choisissez Create role (Créer un rôle).

Étape 2 : créer une fonction Lambda

Créez une fonction Lambda qui servira de cible pour les événements. L'exemple de fonction Lambda, écrit dans Node.js, est invoqué EventBridge lorsqu'un événement correspondant est émis par Amazon EC2 Auto Scaling.

Pour créer une fonction Lambda

1. Ouvrez la page Functions (Fonctions) sur la console Lambda.
2. Sélectionnez Create function (Créer une fonction), puis Author from scratch (Créer à partir de zéro).
3. Sous Basic information (Informations de base), pour Function name (Nom de la fonction), entrez **`LogAutoScalingEvent`**.
4. Pour Exécution, choisissez Node.js 18.x.
5. Faites défiler l'écran vers le bas et choisissez Modifier le rôle d'exécution par défaut, puis dans le champ Rôle d'exécution, choisissez Utiliser un rôle existant.
6. Pour Rôle existant, choisissez LogAutoScalingEvent-role.
7. Laissez les autres valeurs par défaut.
8. Sélectionnez Create function (Créer une fonction). Vous retournez au code et à la configuration de la fonction.
9. Vérifiez que votre fonction `LogAutoScalingEvent` est toujours ouverte dans la console, puis sous Code source, dans l'éditeur, copiez l'exemple de code suivant dans le fichier index.mjs.

```
import { AutoScalingClient, CompleteLifecycleActionCommand } from "@aws-sdk/client-
auto-scaling";
export const handler = async(event) => {
  console.log('LogAutoScalingEvent');
  console.log('Received event:', JSON.stringify(event, null, 2));
  var autoscaling = new AutoScalingClient({ region: event.region });
  var eventDetail = event.detail;
  var params = {
    AutoScalingGroupName: eventDetail['AutoScalingGroupName'], /* required */
    LifecycleActionResult: 'CONTINUE', /* required */
    LifecycleHookName: eventDetail['LifecycleHookName'], /* required */
    InstanceId: eventDetail['EC2InstanceId'],
    LifecycleActionToken: eventDetail['LifecycleActionToken']
  };
  var response;
```

```
  const command = new CompleteLifecycleActionCommand(params);
  try {
    var data = await autoscaling.send(command);
    console.log(data); // successful response
    response = {
      statusCode: 200,
      body: JSON.stringify('SUCCESS'),
    };
  } catch (err) {
    console.log(err, err.stack); // an error occurred
    response = {
      statusCode: 500,
      body: JSON.stringify('ERROR'),
    };
  }
  return response;
};
```

Ce code enregistre simplement l'événement afin qu'à la fin de ce didacticiel, vous puissiez voir un événement apparaître dans le flux de journal CloudWatch des journaux associé à cette fonction Lambda.

10. Choisissez Deploy (Déployer).

Étape 3 : Création d'une EventBridge règle

Créez une EventBridge règle pour exécuter votre fonction Lambda. Pour plus d'informations sur l'utilisation EventBridge, consultez EventBridge À utiliser pour gérer les événements Auto Scaling.

Pour créer une règle avec la console

1. Ouvrez la EventBridge console.
2. Dans le volet de navigation, choisissez Règles.
3. Choisissez Créer une règle.
4. Pour Define rule detail (Définir les détails de la règle), procédez comme suit :

   a. Pour Name (Nom), saisissez **`LogAutoScalingEvent-rule`**.

   b. Pour Event bus (Bus d'événement), choisissez default (défaut). Lorsqu'un événement est généré Service AWS dans votre compte, il est toujours redirigé vers le bus d'événements par défaut de votre compte.

c. Pour Type de règle, choisissez Règle avec un modèle d’événement.

d. Choisissez Suivant.

5. Pour Build event pattern (Créer un modèle d’événement), procédez comme suit :

a. Dans Source de l'événement, choisissez AWS des événements ou des événements EventBridge partenaires.

b. Faites défiler vers le bas jusqu’à Modèle d’événements, puis procédez comme suit :

c. i. Pour Event source (Source d’événement), choisissez Services AWS.

ii. Pour Service AWS, choisissez Auto Scaling.

iii. Dans Event type (Type d'événement), choisissez Instance Launch and Terminate (Lancement et résiliation d'une instance).

iv. Par défaut, la règle correspond à tout événement de mise à l'échelle horizontale ou de montée en puissance. Pour créer une règle qui vous avertit lorsqu'un événement de montée en puissance se produit et qu'une instance est dans un état d'attente en raison d'un hook de cycle de vie, choisissez Specific instance event(s) (Événement[s] d'instance spécifique[s]) et sélectionnez EC2 Instance-launch Lifecycle Action (Action du cycle de vie de l'instance EC2 : lancement).

v. Par défaut, la règle correspond à tout groupe Auto Scaling de la région. Pour que la règle corresponde à un groupe Auto Scaling spécifique, choisissez Nom(s) de groupe spécifique(s), puis sélectionnez le groupe.

vi. Choisissez Next (Suivant).

6. Pour Select target(s) (Sélectionner la ou les cibles), procédez comme suit :

a. Pour Target types (Types de cibles), choisissez Service AWS.

b. Pour Select a target (Sélectionner une cible), choisissez Lambda Function (Fonction Lambda).

c. Pour Fonction, choisissez LogAutoScalingEvent.

d. Choisissez Next (Suivant) deux fois.

7. Sur la page Vérifier et créer, choisissez Créer une règle.

Étape 4 : ajouter un hook de cycle de vie

Dans cette section, vous allez ajouter un hook de cycle de vie afin que Lambda exécute votre fonction sur les instances au lancement.

Ajouter un hook de cycle de vie

1. Ouvrez la page des groupes Auto Scaling de la console Amazon EC2.
2. Cochez la case située en regard de votre groupe Auto Scaling. Un volet fractionné s'ouvre en bas de la page.
3. Dans le volet inférieur, sous l'onglet Gestion des instances, accédez à Hooks de cycle de vie et choisissez Créer un hook de cycle de vie.
4. Pour définir un hook de cycle de vie pour la montée en puissance (lancement d'instances), procédez comme suit :
   a. Dans le champ Nom du hook de cycle de vie, saisissez **`LogAutoScalingEvent-hook`**.
   b. Dans le champ Transition du cycle de vie, choisissez Lancement d'instance.
   c. Dans Délai de pulsation, saisissez **`300`** pour définir le nombre de secondes d'attente d'un rappel de votre fonction Lambda.
   d. Dans le champ Résultat par défaut, choisissez ABANDONNER. Cela signifie que le groupe Auto Scaling résiliera une nouvelle instance si le hook expire sans recevoir de rappel de votre fonction Lambda.
   e. (Facultatif) Laissez le champ Métadonnées de notification vide. Les données d'événement que nous transmettons EventBridge contiennent toutes les informations nécessaires pour appeler la fonction Lambda.
5. Choisissez Créer.

Étape 5 : tester et vérifier l'événement

Pour tester l'événement, mettez à jour le groupe Auto Scaling en augmentant de 1 la capacité souhaitée du groupe Auto Scaling. Votre fonction Lambda est appelée quelques secondes après l'augmentation de la capacité souhaitée.

Pour augmenter la taille du groupe Auto Scaling

1. Ouvrez la page des groupes Auto Scaling de la console Amazon EC2.
2. Cochez la case située en regard de votre groupe Auto Scaling pour afficher les détails dans le volet inférieur tout en continuant à voir les premières lignes du volet supérieur.
3. Dans le volet inférieur, sous l'onglet Détails, choisissez Détails du groupe puis Modifier.
4. Pour Desired capacity (Capacité désirée), augmentez la valeur actuelle de 1.

5. Choisissez Mettre à jour. Pendant le lancement de l'instance, la colonne Statut du volet supérieur affiche le statut Mise à jour de la capacité.

Après avoir augmenté la capacité souhaitée, vous pouvez vérifier que votre fonction Lambda a été appelée.

Pour afficher la sortie de votre fonction Lambda

1. Ouvrez la page Groupes de journaux de la CloudWatch console.
2. Sélectionnez le nom du groupe de journaux pour votre fonction Lambda (`/aws/lambda/LogAutoScalingEvent`).
3. Sélectionnez le nom du flux de journaux pour afficher les données fournies par la fonction pour l'action de cycle de vie.

Vous pouvez ensuite vérifier que votre instance a été lancée avec succès à partir de la description des activités de mise à l'échelle.

Pour afficher l'activité de mise à l'échelle

1. Revenez à la page Groupes Auto Scaling et sélectionnez votre groupe.
2. Dans l'onglet Activité, sous Historique de l'activité, la colonne Statut indique si votre groupe Auto Scaling a réussi à lancer une instance.
   - Si l'action a abouti, le statut de l'activité de mise à l'échelle indique « Succès ».
   - Si elle a échoué, après quelques minutes d'attente, vous verrez une activité de mise à l'échelle accompagnée du statut « Annulée » et du message de statut « L'instance n'a pas réussi à exécuter l'action de cycle de vie de l'utilisateur : l'action de cycle de vie associée au jeton e85eb647-4fe0-4909-b341-a6c42EXAMPLE a été abandonnée : Action de cycle de vie terminée avec le résultat ABANDONNER ».

Pour réduire la taille du groupe Auto Scaling

Si vous n'avez pas besoin de l'instance supplémentaire que vous avez lancée pour ce test, vous pouvez ouvrir l'onglet Détails et réduire la Capacité souhaitée de 1.

Étape 6 : Nettoyer

Si vous n'avez plus besoin des ressources que vous avez créées pour ce didacticiel, procédez comme suit pour les supprimer.

Pour supprimer le hook de cycle de vie

1. Ouvrez la page des groupes Auto Scaling de la console Amazon EC2.
2. Cochez la case située en regard de votre groupe Auto Scaling.
3. Sous l'onglet Gestion des instances, accédez à Hooks de cycle de vie et choisissez le hook de cycle de vie (`LogAutoScalingEvent-hook`).
4. Sélectionnez Actions, Delete (Supprimer).
5. Pour confirmer, choisissez de nouveau Delete (Supprimer).

Pour supprimer la EventBridge règle Amazon

1. Ouvrez la page Règles dans la EventBridge console Amazon.
2. Sous Bus d'événement, choisissez le bus d'événement associé à la règle (`Default`).
3. Ensuite, activez la case à cocher en regard de votre règle (`LogAutoScalingEvent-rule`).
4. Sélectionnez Delete (Supprimer).
5. Lorsque vous êtes invité à confirmer, saisissez le nom de la règle et choisissez Delete (Supprimer).

Si vous avez terminé d'utiliser l'exemple de fonction, supprimez-le. Vous pouvez également supprimer le groupe de journaux qui stocke les journaux de la fonction, ainsi que le rôle d'exécution et la politique d'autorisations que vous avez créés.

Pour supprimer une fonction Lambda

1. Ouvrez la page Functions (Fonctions) sur la console Lambda.
2. Choisissez la fonction (`LogAutoScalingEvent`).
3. Sélectionnez Actions, Supprimer.
4. Lorsque vous êtes invité à confirmer, saisissez **`delete`** pour confirmer la suppression de la fonction spécifiée, puis choisissez Delete (Supprimer).

Pour supprimer le groupe de journaux

1. Ouvrez la page Groupes de journaux de la CloudWatch console.
2. Sélectionnez le groupe de journaux de la fonction (`/aws/lambda/LogAutoScalingEvent`).
3. Sélectionnez Actions, Delete log group(s) (Supprimer le ou les groupes de journaux).
4. Dans la boîte de dialogue Delete log group(s) (Supprimer le ou les groupes de journaux), sélectionnez Delete (Supprimer).

Pour supprimer le rôle d'exécution

1. Ouvrez la page Roles (Rôles) de la console IAM.
2. Sélectionnez le rôle de la fonction (`LogAutoScalingEvent-role`).
3. Sélectionnez Delete (Supprimer).
4. Lorsque vous êtes invité à confirmer, saisissez le nom du rôle et choisissez Delete (Supprimer).

Pour supprimer la politique IAM

1. Ouvrez la page Policies (Politiques) de la console IAM.
2. Sélectionnez la politique que vous avez créée (`LogAutoScalingEvent-policy`).
3. Sélectionnez Actions, Supprimer.
4. Lorsque vous êtes invité à confirmer, saisissez le nom de la politique et choisissez Delete (Supprimer).

Ressources connexes

Les rubriques connexes suivantes peuvent être utiles lorsque vous créez des EventBridge règles basées sur des événements qui se produisent dans les instances de votre groupe Auto Scaling.

- EventBridge À utiliser pour gérer les événements Auto Scaling. Cette section présente des exemples d'événements pour d'autres cas d'utilisation, y compris des événements destinés à la mise à l'échelle horizontale.
- Ajouter des hooks de cycle de vie (console). Cette procédure explique comment ajouter des hooks de cycle de vie pour la montée en puissance (lancement des instances) et la mise à l'échelle horizontale (instances résiliées ou revenant à un groupe chaud).

Pour suivre un didacticiel qui explique comment utiliser le service de métadonnées d'instance (IMDS) afin d'invoquer une action depuis l'instance elle-même, consultez Tutoriel : configurer les données utilisateur pour récupérer l'état du cycle de vie cible via des métadonnées d'instance.

Groupes d'instances pré-initialisées pour Amazon EC2 Auto Scaling

Un groupe d'instances pré-initialisées vous permet de réduire la latence de vos applications dont les temps de démarrage sont exceptionnellement longs, du fait par exemple que les instances doivent écrire des quantités massives de données sur disque. Avec les groupes d'instances pré-initialisées, il n'est plus nécessaire de surprovisionner vos groupes Auto Scaling pour réduire la latence et améliorer ainsi les performances des applications. Pour plus d'informations, consultez l'article de blog Scaling your applications faster with EC2 Auto Scaling Warm Pools.

> **Important**
>
> La création d'un groupe d'instances pré-initialisées quand elle n'est pas nécessaire risque d'entraîner des coûts inutiles. Si le temps de démarrage initial n'entraîne pas de problèmes de latence notables pour votre application, vous n'avez probablement pas besoin d'utiliser ce type de groupe.

Rubriques

Concepts de base

Avant de commencer, familiarisez-vous avec les concepts de base ci-dessous :

Pools d'instances pré-initialisées

Ce type de groupe d'instances EC2 pré-initialisées réside à côté d'un groupe Auto Scaling. Chaque fois que votre application doit monter en puissance, le groupe Auto Scaling peut s'appuyer sur le groupe d'instances pré-initialisées pour atteindre la nouvelle capacité souhaitée. Cela vous permet de vous assurer que les instances sont rapidement disponibles pour gérer le trafic des applications, ce qui accélère la réponse à un événement de montée en puissance. Lorsque des instances sont retirées du groupe d'instances pré-initialisées, elles sont comptabilisées dans la capacité souhaitée du groupe. Il s'agit d'une opération de type démarrage à chaud.

Lorsque les instances sont dans le groupe chaud, vos politiques de mise à l'échelle ne montent en puissance que si la valeur métrique des instances qui sont dans l'état `InService` est supérieure au seuil d'alarme supérieur de la politique de mise à l'échelle (qui est le même que l'utilisation cible d'une politique de suivi des objectifs et d'échelonnement).

Taille d'un groupe d'instances pré-initialisées

Par défaut, la taille d'un groupe d'instances pré-initialisées correspond à la différence entre la capacité maximale du groupe Auto Scaling et sa capacité souhaitée. Par exemple, si la capacité souhaitée de votre groupe Auto Scaling est de 6 et que la capacité maximale correspond à 10, la taille de votre groupe d'instances pré-initialisées est donc de 4 lorsque vous configurez le groupe pour la première fois et qu'il est initialisé.

Pour spécifier séparément la capacité maximale du pool de chaleur, utilisez l'option custom specification (`MaxGroupPreparedCapacity`) et définissez une valeur personnalisée supérieure à la capacité actuelle du groupe. Si vous fournissez une valeur personnalisée, la taille du pool de chaleur est calculée comme la différence entre la valeur personnalisée et la capacité actuelle souhaitée du groupe. Par exemple, si la capacité souhaitée de votre groupe Auto Scaling est de 6, si la capacité maximale est de 20 et si la valeur personnalisée est de 8, la taille de votre pool de chaleur sera de 2 lorsque vous le configurez pour la première fois et que le pool sera initialisé.

Il se peut que vous n'ayez besoin d'utiliser l'option custom specification (`MaxGroupPreparedCapacity`) que lorsque vous travaillez avec de grands groupes Auto Scaling afin de gérer les avantages financiers liés à la mise en place d'un pool de chaleur. Par

exemple, un groupe Auto Scaling avec 1 000 instances, une capacité maximale de 1 500 (pour fournir une capacité supplémentaire en cas de pics de trafic d'urgence) et un groupe d'instances pré-initialisées de 100 instances peut vous aider à atteindre vos objectifs mieux que de garder 500 instances réservées pour une utilisation future dans le groupe d'instances pré-initialisées.

Taille minimale du groupe d'instances pré-initialisées.

Envisagez d'utiliser le paramètre de taille minimale pour définir de manière statique le nombre minimal d'instances à conserver dans le groupe. Aucune taille minimale n'est définie par défaut.

État des instances de groupe d'instances pré-initialisées

Vous pouvez conserver des instances dans le groupe d'instances pré-initialisées dans l'un des trois états suivants : `Stopped`, `Running` ou `Hibernated`. Conserver les instances dans l'état `Stopped` est un moyen efficace de limiter les coûts. Lorsque les instances sont interrompues, vous ne payez que pour les volumes utilisés et les adresses IP élastiques attachées aux instances.

Vous pouvez également conserver les instances dans un état `Hibernated` pour arrêter les instances sans supprimer leur contenu de mémoire (RAM). Lorsqu'une instance est mise en veille prolongée, cela signale au système d'exploitation qu'il doit enregistrer le contenu de votre RAM sur votre volume racine Amazon EBS. Lorsque l'instance est redémarrée, le volume racine est restauré à son état précédent et le contenu de la RAM est rechargé. Pendant que les instances sont en veille prolongée, vous ne payez que pour les volumes EBS, y compris le stockage du contenu RAM, et les adresses IP Elastic attachées aux instances.

Garder des instances dans un état `Running` à l'intérieur du groupe d'instances pré-initialisées est également possible, mais est fortement déconseillé pour éviter d'encourir des frais inutiles. Lorsque les instances sont arrêtées ou mises en veille prolongée, vous économisez le coût des instances elles-mêmes. Vous payez les instances uniquement lorsqu'elles sont en cours d'exécution.

Hooks de cycle de vie

Les hooks de cycle de vie vous permettent de mettre des instances dans un état d’attente afin que vous puissiez effectuer des actions personnalisées sur les instances. Les actions personnalisées sont exécutées au lancement des instances ou avant qu'elles ne se terminent.

Dans une configuration de groupe chaud, les hooks de cycle de vie retardent l’arrêt ou la mise en veille prolongée des instances et leur mise en service pendant un événement de montée en puissance jusqu’à ce que leur initialisation soit terminée. Si vous ajoutez un groupe d'instances

pré-initialisées à votre groupe Auto Scaling sans hook de cycle de vie, les instances dont l'initialisation prend beaucoup de temps peuvent être arrêtées ou mises en veille prolongée, puis mises en service pendant un événement de montée en puissance avant qu'elles ne soient prêtes.

Politique de réutilisation d'instance

Par défaut, Amazon EC2 Auto Scaling résilie vos instances lors de la mise à l'échelle horizontale de votre groupe Auto Scaling. Ensuite, il lance de nouvelles instances dans le groupe d'instances pré-initialisées pour remplacer celles qui ont été résiliées.

Si vous souhaitez plutôt renvoyer des instances vers le groupe d'instances pré-initialisées, vous pouvez spécifier une politique de réutilisation d'instance. Cela vous permet de réutiliser des instances déjà configurées pour servir le trafic des applications. Pour s'assurer que votre groupe d'instances pré-initialisées n'est pas surapprovisionné, Amazon EC2 Auto Scaling peut résilier des instances dans le groupe d'instances pré-initialisées pour réduire sa taille lorsque celle-ci est plus grande que nécessaire en fonction de ses paramètres. Lors de la résiliation d'instances dans le groupe d'instances pré-initialisées, il utilise la politique de résiliation par défaut pour choisir les instances à résilier en premier.

> **Important**
>
> Si vous souhaitez mettre en veille prolongée des instances mises à l'échelle horizontale et que le groupe Auto Scaling contient déjà des instances, celles-ci doivent répondre aux exigences de la mise en veille prolongée des instances. Si ce n'est pas le cas, lorsque les instances retourneront dans le groupe d'instances pré-initialisées, elles seront arrêtées au lieu d'être mises en veille prolongée.

> **Note**
>
> Actuellement, vous ne pouvez spécifier une politique de réutilisation d'instance qu'à l'aide de la AWS CLI ou d'un kit SDK. Cette fonction n'est pas disponible depuis la console.

Prérequis

Avant de créer un groupe chaud pour votre groupe Auto Scaling, déterminez comment vous allez utiliser les hooks de cycle de vie pour initialiser de nouvelles instances avec un état initial approprié.

Pour effectuer des actions personnalisées sur des instances alors qu'elles sont en état d'attente à cause d'un hook de cycle de vie, deux options s'offrent à vous :

- Pour les scénarios simples où vous souhaitez exécuter des commandes sur vos instances au lancement, vous pouvez inclure un script de données utilisateur lorsque vous créez un modèle de lancement ou une configuration de lancement pour votre groupe Auto Scaling. Les scripts de données utilisateur ne sont que des scripts shell standards ou des directives cloud-init exécutées par cloud-init au démarrage de vos instances. Le script peut également contrôler le moment où vos instances passent à l'état suivant en utilisant l'ID de l'instance sur laquelle il s'exécute. Si ce n'est pas le cas, mettez à jour le script pour récupérer l'ID de l'instance à partir de ses métadonnées. Pour plus d'informations, consultez la section Récupérer les métadonnées d'une instance dans le guide de l'utilisateur Amazon EC2.

  > **Tip**
  >
  > Pour exécuter des scripts de données utilisateur lors du redémarrage d'une instance, les données utilisateur doivent être au format MIME en plusieurs parties et spécifier les éléments suivants dans la section `#cloud-config` des données utilisateur :
  >
  > ```
  > #cloud-config
  > cloud_final_modules:
  >  - [scripts-user, always]
  > ```

- Pour les scénarios avancés dans lesquels vous avez besoin d'un service, par exemple AWS Lambda pour agir lorsque des instances entrent ou sortent du pool de chaleur, vous pouvez créer un lien de cycle de vie pour votre groupe Auto Scaling et configurer le service cible pour effectuer des actions personnalisées en fonction des notifications relatives au cycle de vie. Pour plus d'informations, consultez Cibles de notification prises en charge.

Préparer les instances à la mise en veille prolongée

Pour préparer les instances Auto Scaling à utiliser l'état du `Hibernated` pool, créez un nouveau modèle de lancement ou une nouvelle configuration de lancement correctement configuré pour prendre en charge l'hibernation des instances, comme décrit dans la rubrique Conditions préalables à l'hibernation du guide de l'utilisateur Amazon EC2. Ensuite, associez le nouveau modèle de lancement ou la nouvelle configuration de lancement au groupe Auto Scaling et lancez une actualisation d'instance pour remplacer les instances associées à un précédent modèle de lancement

ou configuration de lancement. Pour plus d'informations, consultez Utiliser une actualisation d'instance pour mettre à jour les instances d'un groupe Auto Scaling.

Mise à jour les instances d'un groupe chaud

Pour mettre à jour les instances d'un groupe chaud, vous devez créer un nouveau modèle de lancement ou une nouvelle configuration de lancement et l'associer au groupe Auto Scaling. Toutes les nouvelles instances sont lancées à l'aide de la nouvelle AMI et d'autres mises à jour spécifiées dans le modèle de lancement ou la configuration de lancement, mais les instances existantes ne sont pas affectées.

Pour forcer le lancement d'instances de groupe chaud de remplacement qui utilisent le nouveau modèle de lancement ou la nouvelle configuration de lancement, vous pouvez lancer une actualisation de l'instance pour effectuer une mise à jour progressive de votre groupe. Une actualisation d'instance remplace d'abord les instances `InService`. Elle remplace ensuite les instances du groupe d'instances pré-initialisées. Pour plus d'informations, consultez Utiliser une actualisation d'instance pour mettre à jour les instances d'un groupe Auto Scaling.

Ressources connexes

Vous pouvez consulter notre GitHubréférentiel pour obtenir des exemples de crochets de cycle de vie pour les piscines chaudes.

Limites

- Vous ne pouvez pas ajouter un pool de chaleur à un groupe Auto Scaling doté d'une politique d'instances mixtes. Vous ne pouvez pas non plus ajouter un warm pool à un groupe Auto Scaling doté d'un modèle de lancement ou d'une configuration de lancement qui demande des instances Spot.
- Amazon EC2 Auto Scaling peut placer une instance dans l'état `Stopped` ou `Hibernated` que si elle est dotée d'un volume Amazon EBS comme périphérique racine. Les instances qui utilisent des stockages d'instance pour le périphérique racine ne peuvent pas être arrêtées ou mises en veille prolongée.
- Amazon EC2 Auto Scaling ne peut mettre une instance dans un *`Hibernated`* état que si elle répond à toutes les exigences répertoriées dans la rubrique relative aux conditions préalables à l'hibernation du guide de l'utilisateur Amazon EC2.
- Si votre groupe d'instances pré-initialisées est épuisé lors d'un événement de montée en puissance, les instances sont lancées directement dans le groupe Auto Scaling (démarrage à

froid). Le démarrage à froid se produit également en cas d'épuisement de la capacité d'une zone de disponibilité.

- Si une instance du warm pool rencontre un problème pendant le processus de lancement, l'empêchant d'atteindre `InService` cet état, l'instance sera considérée comme un échec de lancement et sera interrompue. Cela s'applique quelle que soit la cause sous-jacente, telle qu'une erreur de capacité insuffisante ou tout autre facteur.
- Si vous essayez d'utiliser un pool d'instances pré-initialisées avec Amazon Elastic Kubernetes Service (Amazon EKS), les instances qui sont toujours en cours d'initialisation peuvent s'enregistrer auprès de votre cluster Amazon EKS. Par conséquent, le cluster peut planifier des tâches sur une instance alors qu'il se prépare à être arrêté ou mis en veille prolongée.
- De même, si vous essayez d'utiliser un groupe d'instances pré-initialisées avec un cluster Amazon ECS, les instances peuvent s'enregistrer auprès du cluster avant la fin de leur initialisation. Pour résoudre ce problème, vous devez configurer un modèle de lancement ou une configuration de lancement qui inclut une variable de configuration d'agent spéciale dans les données utilisateur. Pour plus d'informations, consultez Utilisation d'un groupe d'instances pré-initialisées pour votre groupe Auto Scaling dans le Guide du développeur Amazon Elastic Container Service.
- La prise en charge de l'hibernation pour les pools chauds est disponible dans toutes les zones commerciales Régions AWS où Amazon EC2 Auto Scaling et Hibernation sont disponibles, à l'exception des suivantes :
  - Asie-Pacifique (Hyderabad)
  - Asie-Pacifique (Melbourne)
  - Canada Ouest (Calgary)
  - Région Chine (Beijing)
  - Région Chine (Ningxia)
  - Europe (Espagne)
  - Israël (Tel Aviv)

Utiliser des hooks de cycle de vie avec un groupe d'instances pré-initialisées

Les instances du groupe d'instances pré-initialisées gèrent leur propre cycle de vie indépendant, ce qui vous permet de créer l'action personnalisée appropriée pour chaque transition. Ce cycle de vie est conçu pour vous aider à appeler des actions dans un service cible (par exemple, une fonction

Lambda) pendant qu'une instance est encore en cours d'initialisation et avant qu'elle ne soit mise en service.

> **Note**
>
> Les opérations d'API que vous utilisez pour ajouter et gérer des hooks de cycle de vie et des actions de cycle de vie complètes ne sont pas modifiées. Seul le cycle de vie de l'instance est modifié.

Pour plus d'informations sur l'ajout d'un hook de cycle de vie, consultez Ajouter des hooks de cycle de vie. Pour plus d'informations sur l'exécution d'une action de cycle de vie, consultez Effectuer une action de cycle de vie.

Pour les instances entrant dans le groupe d'instances pré-initialisées, vous aurez peut-être besoin d'un hook de cycle de vie pour l'une des raisons suivantes :

- Vous souhaitez lancer des instances EC2 à partir d'une AMI dont l'initialisation est très longue.
- Vous souhaitez exécuter des scripts de données utilisateur pour l'amorçage des instances EC2.

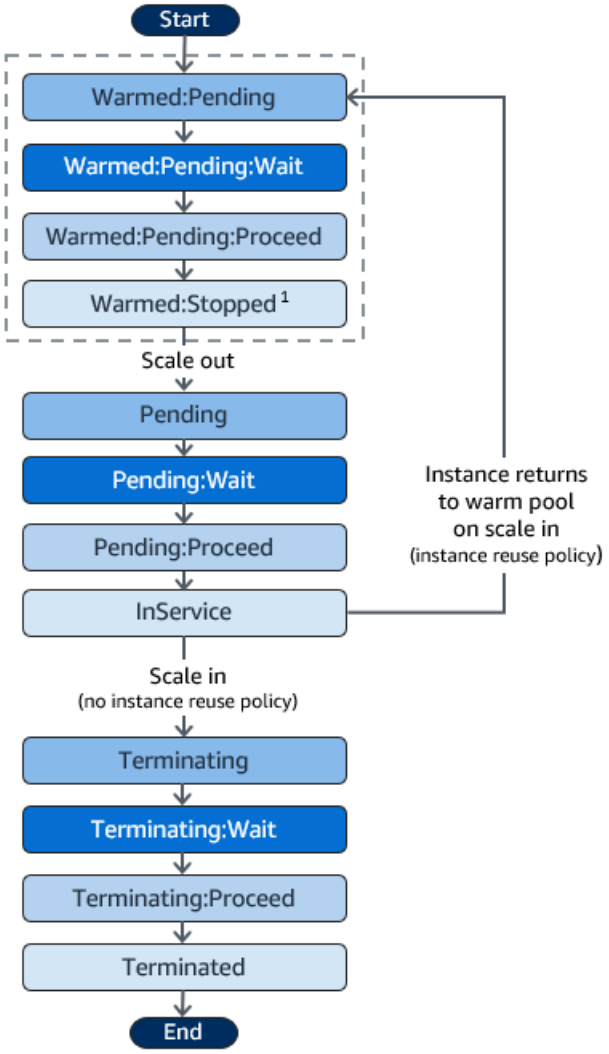
Pour les instances quittant le groupe d'instances pré-initialisées, vous aurez peut-être besoin d'un hook de cycle de vie pour l'une des raisons suivantes :

- Vous pouvez utiliser un peu plus de temps pour préparer les instances EC2 à utiliser. Par exemple, certains de vos services pourraient devoir démarrer lorsqu'une instance redémarre pour que votre application puisse fonctionner correctement.
- Vous souhaitez pré-remplir les données du cache pour vous assurer qu'un nouveau serveur n'est pas lancé avec un cache vide.
- Vous souhaitez enregistrer de nouvelles instances en tant qu'instances gérées auprès de votre service de gestion de la configuration.

Transitions de l'état du cycle de vie pour les instances dans un groupe d'instances pré-initialisées

Une instance Auto Scaling peut passer par plusieurs états dans le cadre de son cycle de vie.

Le schéma suivant montre la transition entre les états Auto Scaling lorsque vous utilisez un groupe d'instances pré-initialisées :



[1] Cet état varie en fonction du paramètre d'état du groupe d'instances pré-initialisées. Si l'état du groupe est défini sur `Running`, alors cet état est à la place `Warmed:Running`. Si l'état du groupe est défini sur `Hibernated`, alors cet état est à la place `Warmed:Hibernated`.

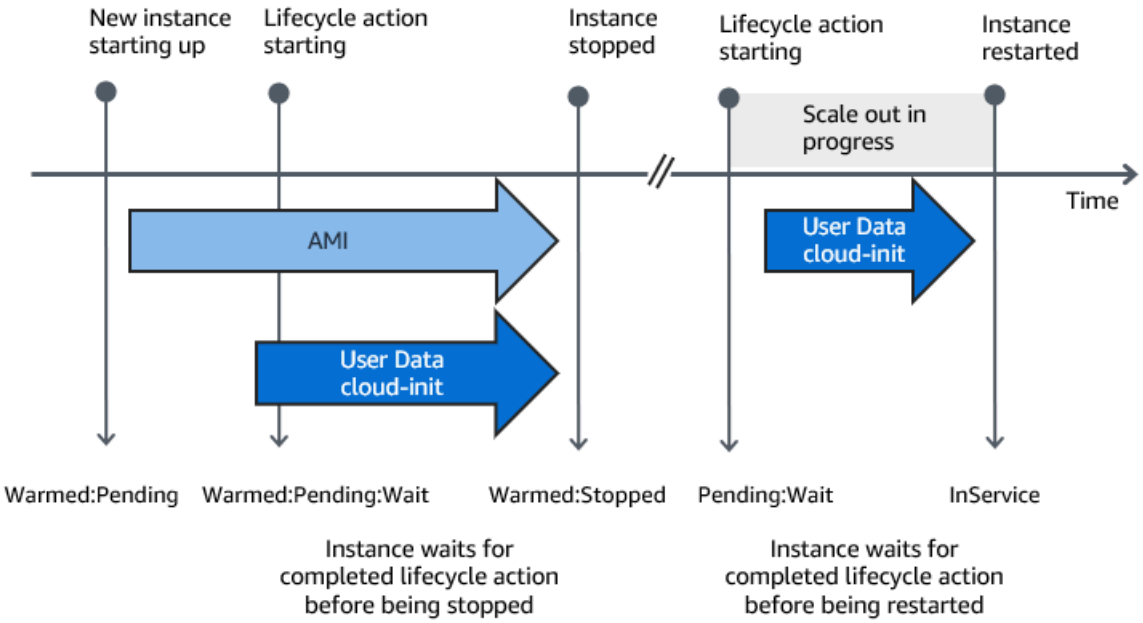
Lorsque vous ajoutez des hooks de cycle de vie, tenez compte des éléments suivants :

- Lorsqu'un hook de cycle de vie est configuré pour l'action `autoscaling:EC2_INSTANCE_LAUNCHING` du cycle de vie, une instance nouvellement lancée s'arrête d'abord pour effectuer une action personnalisée lorsqu'elle atteint l'état `Warmed:Pending:Wait`, puis de nouveau lorsque l'instance redémarre et atteint l'état `Pending:Wait`.
- Lorsqu'un hook de cycle de vie est configuré pour l'action `EC2_INSTANCE_TERMINATING` du cycle de vie, une instance en cours de résiliation s'arrête pour effectuer une action personnalisée lorsqu'elle atteint l'état `Terminating:Wait`. Toutefois, si vous spécifiez une politique de

réutilisation des instances pour renvoyer les instances dans le groupe d'instances pré-initialisées à grande échelle au lieu de les arrêter, une instance qui y revient s'arrête pour effectuer une action personnalisée à l'état `EC2_INSTANCE_TERMINATING` pour l'action de cycle de vie `Warmed:Pending:Wait`.

- Si la demande de votre application vide le groupe d'instances pré-initialisées, Amazon EC2 Auto Scaling peut lancer des instances directement dans le groupe Auto Scaling tant que le groupe n'atteint pas encore sa capacité maximale. Si les instances se lancent directement dans le groupe, elles ne sont mises en attente pour effectuer une action personnalisée à l'état `Pending:Wait`.
- Pour contrôler la durée pendant laquelle une instance reste en état d'attente avant de passer à l'état suivant, configurez votre action personnalisée de manière à utiliser la commande complete-lifecycle-action. Avec les hooks de cycle de vie, les instances restent en attente soit jusqu'à ce que vous signaliez à Amazon EC2 Auto Scaling que l'action du cycle de vie est terminée, soit jusqu'à ce que le délai d'attente (défini par défaut sur une heure) soit écoulé.

Ce qui suit résume le processus d'un événement de montée en puissance.



Lorsque les instances atteignent un état d'attente, Amazon EC2 Auto Scaling envoie une notification. Des exemples de ces notifications sont disponibles dans la EventBridge section de ce guide. Pour plus d'informations, consultez Exemples d'événements et de modèles de groupe chaud.

Cibles de notification prises en charge

Amazon EC2 Auto Scaling prend en charge la définition de l'un des éléments suivants en tant que cibles de notification pour les notifications de cycle de vie :

- EventBridge règles
- Rubriques Amazon SNS
- Files d'attente Amazon SQS

> **Important**
>
> N'oubliez pas que si vous avez un script (cloud-init) de données utilisateur dans votre modèle de lancement ou configuration de lancement qui configure vos instances lors de leur lancement, vous n'avez pas besoin de notifications pour effectuer des actions personnalisées sur vos instances qui sont lancées ou relancées.

Les sections suivantes contiennent des liens vers la documentation décrivant comment configurer les cibles de notification :

EventBridge règles : pour exécuter du code lorsqu'Amazon EC2 Auto Scaling met une instance en état d'attente, vous pouvez créer une EventBridge règle et spécifier une fonction Lambda comme cible. Pour appeler différentes fonctions Lambda en fonction de différentes notifications de cycle de vie, vous pouvez créer plusieurs règles et associer chaque règle à un modèle d'événement et à une fonction Lambda spécifiques. Pour plus d’informations, consultez Créez des EventBridge règles pour les événements en piscine chaude.

Rubriques Amazon SNS : pour recevoir une notification lorsqu'une instance est mise en état d'attente, vous créez une rubrique Amazon SNS, puis configurez le filtrage des messages Amazon SNS pour fournir des notifications de cycle de vie différemment en fonction d'un attribut de message. Pour plus d’informations, consultez Recevoir des notifications à l'aide d'Amazon SNS.

Files d'attente Amazon SQS : pour configurer un point de livraison pour les notifications de cycle de vie où un consommateur pertinent peut les récupérer et les traiter, vous pouvez créer une file d'attente Amazon SQS et un consommateur de file d'attente qui traite les messages de la file d'attente SQS. Si vous souhaitez que le consommateur de file d'attente traite différemment les notifications de cycle de vie en fonction d'un attribut de message, vous devez également configurer le consommateur de file d'attente pour analyser le message, puis agir sur le message lorsqu'un

attribut spécifique correspond à la valeur souhaitée. Pour plus d'informations, consultez Recevoir des notifications à l'aide d'Amazon SQS.

Créez un groupe chaud pour un groupe Auto Scaling

Cette rubrique explique comment créer un groupe chaud pour votre groupe Auto Scaling.

> **Important**
>
> Avant de continuer, renseignez les prérequis de création d'un groupe chaud et confirmez que vous avez créé un hook de cycle de vie pour votre groupe Auto Scaling.

Créer un groupe d'instances pré-initialisées

Utilisez la procédure suivante pour créer un groupe chaud pour votre groupe Auto Scaling.

Pour créer un groupe d'instances pré-initialisées (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Activez la case à cocher en regard d'un groupe existant.

   Un volet fractionné s'ouvre en bas de la page.
3. Cliquez sur l'onglet Instance management (Gestion des instances).
4. Sous Warm pool (Groupe d'instances pré-initialisées), sélectionnez Create warm pool (Créer un groupe d'instances pré-initialisées).
5. Pour configurer un groupe d'instances pré-initialisées, procédez comme suit :

   a. Pour Warm pool instance state (État de l'instance du groupe d'instances pré-initialisées), choisissez l'état dans lequel vous souhaitez transférer vos instances lorsqu'elles intègrent le groupe. L'argument par défaut est `Stopped`.

   b. Pour Minimum warm pool size (Taille minimale du groupe d'instances pré-initialisées), entrez le nombre minimal d'instances à conserver dans le groupe.

   c. Pour la réutilisation des instances, cochez la case Reuse on scale in pour permettre aux instances du groupe Auto Scaling de retourner dans le pool de chaleur à l'échelle intégrée.

   d. Pour la taille de la piscine chaude, choisissez l'une des options disponibles :

- Spécification par défaut : La taille du pool de chaleur est déterminée par la différence entre la capacité maximale et la capacité souhaitée du groupe Auto Scaling. Cette option rationalise la gestion des piscines d'eau chaude. Après avoir créé le bassin d'eau chaude, sa taille peut être facilement mise à jour en ajustant simplement la capacité maximale du groupe.
- Spécification personnalisée : La taille du pool de chaleur est déterminée par la différence entre une valeur personnalisée et la capacité souhaitée du groupe Auto Scaling. Cette option vous donne la flexibilité de gérer la taille de votre piscine chaude indépendamment de la capacité maximale du groupe.

6. Consultez la section Taille estimée de la piscine chaude en fonction des paramètres actuels pour confirmer comment les spécifications par défaut ou personnalisées s'appliquent à la taille de la piscine chaude. N'oubliez pas que la taille du pool de chaleur dépend de la capacité souhaitée du groupe Auto Scaling, qui changera si le groupe évolue.
7. Choisissez Créer.

Supprimer un groupe d'instances pré-initialisées

Quand vous n'avez plus besoin du groupe d'instances pré-initialisées, utilisez la procédure suivante pour le supprimer.

Pour supprimer votre groupe d'instances pré-initialisées (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Activez la case à cocher en regard d'un groupe existant.

   Un volet fractionné s'ouvre en bas de la page.
3. Cliquez sur l'onglet Instance management (Gestion des instances).
4. Pour Warm pool (Groupe d'instances pré-initialisées), choisissez Actions, Delete (Supprimer).
5. Lorsque vous êtes invité à confirmer l’opération, choisissez Supprimer.

Afficher le statut de surveillance de l'état et les motifs des échecs de surveillances de l'état

La surveillance de l'état permet à Amazon EC2 Auto Scaling de déterminer si une instance est malsaine et doit être interrompue. Pour les instances de groupe d'instances pré-initialisées maintenues dans l'état `Stopped`, il utilise la connaissance qu'Amazon EBS a de la disponibilité d'une instance `Stopped` pour identifier les instances malsaines. Il le fait en appelant l'API `DescribeVolumeStatus` pour déterminer l'état du volume EBS qui est attaché à l'instance. Pour les instances de groupe d'instances pré-initialisées maintenues dans l'état `Running`, il s'appuie sur les vérifications d'état EC2 pour déterminer l'état de l'instance. Bien qu'il n'y ait pas de période de grâce de surveillance de l'état pour les instances de groupe d'instances pré-initialisées, Amazon EC2 Auto Scaling ne commence pas à surveiller l'état de l'instance tant que le hook de cycle de vie n'est pas terminé.

Lorsqu'une instance est jugée malsaine, Amazon EC2 Auto Scaling la supprime automatiquement et en crée une nouvelle pour la remplacer. Généralement, les instances sont interrompues quelques minutes après l'échec de la surveillance de leur état. Pour plus d'informations, consultez Afficher le motif des échecs d'une surveillance de l'état.

La surveillance personnalisée de l'état est également prise en charge. Cela peut être utile si vous disposez de votre propre système de surveillance de l'état qui peut détecter l'état d'une instance et envoyer ces informations à Amazon EC2 Auto Scaling. Pour plus d'informations, consultez Surveillances d'état personnalisées.

Dans la console Amazon EC2 Auto Scaling, vous pouvez afficher le statut (sain ou non sain) de vos instances de groupe d'instances pré-initialisées. Vous pouvez également consulter leur état de santé à l'aide du AWS CLI ou de l'un des SDK.

Pour afficher le statut de vos instances du groupe d'instances pré-initialisées (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page Auto Scaling groups (Groupes Auto Scaling).
3. Sous l'onglet Instance management (Gestion des instances) dans Warm groupe instances (Instances de groupe d'instances pré-initialisées), la colonne Lifecycle (Cycle de vie) affiche l'état de vos instances.

La colonne Health status (État d'intégrité) indique l'évaluation d'Amazon EC2 Auto Scaling portant sur l'état de l'instance.

> **Note**
>
> Les nouvelles instances commencent dans un état sain. Tant que le hook de cycle de vie n'est pas terminé, l'état d'une instance n'est pas surveillé.

Pour afficher le motif des échecs d'une surveillance de l'état (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page Auto Scaling groups (Groupes Auto Scaling).
3. Sous l'onglet Activity (Activité) sous Activity history (Historique des activités), la colonne Status (État) indique si votre groupe Auto Scaling a réussi à lancer ou à résilier des instances.

   Si des instances malsaines sont interrompues, la colonne Cause indique la date et l'heure de l'interruption et le motif de l'échec de la surveillance de l'état. Par exemple, « Sur 2021-04-01T 21:48:35 Z, une instance a été mise hors service en réponse à l'échec de la surveillance de l'état du volume EBS ».

Pour afficher le statut de vos instances du groupe d'instances pré-initialisées (AWS CLI)

Affichez le groupe d'instances pré-initialisées d'un groupe Auto Scaling à l'aide de la commande suivante : describe-warm-pool.

```
aws autoscaling describe-warm-pool --auto-scaling-group-name my-asg
```

Exemple de sortie.

```
{
    "WarmPoolConfiguration": {
        "MinSize": 0,
        "PoolState": "Stopped"
    },
```

```
    "Instances": [
        {
            "InstanceId": "i-0b5e5e7521cfaa46c",
            "InstanceType": "t2.micro",
            "AvailabilityZone": "us-west-2a",
            "LifecycleState": "Warmed:Stopped",
            "HealthStatus": "Healthy",
            "LaunchTemplate": {
                "LaunchTemplateId": "lt-08c4cd42f320d5dcd",
                "LaunchTemplateName": "my-template-for-auto-scaling",
                "Version": "1"
            }
        },
        {
            "InstanceId": "i-0e21af9dcfb7aa6bf",
            "InstanceType": "t2.micro",
            "AvailabilityZone": "us-west-2a",
            "LifecycleState": "Warmed:Stopped",
            "HealthStatus": "Healthy",
            "LaunchTemplate": {
                "LaunchTemplateId": "lt-08c4cd42f320d5dcd",
                "LaunchTemplateName": "my-template-for-auto-scaling",
                "Version": "1"
            }
        }
    ]
}
```

Pour afficher le motif des échecs d'une surveillance de l'état (AWS CLI)

Utilisez la commande describe-scaling-activities suivante.

```
aws autoscaling describe-scaling-activities --auto-scaling-group-name my-asg
```

Voici un exemple de réponse, où `Description` indique que votre groupe Auto Scaling a mis fin à une instance et `Cause` indique le motif de l'échec de la surveillance de l'état.

Les activités de mise à l'échelle sont classées par heure de début. Les activités toujours en cours sont décrites en premier lieu.

```
{
  "Activities": [
```

```
    {
      "ActivityId": "4c65e23d-a35a-4e7d-b6e4-2eaa8753dc12",
      "AutoScalingGroupName": "my-asg",
      "Description": "Terminating EC2 instance: i-04925c838b6438f14",
      "Cause": "At 2021-04-01T21:48:35Z an instance was taken out of service in
 response to EBS volume health check failure.",
      "StartTime": "2021-04-01T21:48:35.859Z",
      "EndTime": "2021-04-01T21:49:18Z",
      "StatusCode": "Successful",
      "Progress": 100,
      "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":\"us-west-2a
\"...}",
      "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:283179a2-
f3ce-423d-93f6-66bb518232f7:autoScalingGroupName/my-asg"
    },
...
  ]
}
```

Exemples de création et de gestion de piscines d'eau chaude à l'aide du AWS CLI

Vous pouvez créer et gérer des pools de chaleur à l'aide des kits de AWS Management Console développement logiciel AWS Command Line Interface (AWS CLI) ou des kits de développement logiciel (SDK).

Les exemples suivants vous montrent comment créer et gérer des groupes d'instances pré-initialisées à l'aide de la AWS CLI.

Table des matières

Exemple 1 : conserver des instances dans l'état **`Stopped`**

L'exemple de put-warm-pool suivant crée un groupe d'instances pré-initialisées qui maintient les instances dans un état `Stopped`.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Stopped
```

Exemple 2 : conserver des instances dans l'état **`Running`**

L'exemple put-warm-pool suivant crée un pool d'instances pré-initialisées qui maintient les instances dans une pile `Running` au lieu d'un état `Stopped`.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Running
```

Exemple 3 : conserver des instances dans l'état **`Hibernated`**

L'exemple de put-warm-pool suivant crée un groupe d'instances pré-initialisées qui maintient les instances dans une pile `Hibernated` au lieu d'un état `Stopped`. Cela vous permet d'arrêter les instances sans supprimer leur contenu en mémoire (RAM).

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Hibernated
```

Exemple 4 : renvoyer des instances au groupe d'instances pré-initialisées lors de la mise à l'échelle horizontale

L'exemple put-warm-pool suivant crée un groupe d'instances pré-initialisées qui maintient les instances dans un état `Stopped` et inclut l'option `--instance-reuse-policy`. La valeur `'{"ReuseOnScaleIn": true}'` de la politique de réutilisation d'instance indique à Amazon EC2 Auto Scaling de renvoyer les instances vers le groupe d'instances pré-initialisées lors de la mise à l'échelle horizontale de votre groupe Auto Scaling.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
```

```
  --pool-state Stopped --instance-reuse-policy '{"ReuseOnScaleIn": true}'
```

Exemple 5 : spécifier le nombre minimal d'instances dans le groupe d'instances pré-initialisées

L'exemple de put-warm-pool suivant crée un groupe d'instances pré-initialisées qui conserve un minimum de 4 instances, de sorte qu'il y ait au moins 4 instances disponibles pour gérer les pics de trafic.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Stopped --min-size 4
```

Exemple 6 : définir la taille de la piscine chaude à l'aide d'une spécification personnalisée

Par défaut, Amazon EC2 Auto Scaling gère la taille de votre pool de chaleur comme la différence entre la capacité maximale et la capacité souhaitée du groupe Auto Scaling. Cependant, vous pouvez gérer la taille de la piscine chaude indépendamment de la capacité maximale du groupe en utilisant `--max-group-prepared-capacity` cette option.

L'exemple put-warm-pool suivant crée un pool de chaleur et définit le nombre maximum d'instances pouvant exister simultanément dans le pool de chauffage et dans le groupe Auto Scaling. Si le groupe a une capacité souhaitée de 800 personnes, le pool de chaleur aura initialement une taille de 100 lorsqu'il s'initialisera après avoir exécuté cette commande.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Stopped --max-group-prepared-capacity 900
```

Pour conserver un nombre minimal d'instances dans le groupe, ajoutez l'option `--min-size` à la commande, comme suit.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Stopped --max-group-prepared-capacity 900 --min-size 25
```

Exemple 7 : définir une taille absolue de groupe d'instances pré-initialisées

Si vous définissez les mêmes valeurs pour les options `--max-group-prepared-capacity` et `--min-size`, le groupe d'instances pré-initialisées a une taille absolue. L'exemple put-warm-pool

suivant crée un groupe d'instances pré-initialisées qui maintient une taille de groupe d'instances pré-initialisées constante de 10 instances.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Stopped --min-size 10 --max-group-prepared-capacity 10
```

Exemple 8 : supprimer un groupe d'instances pré-initialisées

Utilisez la commande delete-warm-pool pour supprimer un groupe d'instances pré-initialisées.

```
aws autoscaling delete-warm-pool --auto-scaling-group-name my-asg
```

Si le groupe d'instances pré-initialisées comprend des instances, ou si des activités de mise à l'échelle sont en cours, utilisez la commande delete-warm-pool avec l'option `--force-delete`. Cette option résilie également les instances Amazon EC2 et toutes les actions de cycle de vie en attente.

```
aws autoscaling delete-warm-pool --auto-scaling-group-name my-asg --force-delete
```

Détacher ou attacher des instances

Vous pouvez détacher des instances de votre groupe Auto Scaling. Une fois qu'une instance est détachée, elle devient indépendante et peut être gérée seule ou attachée à un autre groupe Auto Scaling, distinct du groupe d'origine auquel elle appartenait. Cela peut être utile, par exemple, lorsque vous souhaitez effectuer des tests à l'aide d'instances existantes qui exécutent déjà votre application.

Cette rubrique fournit des instructions sur la manière de détacher et d'attacher des instances. Lorsque vous attachez des instances, vous pouvez également utiliser une instance existante plutôt qu'une instance détachée.

Au lieu de détacher et de rattacher une instance au même groupe, nous vous recommandons d'utiliser la procédure de mise en veille pour supprimer temporairement l'instance du groupe. Pour plus d'informations, consultez Supprimer temporairement des instances du groupe Auto Scaling.

Table des matières

Considérations relatives au détachement des instances

Lorsque vous détachez des instances, gardez les points suivants à l'esprit :

- Vous ne pouvez détacher une instance que lorsqu'elle est dans son `InService` état actuel.
- Une fois que vous avez détaché une instance, elle continue de fonctionner et d'être facturée. Pour éviter des frais inutiles, veillez à rattacher ou à résilier les instances détachées lorsqu'elles ne sont plus nécessaires.
- Vous pouvez choisir de réduire la capacité souhaitée en fonction du nombre d'instances que vous souhaitez détacher. Si vous choisissez de ne pas réduire la capacité, Amazon EC2 Auto Scaling lance de nouvelles instances pour remplacer les instances détachées afin de maintenir la capacité souhaitée.
- Si le nombre d'instances que vous détachez amène le groupe Auto Scaling en dessous de sa capacité minimale, vous devez réduire la capacité minimale.
- Si vous détachez plusieurs instances de la même zone de disponibilité sans réduire la capacité souhaitée, le groupe se rééquilibrera à moins que vous ne suspendiez le processus. `AZRebalance` Pour plus d'informations, consultez Suspendre et reprendre les processus Amazon EC2 Auto Scaling.
- Si vous détachez une instance d'un groupe Auto Scaling qui possède un groupe cible d'équilibreur de charge attaché ou un Classic Load Balancer, l'instance est désenregistrée de l'équilibreur de charge. Si le drainage de la connexion (délai de désinscription) est activé pour l'équilibreur de charge, Amazon EC2 Auto Scaling attend la fin des demandes à la volée.

> **Note**
>
> Si vous détachez des instances qui se trouvent dans l'état `Standby`, faites preuve de prudence. Une tentative de détachement des instances après les avoir placées dans l'état `Standby` peut entraîner la fermeture inattendue d'autres instances.

Considérations relatives à l'attachement d'instances

Notez les points suivants lorsque vous attachez des instances :

- Amazon EC2 Auto Scaling traite les instances jointes de la même manière que les instances lancées par le groupe lui-même. Cela signifie que les instances attachées peuvent être résiliées

lors d'événements d'extension si elles sont sélectionnées. Les autorisations accordées par le rôle AWSServiceRoleForAutoScaling lié au service permettent à Amazon EC2 Auto Scaling de le faire.

- Lorsque vous attachez des instances, la capacité souhaitée du groupe augmente en fonction du nombre d'instance attachées. Si la capacité souhaitée après l'ajout des nouvelles instances dépasse la taille maximale du groupe, la demande d'attachement d'autres instances échoue.
- Si vous ajoutez des instances à votre groupe, ce qui entraîne une répartition inégale entre les zones de disponibilité, Amazon EC2 Auto Scaling rééquilibre le groupe pour rétablir une distribution uniforme, sauf si vous suspendez le processus. `AZRebalance` Pour plus d'informations, consultez Suspendre et reprendre les processus Amazon EC2 Auto Scaling.
- Si vous attachez une instance à un groupe Auto Scaling qui possède un groupe cible d'équilibreur de charge attaché ou un Classic Load Balancer, l'instance est enregistrée avec l'équilibreur de charge.

Pour pouvoir attacher une instance, celle-ci doit répondre aux critères suivants :

- L'instance est dans l'état `running` avec Amazon EC2.
- L'AMI utilisée pour lancer l'instance doit toujours exister.
- L'instance ne fait pas partie d'un autre groupe Auto Scaling.
- L'instance est lancée dans l'une des zones de disponibilité définies dans le groupe Auto Scaling.
- Si le groupe Auto Scaling possède un groupe cible de l'équilibreur de charge attaché ou Classic Load Balancer, l'instance et l'équilibreur de charge doivent tous les deux se trouver dans le même VPC.

Déplacer une instance vers un autre groupe à l'aide de la fonction détacher et attacher

Utilisez l'une des procédures suivantes pour détacher une instance de votre groupe Auto Scaling et l'associer à un autre groupe Auto Scaling.

Pour créer un nouveau groupe Auto Scaling à partir d'une instance détachée, voir Créer un groupe Auto Scaling à l'aide des paramètres d'une instance existante (non recommandé, crée une configuration de lancement).

Console

Pour détacher une instance d'un groupe Auto Scaling

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Sous l'onglet Instance management (Gestion des instances) dans Instances, sélectionnez une instance et choisissez Actions, Detach (Détacher).
4. Dans la boîte de dialogue Détacher l'instance, gardez la case à cocher Remplacer l'instance sélectionnée pour lancer une instance de remplacement. Désactivez la case à cocher pour réduire la capacité souhaitée.
5. Lorsque vous êtes invité à confirmer l'opération, saisissez **`detach`** pour confirmer la suppression de l'instance spécifiée du groupe Auto Scaling, puis choisissez Détacher l'instance.

Vous pouvez désormais associer l'instance à un autre groupe Auto Scaling.

Pour attacher une instance à un groupe Auto Scaling

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. (Facultatif) Dans le panneau de navigation, sous Auto Scaling, choisissez Auto Scaling Groups (Groupes Auto Scaling). Sélectionnez le groupe Auto Scaling et vérifiez que sa taille maximale est suffisamment grande pour pouvoir ajouter une autre instance. Sinon, dans l'onglet Details (Détails), augmentez la capacité maximale.
3. Dans le panneau de navigation, sous Instances, choisissez Instances, puis sélectionnez une instance.
4. Choisissez Actions, Instance Settings (Paramètres de l'instance), puis Attach to Auto Scaling Group (Attacher à un groupe Auto Scaling).
5. Sur la page Attach to Auto Scaling group (Attacher à un groupe Auto Scaling), sous Auto Scaling Group (Groupe Auto Scaling) sélectionnez le groupe Auto Scaling, puis choisissez Attach (Attacher).
6. Si l'instance ne répond pas aux critères, un message d'erreur détaillé s'affiche. Par exemple, l'instance peut se trouver dans une zone de disponibilité différente de celle du groupe Auto

Scaling. Choisissez Fermer et réessayez avec un groupe Auto Scaling répondant aux critères.

AWS CLI

Pour détacher et attacher une instance, utilisez les exemples de commandes suivants. Remplacez chaque *espace réservé à la saisie de l'utilisateur* par vos propres informations.

Pour détacher une instance d'un groupe Auto Scaling

1. Pour décrire les instances actuelles, utilisez la commande describe-auto-scaling-instances suivante.

```
aws autoscaling describe-auto-scaling-instances \
  --query 'AutoScalingInstances[?AutoScalingGroupName==`my-asg`]'
```

L'exemple suivant montre le résultat produit lorsque vous exécutez cette commande.

Prenez note de l'ID de l'instance que vous souhaitez supprimer du groupe. Vous aurez besoin de cet identifiant à l'étape suivante.


```
{
    "AutoScalingInstances": [
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-05b4f7d5be44822a6",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
```

```
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-0c20ac468fa3049e8",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-0787762faf1c28619",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-0f280a4c58d319a8a",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        }
    ]
}
```


2. Pour détacher une instance sans réduire la capacité souhaitée, utilisez la commande detach-instances suivante.

```
aws autoscaling detach-instances --instance-ids i-05b4f7d5be44822a6 \
  --auto-scaling-group-name my-asg
```

Pour détacher une instance et réduire la capacité souhaitée, incluez l'`--should-decrement-desired-capacity`option.

```
aws autoscaling detach-instances --instance-ids i-05b4f7d5be44822a6 \
  --auto-scaling-group-name my-asg --should-decrement-desired-capacity
```

Vous pouvez désormais associer l'instance à un autre groupe Auto Scaling.

Pour attacher une instance à un groupe Auto Scaling

1. Pour associer l'instance à un autre groupe Auto Scaling, utilisez la commande attach-instances suivante.

```
aws autoscaling attach-instances --instance-ids i-05b4f7d5be44822a6 --auto-
scaling-group-name my-asg-for-testing
```

2. Pour vérifier la taille du groupe Auto Scaling après avoir attaché une instance, utilisez la commande describe-auto-scaling-groups suivante.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-names my-asg-
for-testing
```

L'exemple de réponse suivant montre que le groupe possède deux instances en cours d'exécution, dont l'une est l'instance que vous avez attachée.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg-for-testing",
            "AutoScalingGroupARN": "arn",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "2",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "MinSize": 1,
```

```
            "MaxSize": 5,
            "DesiredCapacity": 2,
            ...
            "Instances": [
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
                        "LaunchTemplateId": "lt-050555ad16a3f9c7f"
                    },
                    "InstanceId": "i-05b4f7d5be44822a6",
                    "InstanceType": "t3.micro",
                    "HealthStatus": "Healthy",
                    "LifecycleState": "InService"
                },
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "2",
                        "LaunchTemplateId": "lt-050555ad16a3f9c7f"
                    },
                    "InstanceId": "i-00dcdfffdf5175890",
                    "InstanceType": "t3.micro",
                    "HealthStatus": "Healthy",
                    "LifecycleState": "InService"
                }
            ],
            ...
        }
    ]
}
```

Supprimer temporairement des instances du groupe Auto Scaling

Vous pouvez faire passer une instance du statut `InService` au statut `Standby`, mettre à jour ou dépanner l'instance, puis la remettre en service. Les instances en veille font toujours partie du groupe Auto Scaling, mais elles ne gèrent pas activement le trafic de l'équilibreur de charge.

Cette fonction vous permet d'arrêter et de démarrer les instances ou de les redémarrer sans vous soucier qu'Amazon EC2 Auto Scaling résilie les instances dans le cadre de ses surveillances de l'état ou lors d'événements de mise à l'échelle horizontale.

Par exemple, vous pouvez modifier l'image Amazon Machine Image (AMI) pour un groupe Auto Scaling à tout moment en modifiant le modèle de lancement ou la configuration de lancement. Toutes les instances ultérieures lancées par le groupe Auto Scaling utilisent cette AMI. Cependant, le groupe Auto Scaling ne met pas à jour les instances actuellement en service. Vous pouvez résilier ces instances et laisser Amazon EC2 Auto Scaling les remplacer ou utiliser la fonction d'actualisation de l'instance pour résilier les instances et les remplacer. Vous pouvez également placer les instances en veille, mettre à jour le logiciel, puis remettre les instances en service.

Le détachement des instances d'un groupe Auto Scaling est similaire à la mise en veille des instances. Le détachement d'instances peut être utile si vous souhaitez les associer à un autre groupe ou gérer les instances comme des instances EC2 autonomes et éventuellement les mettre hors service. Pour plus d’informations, consultez Détacher ou attacher des instances.

Table des matières

Comment fonctionne l'état de veille ?

L'état de veille fonctionne comme suit pour vous aider à temporairement supprimer une instance d'un groupe Auto Scaling :

1. Vous mettez une instance en état de veille. L'instance reste dans cet état jusqu'à ce qu'elle change de statut.
2. Si un groupe cible d'équilibreur de charge ou Classic Load Balancer est attaché à votre groupe Auto Scaling, l'instance est désenregistrée de l'équilibreur de charge. Si Connection Draining est activée pour l'équilibreur de charge, Elastic Load Balancing attend 300 secondes par défaut avant de terminer le processus de désenregistration, ce qui permet aux demandes en cours d'exécution de se terminer.
3. Vous pouvez mettre à jour ou dépanner l'instance.

4. Vous remettez l'instance en service en la sortant de l'état de veille.
5. Si un groupe cible d'équilibreur de charge ou un Classic Load Balancer est attaché au groupe Auto Scaling, l'instance est enregistrée avec l'équilibreur de charge.

Pour plus d'informations sur le cycle de vie des instances dans un groupe Auto Scaling, consultez Cycle de vie d'une instance Amazon EC2 Auto Scaling.

Considérations

Les points suivants doivent être pris en compte lors du placement d'instances dans et hors de l'état de veille :

- Lorsque vous mettez une instance en veille, vous pouvez soit réduire la capacité souhaitée par le biais de cette opération, soit conserver la même valeur.
  - Si vous choisissez de ne pas réduire la capacité souhaitée du groupe Auto Scaling, Amazon EC2 Auto Scaling lance une instance pour remplacer l'instance en veille. L'objectif est de vous aider à préserver la capacité pour votre application tandis qu'une ou plusieurs instances sont en veille.
  - Si vous choisissez de réduire la capacité souhaitée du groupe Auto Scaling, cela empêche le lancement d'une instance pour remplacer l'instance en veille.
- Une fois l'instance remise en service, la capacité souhaitée est incrémentée pour refléter le nombre d'instances du groupe Auto Scaling.
- Pour procéder à l'incrémentation (et à la décrémentation), la nouvelle capacité souhaitée doit être comprise entre la taille de groupe minimale et maximale. Sinon, l'opération échoue.
- Si, après avoir mis une instance en veille ou remis l'instance en service en la sortant de l'état de veille, il s'avère que votre groupe Auto Scaling n'est pas équilibré entre les zones de disponibilité, Amazon EC2 Auto Scaling compense en rééquilibrant les zones de disponibilité, sauf si vous suspendez le processus `AZRebalance`. Pour plus d'informations, consultez Suspendre et reprendre les processus Amazon EC2 Auto Scaling.
- Vous êtes facturé pour les instances en état de veille.

État de santé d'une instance en veille

Amazon EC2 Auto Scaling ne réalise pas de surveillance de l'état sur des instances en veille. Lorsque l'instance est en veille, son état de santé reflète le statut qu'elle avait avant que vous ne la

mettiez en veille. Amazon EC2 Auto Scaling ne réalise pas de surveillance de l'état sur l'instance tant que vous ne la remettez pas en service.

Par exemple, si vous mettez une instance saine en veille et que vous y mettez fin, Amazon EC2 Auto Scaling continue de signaler l'instance comme saine. Si vous tentez de remettre en service une instance arrêtée qui était en veille, Amazon EC2 Auto Scaling effectue une surveillance de l'état de l'instance, détermine qu'elle est arrêtée et non saine, et lance une instance de remplacement. Pour plus d'informations, consultez Surveillance de l'état des instances dans un groupe Auto Scaling.

Supprimer temporairement une instance en la mettant en veille

Utilisez l'une des procédures suivantes pour mettre temporairement une instance hors service en la plaçant en mode veille.

Console

Pour supprimer temporairement une instance

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Sous l'onglet Gestion des instances) dans Instances, sélectionnez une instance.
4. Choisissez Actions, Set to Standby (Régler sur Veille).
5. Dans la boîte de dialogue Régler sur Veille, gardez la case à cocher Remplacer l'instance sélectionnée pour lancer une instance de remplacement. Désactivez la case à cocher pour réduire la capacité souhaitée.
6. Lorsque vous êtes invité à confirmer, tapez **`standby`** pour confirmer le placement de l'instance spécifiée dans l'état `Standby`, puis choisissez Régler sur Veille.
7. Vous pouvez mettre à jour ou dépanner l'instance le cas échéant. Lorsque vous avez terminé, continuez avec l'étape suivante pour remettre l'instance en service.
8. Sélectionnez l'instance, choisissez Actions, Définir sur InService. Dans la InService boîte de dialogue Définir sur, choisissez Définir sur InService.

AWS CLI

Pour supprimer temporairement une instance de votre groupe Auto Scaling, utilisez les exemples de commandes suivants. Remplacez chaque *espace réservé à la saisie de l'utilisateur* par vos propres informations.

Pour supprimer temporairement une instance

1. Utilisez la commande describe-auto-scaling-instances suivante pour identifier l'instance à mettre à jour.

```
aws autoscaling describe-auto-scaling-instances \
  --query 'AutoScalingInstances[?AutoScalingGroupName==`my-asg`]'
```

L'exemple suivant montre le résultat produit lorsque vous exécutez cette commande.

Prenez note de l'ID de l'instance que vous souhaitez supprimer du groupe. Vous aurez besoin de cet identifiant à l'étape suivante.

```
{
    "AutoScalingInstances": [
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-05b4f7d5be44822a6",
            "InstanceId": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
       ...
    ]
}
```

2. Passez l'instance à l'état `Standby` avec la commande enter-standby suivante. L'option `--should-decrement-desired-capacity` réduit la capacité souhaitée afin que le groupe Auto Scaling ne lance pas d'instance de remplacement.

```
aws autoscaling enter-standby --instance-ids i-05b4f7d5be44822a6 \
  --auto-scaling-group-name my-asg --should-decrement-desired-capacity
```

Voici un exemple de réponse.

```
{
    "Activities": [
        {
            "ActivityId": "3b1839fe-24b0-40d9-80ae-bcd883c2be32",
            "AutoScalingGroupName": "my-asg",
            "Description": "Moving EC2 instance to Standby:
 i-05b4f7d5be44822a6",
            "Cause": "At 2023-12-15T21:31:26Z instance i-05b4f7d5be44822a6 was
 moved to standby
              in response to a user request, shrinking the capacity from 4 to
 3.",
            "StartTime": "2023-12-15T21:31:26.150Z",
            "StatusCode": "InProgress",
            "Progress": 50,
            "Details": "{\"Subnet ID\":\"subnet-c934b782\",\"Availability Zone
\":\"us-west-2a\"}"
        }
    ]
}
```

3. (Facultatif) Vérifiez que l'instance est en `Standby` à l'aide de la commande describe-auto-scaling-instances.

```
aws autoscaling describe-auto-scaling-instances --instance-
ids i-05b4f7d5be44822a6
```

Voici un exemple de réponse. Notez que l'état de l'instance est désormais `Standby`.

```
{
    "AutoScalingInstances": [
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
```

```
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-05b4f7d5be44822a6",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "Standby"
        },
        ...
    ]
}
```

4. Vous pouvez mettre à jour ou dépanner l'instance le cas échéant. Lorsque vous avez terminé, continuez avec l'étape suivante pour remettre l'instance en service.
5. Remettez l'instance en service avec la commande exit-standby suivante.

```
aws autoscaling exit-standby --instance-ids i-05b4f7d5be44822a6 --auto-scaling-
group-name my-asg
```

Voici un exemple de réponse.

```
{
    "Activities": [
        {
            "ActivityId": "db12b166-cdcc-4c54-8aac-08c5935f8389",
            "AutoScalingGroupName": "my-asg",
            "Description": "Moving EC2 instance out of Standby:
 i-05b4f7d5be44822a6",
            "Cause": "At 2023-12-15T21:46:14Z instance i-05b4f7d5be44822a6 was
 moved out of standby in
               response to a user request, increasing the capacity from 3 to
 4.",
            "StartTime": "2023-12-15T21:46:14.678Z",
            "StatusCode": "PreInService",
            "Progress": 30,
            "Details": "{\"Subnet ID\":\"subnet-c934b782\",\"Availability Zone
\":\"us-west-2a\"}"
        }
    ]
}
```

6. (Facultatif) Vérifiez que l'instance est remise en service avec la commande `describe-auto-scaling-instances` suivante.

```
aws autoscaling describe-auto-scaling-instances --instance-
ids i-05b4f7d5be44822a6
```

Voici un exemple de réponse. Notez que l'état de l'instance est `InService`.

```
{
    "AutoScalingInstances": [
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-05b4f7d5be44822a6",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
       ...
    ]
}
```

Supprimer votre infrastructure Auto Scaling

Pour supprimer totalement l'infrastructure de mise à l'échelle, exécutez les tâches suivantes.

Tâches

- Supprimer votre groupe Auto Scaling
- (Facultatif) Supprimer la configuration du lancement
- (Facultatif) Suppression du modèle de lancement
- (Facultatif) Supprimer l'équilibreur de charge et les groupes cibles
- (Facultatif) Supprimer les CloudWatch alarmes

Supprimer votre groupe Auto Scaling

Lorsque vous supprimez un groupe Auto Scaling, ses valeurs minimales et maximales souhaitées sont définies sur 0. Les instances sont alors résiliées. La suppression d'une instance supprime également les journaux ou données associés, ainsi que tous les volumes de l'instance. Si vous ne souhaitez pas résilier une ou plusieurs instances, vous pouvez les détacher avant de supprimer le groupe Auto Scaling. Si le groupe a des politiques de mise à l'échelle, la suppression du groupe entraîne la suppression des politiques, des actions d'alarme sous-jacentes et de toute alarme qui n'a plus d'action associée.

Pour supprimer votre groupe Auto Scaling (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case en regard de votre groupe Auto Scaling et choisissez Actions, puis Supprimer.
3. Lorsque vous êtes invité à confirmer l'opération, saisissez **`delete`** pour confirmer la suppression du groupe Auto Scaling spécifié, puis choisissez Delete (Supprimer).

   Une icône de chargement dans la colonne Name (Nom) indique que le groupe Auto Scaling est en cours de suppression. Les colonnes Desired (Souhaitée), Min et Max affichent `0` instance pour le groupe Auto Scaling. Quelques minutes sont nécessaires pour résilier l'instance et supprimer le groupe. Actualisez la liste pour afficher l'état actuel.

Pour supprimer votre groupe Auto Scaling (AWS CLI)

Utilisez la commande delete-auto-scaling-group suivante pour supprimer le groupe Auto Scaling. Cette opération ne fonctionne pas si le groupe possède des instances EC2 ; elle concerne uniquement les groupes ne comportant aucune instance.

```
aws autoscaling delete-auto-scaling-group --auto-scaling-group-name my-asg
```

Si le groupe a des instances ou des activités de mise à l'échelle en cours, utilisez la commande delete-auto-scaling-group avec l'option `--force-delete`. Cette action entraînera également une résiliation des instances EC2. Lorsque vous supprimez un groupe Auto Scaling de la console Amazon EC2 Auto Scaling, la console utilise cette opération pour mettre fin à toutes les instances EC2 et supprimer le groupe en même temps.

```
aws autoscaling delete-auto-scaling-group --auto-scaling-group-name my-asg --force-
delete
```

(Facultatif) Supprimer la configuration du lancement

Vous pouvez passer cette étape pour conserver la configuration du lancement pour une utilisation ultérieure.

Pour supprimer la configuration du lancement (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Dans le volet de navigation à gauche, sous Auto Scaling, choisissez Groupes Auto Scaling.
3. Sélectionnez Configurations de lancement en haut de la page. Lorsque vous êtes invité à confirmer, choisissez Afficher les configurations de lancement pour confirmer que vous souhaitez consulter la page Configurations de lancement.
4. Sélectionnez votre configuration du lancement et cliquez sur Actions, Supprimer la configuration de lancement.
5. Lorsque vous êtes invité à confirmer l'opération, choisissez Supprimer.

Pour supprimer la configuration du lancement (AWS CLI)

Utilisez la commande delete-launch-configuration suivante.

```
aws autoscaling delete-launch-configuration --launch-configuration-name my-launch-
config
```

(Facultatif) Suppression du modèle de lancement

Vous pouvez supprimer votre modèle de lancement ou juste une version de votre modèle de lancement. Lorsque vous supprimez un modèle de lancement, toutes ses versions sont supprimées.

Vous pouvez ignorer cette étape pour conserver le modèle de lancement en vue d'une utilisation ultérieure.

Pour supprimer votre modèle de lancement (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.

2. Dans le volet de navigation, sous Instances, choisissez Launch Templates (Modèles de lancement).
3. Sélectionnez votre modèle de lancement, puis effectuez l'une des actions suivantes :
   - Choisissez Actions, puis Delete template (Supprimer le modèle). Lorsque vous êtes invité à confirmer l'opération, saisissez **Delete** pour confirmer la suppression du modèle de lancement spécifié, puis choisissez Delete (Supprimer).
   - Choisissez Actions, puis Delete template version (Supprimer la version du modèle). Sélectionnez la version à supprimer et choisissez Supprimer.

Pour supprimer le modèle de lancement (AWS CLI)

Utilisez la commande delete-launch-template suivante pour supprimer votre modèle et toutes ses versions.

```
aws ec2 delete-launch-template --launch-template-id lt-068f72b72934aff71
```

Vous pouvez également utiliser la commande delete-launch-template-versions pour supprimer une version spécifique d'un modèle de lancement.

```
aws ec2 delete-launch-template-versions --launch-template-id lt-068f72b72934aff71 --
versions 1
```

(Facultatif) Supprimer l'équilibreur de charge et les groupes cibles

Ignorez cette étape si votre groupe Auto Scaling n'est pas associé à un équilibreur de charge Elastic Load Balancing ou si vous souhaitez conserver ce dernier pour une utilisation ultérieure.

Pour supprimer l'équilibreur de charge (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Dans le panneau de navigation, sous Load Balancing (Équilibrage de charge), choisissez Load Balancers (Équilibreurs de charge).
3. Sélectionnez l'équilibreur de charge et choisissez Actions, Delete (Supprimer).
4. Lorsque vous êtes invité à confirmer l'opération, choisissez Oui, supprimer.

Pour supprimer votre groupe cible (console)

1. Dans le panneau de navigation, sous Load Balancing (Répartition de charge), choisissez Target Groups (Groupes cibles).
2. Sélectionnez le groupe cible et choisissez Actions, Delete (Supprimer).
3. Lorsque vous êtes invité à confirmer l'opération, choisissez Yes, Delete.

Pour supprimer l'équilibreur de charge associé au groupe Auto Scaling (AWS CLI)

Pour les équilibreurs de charge d'application (Application Load Balancer) et les équilibreurs de charge du réseau (Network Load Balancer), utilisez les commandes delete-load-balancer et delete-target-group.

```
aws elbv2 delete-load-balancer --load-balancer-arn my-load-balancer-arn
aws elbv2 delete-target-group --target-group-arn my-target-group-arn
```

Pour les Classic Load Balancers, utilisez la commande delete-load-balancer.

```
aws elb delete-load-balancer --load-balancer-name my-load-balancer
```

(Facultatif) Supprimer les CloudWatch alarmes

Pour supprimer les CloudWatch alarmes associées à votre groupe Auto Scaling, procédez comme suit. Par exemple, des alarmes peuvent être associées à la mise à l'échelle d'étape ou à de simples politiques de mise à l'échelle.

> **Note**
>
> La suppression d'un groupe Auto Scaling supprime automatiquement les CloudWatch alarmes gérées par Amazon EC2 Auto Scaling dans le cadre d'une politique de dimensionnement du suivi des cibles.

Vous pouvez ignorer cette étape si votre groupe Auto Scaling n'est associé à aucune CloudWatch alarme ou si vous souhaitez conserver les alarmes pour une utilisation future.

Pour supprimer les CloudWatch alarmes (console)

1. Ouvrez la CloudWatch console à l'adresse https://console.aws.amazon.com/cloudwatch/.

2. Dans le panneau de navigation, choisissez Alarms (Alarmes).
3. Sélectionnez les alarmes et choisissez Action, Delete (Supprimer).
4. Lorsque vous êtes invité à confirmer l'opération, choisissez Supprimer.

Pour supprimer les CloudWatch alarmes (AWS CLI)

Utilisez la commande delete-alarms suivante. Vous pouvez supprimer une ou plusieurs alarmes en même temps. Par exemple, utilisez la commande suivante pour supprimer les alarmes `Step-Scaling-AlarmHigh-AddCapacity` et `Step-Scaling-AlarmLow-RemoveCapacity`.

```
aws cloudwatch delete-alarms --alarm-name Step-Scaling-AlarmHigh-AddCapacity Step-Scaling-AlarmLow-RemoveCapacity
```

Exemples de création et de gestion de groupes Auto Scaling avec les AWS SDK

Vous pouvez créer un groupe Auto Scaling en utilisant le AWS Management Console, le AWS CLI, un AWS SDK, et AWS CloudFormation.

Les exemples de code suivants montrent comment créer, mettre à jour, décrire et supprimer un groupe Auto Scaling dans votre langage de programmation compatible préféré à l'aide AWS des SDK.

Table des matières

Création d'un groupe Auto Scaling à l'aide d'un AWS SDK

Les exemples de code suivants montrent comment utiliser`CreateAutoScalingGroup`.

.NET

AWS SDK for .NET

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
/// <summary>
/// Create a new Amazon EC2 Auto Scaling group.
/// </summary>
/// <param name="groupName">The name to use for the new Auto Scaling
/// group.</param>
/// <param name="launchTemplateName">The name of the Amazon EC2 Auto Scaling
/// launch template to use to create instances in the group.</param>
/// <returns>A Boolean value indicating the success of the action.</returns>
public async Task<bool> CreateAutoScalingGroupAsync(
    string groupName,
    string launchTemplateName,
    string availabilityZone)
{
    var templateSpecification = new LaunchTemplateSpecification
    {
        LaunchTemplateName = launchTemplateName,
    };

    var zoneList = new List<string>
        {
            availabilityZone,
        };

    var request = new CreateAutoScalingGroupRequest
    {
        AutoScalingGroupName = groupName,
        AvailabilityZones = zoneList,
        LaunchTemplate = templateSpecification,
        MaxSize = 6,
        MinSize = 1
    };
```

```
        var response = await
_amazonAutoScaling.CreateAutoScalingGroupAsync(request);
        Console.WriteLine($"{groupName} Auto Scaling Group created");
        return response.HttpStatusCode == System.Net.HttpStatusCode.OK;
    }
```

- Pour plus de détails sur l'API, reportez-vous CreateAutoScalingGroupà la section Référence des AWS SDK for .NET API.

C++

SDK pour C++

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

        Aws::AutoScaling::Model::CreateAutoScalingGroupRequest request;
        request.SetAutoScalingGroupName(groupName);
        Aws::Vector<Aws::String> availabilityGroupZones;
        availabilityGroupZones.push_back(
                availabilityZones[availabilityZoneChoice - 1].GetZoneName());
        request.SetAvailabilityZones(availabilityGroupZones);
        request.SetMaxSize(1);
        request.SetMinSize(1);

        Aws::AutoScaling::Model::LaunchTemplateSpecification
launchTemplateSpecification;
        launchTemplateSpecification.SetLaunchTemplateName(templateName);
        request.SetLaunchTemplate(launchTemplateSpecification);
```

```
        Aws::AutoScaling::Model::CreateAutoScalingGroupOutcome outcome =
                autoScalingClient.CreateAutoScalingGroup(request);

        if (outcome.IsSuccess()) {
            std::cout << "Created Auto Scaling group '" << groupName << "'..."
                      << std::endl;
        }
        else if (outcome.GetError().GetErrorType() ==
                 Aws::AutoScaling::AutoScalingErrors::ALREADY_EXISTS_FAULT) {
            std::cout << "Auto Scaling group '" << groupName << "' already
exists."
                      << std::endl;
        }
        else {
            std::cerr << "Error with AutoScaling::CreateAutoScalingGroup. "
                      << outcome.GetError().GetMessage()
                      << std::endl;

        }
```

- Pour plus de détails sur l'API, reportez-vous CreateAutoScalingGroupà la section Référence des AWS SDK for C++ API.

CLI

AWS CLI

Exemple 1 : pour créer un groupe Auto Scaling

L'`create-auto-scaling-group`exemple suivant crée un groupe Auto Scaling dans des sous-réseaux de plusieurs zones de disponibilité au sein d'une région. Les instances sont lancées avec la version par défaut du modèle de lancement spécifié. Notez que les valeurs par défaut sont utilisées pour la plupart des autres paramètres, tels que les politiques de résiliation et la configuration du bilan de santé.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateId=lt-1234567890abcde12 \
    --min-size 1 \
    --max-size 5 \
```

```
    --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, voir Groupes Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.

Exemple 2 : pour associer un Application Load Balancer, un Network Load Balancer ou un Gateway Load Balancer

Cet exemple indique l'ARN d'un groupe cible pour un équilibreur de charge qui prend en charge le trafic attendu. Le type de bilan de santé indique `ELB` que lorsqu'Elastic Load Balancing signale qu'une instance est défectueuse, le groupe Auto Scaling la remplace. La commande définit également un délai de grâce en `600` secondes pour le contrôle de santé. Le délai de grâce permet d'éviter la résiliation prématurée des instances nouvellement lancées.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateId=lt-1234567890abcde12 \
    --target-group-arns arn:aws:elasticloadbalancing:us-
west-2:123456789012:targetgroup/my-targets/943f017f100becff \
    --health-check-type ELB \
    --health-check-grace-period 600 \
    --min-size 1 \
    --max-size 5 \
    --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, consultez Elastic Load Balancing et Amazon EC2 Auto Scaling dans le manuel Guide l'utilisateur Amazon EC2 Auto Scaling.

Exemple 3 : pour spécifier un groupe de placement et utiliser la dernière version du modèle de lancement

Cet exemple lance des instances dans un groupe de placement au sein d'une seule zone de disponibilité. Cela peut être utile pour les groupes à faible latence soumis à des charges de travail HPC. Cet exemple indique également la taille minimale, la taille maximale et la capacité souhaitée du groupe.

```
aws autoscaling create-auto-scaling-group \
```

```
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateId=lt-1234567890abcde12,Version='$Latest' \
    --min-size 1 \
    --max-size 5 \
    --desired-capacity 3 \
    --placement-group my-placement-group \
    --vpc-zone-identifier "subnet-6194ea3b"
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, veuillez consulter la rubrique Groupes de placement dans le Guide de l'utilisateur Amazon EC2 pour les instances Linux.

Exemple 4 : Pour spécifier une instance unique, le groupe Auto Scaling et utiliser une version spécifique du modèle de lancement

Cet exemple crée un groupe Auto Scaling dont les capacités minimale et maximale sont définies 1 de manière à garantir l'exécution d'une instance. La commande spécifie également la v1 d'un modèle de lancement dans lequel l'ID d'un ENI existant est spécifié. Lorsque vous utilisez un modèle de lancement qui spécifie une ENI existante pour eth0, vous devez spécifier une zone de disponibilité pour le groupe Auto Scaling qui correspond à l'interface réseau, sans également spécifier d'ID de sous-réseau dans la demande.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg-single-instance \
    --launch-template LaunchTemplateName=my-template-for-auto-scaling,Version='1'
 \
    --min-size 1 \
    --max-size 1 \
    --availability-zones us-west-2a
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, voir Groupes Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.

Exemple 5 : pour définir une politique de résiliation différente

Cet exemple crée un groupe Auto Scaling à l'aide d'une configuration de lancement et définit la politique de résiliation pour mettre fin aux instances les plus anciennes en premier. La commande applique également une balise au groupe et à ses instances, avec une clé `Role` et une valeur de`WebServer`.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-configuration-name my-lc \
    --min-size 1 \
    --max-size 5 \
    --termination-policies "OldestInstance" \
    --tags "ResourceId=my-asg,ResourceType=auto-scaling-
group,Key=Role,Value=WebServer,PropagateAtLaunch=true" \
    --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, consultez la section Travailler avec les politiques de résiliation d'Amazon EC2 Auto Scaling dans le guide de l'utilisateur d'Amazon EC2 Auto Scaling.

Exemple 6 : pour spécifier un hook du cycle de vie de lancement

Cet exemple crée un groupe Auto Scaling avec un hook de cycle de vie qui prend en charge une action personnalisée lors du lancement de l'instance.

```
aws autoscaling create-auto-scaling-group \
    --cli-input-json file://~/config.json
```

Contenu du `config.json` fichier :

```
{
    "AutoScalingGroupName": "my-asg",
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-1234567890abcde12"
    },
    "LifecycleHookSpecificationList": [{
        "LifecycleHookName": "my-launch-hook",
        "LifecycleTransition": "autoscaling:EC2_INSTANCE_LAUNCHING",
        "NotificationTargetARN": "arn:aws:sqs:us-west-2:123456789012:my-sqs-
queue",
        "RoleARN": "arn:aws:iam::123456789012:role/my-notification-role",
        "NotificationMetadata": "SQS message metadata",
        "HeartbeatTimeout": 4800,
        "DefaultResult": "ABANDON"
    }],
    "MinSize": 1,
    "MaxSize": 5,
```

```
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
    "Tags": [{
        "ResourceType": "auto-scaling-group",
        "ResourceId": "my-asg",
        "PropagateAtLaunch": true,
        "Value": "test",
        "Key": "environment"
    }]
}
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, consultez Hooks du cycle de vie d'Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.

Exemple 7 : Pour spécifier un hook du cycle de vie de terminaison

Cet exemple crée un groupe Auto Scaling avec un hook de cycle de vie qui prend en charge une action personnalisée lors de la fermeture de l'instance.

```
aws autoscaling create-auto-scaling-group \
    --cli-input-json file://~/config.json
```

Contenu de `config.json` :

```
{
    "AutoScalingGroupName": "my-asg",
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-1234567890abcde12"
    },
    "LifecycleHookSpecificationList": [{
        "LifecycleHookName": "my-termination-hook",
        "LifecycleTransition": "autoscaling:EC2_INSTANCE_TERMINATING",
        "HeartbeatTimeout": 120,
        "DefaultResult": "CONTINUE"
    }],
    "MinSize": 1,
    "MaxSize": 5,
    "TargetGroupARNs": [
        "arn:aws:elasticloadbalancing:us-west-2:123456789012:targetgroup/my-
targets/73e2d6bc24d8a067"
    ],
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

```
}
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, consultez Hooks du cycle de vie d'Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.

Exemple 8 : pour définir une politique de résiliation personnalisée

Cet exemple crée un groupe Auto Scaling qui spécifie une politique d'arrêt de fonction Lambda personnalisée qui indique à Amazon EC2 Auto Scaling quelles instances peuvent être interrompues en toute sécurité à grande échelle.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg-single-instance \
    --launch-template LaunchTemplateName=my-template-for-auto-scaling \
    --min-size 1 \
    --max-size 5 \
    --termination-policies "arn:aws:lambda:us-
west-2:123456789012:function:HelloFunction:prod" \
    --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, consultez la section Création d'une politique de résiliation personnalisée avec Lambda dans le guide de l'utilisateur Amazon EC2 Auto Scaling.

- Pour plus de détails sur l'API, reportez-vous CreateAutoScalingGroupà la section Référence des AWS CLI commandes.

Java

SDK pour Java 2.x

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
import software.amazon.awssdk.core.waiters.WaiterResponse;
```

```
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.autoscaling.AutoScalingClient;
import software.amazon.awssdk.services.autoscaling.model.AutoScalingException;
import
 software.amazon.awssdk.services.autoscaling.model.CreateAutoScalingGroupRequest;
import
 software.amazon.awssdk.services.autoscaling.model.DescribeAutoScalingGroupsRequest;
import
 software.amazon.awssdk.services.autoscaling.model.DescribeAutoScalingGroupsResponse;
import
 software.amazon.awssdk.services.autoscaling.model.LaunchTemplateSpecification;
import software.amazon.awssdk.services.autoscaling.waiters.AutoScalingWaiter;

/**
 * Before running this SDK for Java (v2) code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 */
public class CreateAutoScalingGroup {
    public static void main(String[] args) {
        final String usage = """

                Usage:
                    <groupName> <launchTemplateName> <serviceLinkedRoleARN>
 <vpcZoneId>

                Where:
                    groupName - The name of the Auto Scaling group.
                    launchTemplateName - The name of the launch template.\s
                    vpcZoneId - A subnet Id for a virtual private cloud (VPC)
 where instances in the Auto Scaling group can be created.
                """;

        if (args.length != 3) {
            System.out.println(usage);
            System.exit(1);
        }

        String groupName = args[0];
        String launchTemplateName = args[1];
```

```
        String vpcZoneId = args[2];
        AutoScalingClient autoScalingClient = AutoScalingClient.builder()
                .region(Region.US_EAST_1)
                .build();

        createAutoScalingGroup(autoScalingClient, groupName, launchTemplateName,
vpcZoneId);
        autoScalingClient.close();
    }

    public static void createAutoScalingGroup(AutoScalingClient
autoScalingClient,
            String groupName,
            String launchTemplateName,
            String vpcZoneId) {

        try {
            AutoScalingWaiter waiter = autoScalingClient.waiter();
            LaunchTemplateSpecification templateSpecification =
LaunchTemplateSpecification.builder()
                    .launchTemplateName(launchTemplateName)
                    .build();

            CreateAutoScalingGroupRequest request =
CreateAutoScalingGroupRequest.builder()
                    .autoScalingGroupName(groupName)
                    .availabilityZones("us-east-1a")
                    .launchTemplate(templateSpecification)
                    .maxSize(1)
                    .minSize(1)
                    .vpcZoneIdentifier(vpcZoneId)
                    .build();

            autoScalingClient.createAutoScalingGroup(request);
            DescribeAutoScalingGroupsRequest groupsRequest =
DescribeAutoScalingGroupsRequest.builder()
                    .autoScalingGroupNames(groupName)
                    .build();

            WaiterResponse<DescribeAutoScalingGroupsResponse> waiterResponse =
waiter
                    .waitUntilGroupExists(groupsRequest);
            waiterResponse.matched().response().ifPresent(System.out::println);
            System.out.println("Auto Scaling Group created");
```

```
        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
}
```

- Pour plus de détails sur l'API, reportez-vous CreateAutoScalingGroupà la section Référence des AWS SDK for Java 2.x API.

Kotlin

SDK pour Kotlin

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
suspend fun createAutoScalingGroup(
    groupName: String,
    launchTemplateNameVal: String,
    serviceLinkedRoleARNVal: String,
    vpcZoneIdVal: String
) {
    val templateSpecification =
        LaunchTemplateSpecification {
            launchTemplateName = launchTemplateNameVal
        }

    val request =
        CreateAutoScalingGroupRequest {
            autoScalingGroupName = groupName
            availabilityZones = listOf("us-east-1a")
            launchTemplate = templateSpecification
            maxSize = 1
            minSize = 1
            vpcZoneIdentifier = vpcZoneIdVal
```

```
            serviceLinkedRoleArn = serviceLinkedRoleARNVal
        }

    // This object is required for the waiter call.
    val groupsRequestWaiter =
        DescribeAutoScalingGroupsRequest {
            autoScalingGroupNames = listOf(groupName)
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.createAutoScalingGroup(request)
        autoScalingClient.waitUntilGroupExists(groupsRequestWaiter)
        println("$groupName was created!")
    }
}
```

- Pour plus de détails sur l'API, reportez-vous CreateAutoScalingGroupà la section AWS SDK pour la référence de l'API Kotlin.

PHP

Kit SDK pour PHP

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
    public function createAutoScalingGroup(
        $autoScalingGroupName,
        $availabilityZones,
        $minSize,
        $maxSize,
        $launchTemplateId
    ) {
        return $this->autoScalingClient->createAutoScalingGroup([
            'AutoScalingGroupName' => $autoScalingGroupName,
            'AvailabilityZones' => $availabilityZones,
            'MinSize' => $minSize,
```

```
            'MaxSize' => $maxSize,
            'LaunchTemplate' => [
                'LaunchTemplateId' => $launchTemplateId,
            ],
        ]);
    }
```

- Pour plus de détails sur l'API, reportez-vous CreateAutoScalingGroupà la section Référence des AWS SDK for PHP API.

PowerShell

Outils pour PowerShell

Exemple 1 : Cet exemple crée un groupe Auto Scaling avec le nom et les attributs spécifiés. La capacité souhaitée par défaut est la taille minimale. Par conséquent, ce groupe Auto Scaling lance deux instances, une dans chacune des deux zones de disponibilité spécifiées.

```
New-ASAutoScalingGroup -AutoScalingGroupName my-asg -LaunchConfigurationName my-
lc -MinSize 2 -MaxSize 6 -AvailabilityZone @("us-west-2a", "us-west-2b")
```

- Pour plus de détails sur l'API, reportez-vous CreateAutoScalingGroupà la section Référence des AWS Tools for PowerShell applets de commande.

Python

SDK pour Python (Boto3)

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
class AutoScalingWrapper:
    """Encapsulates Amazon EC2 Auto Scaling actions."""

    def __init__(self, autoscaling_client):
        """
```

```
        :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client.
        """
        self.autoscaling_client = autoscaling_client


    def create_group(
        self, group_name, group_zones, launch_template_name, min_size, max_size
    ):
        """
        Creates an Auto Scaling group.

        :param group_name: The name to give to the group.
        :param group_zones: The Availability Zones in which instances can be
created.
        :param launch_template_name: The name of an existing Amazon EC2 launch
template.
                                     The launch template specifies the
configuration of
                                     instances that are created by auto scaling
activities.
        :param min_size: The minimum number of active instances in the group.
        :param max_size: The maximum number of active instances in the group.
        """
        try:
            self.autoscaling_client.create_auto_scaling_group(
                AutoScalingGroupName=group_name,
                AvailabilityZones=group_zones,
                LaunchTemplate={
                    "LaunchTemplateName": launch_template_name,
                    "Version": "$Default",
                },
                MinSize=min_size,
                MaxSize=max_size,
            )
        except ClientError as err:
            logger.error(
                "Couldn't create group %s. Here's why: %s: %s",
                group_name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
```

- Pour plus de détails sur l'API, consultez CreateAutoScalingGroupe AWS manuel de référence de l'API SDK for Python (Boto3).

Rust

SDK pour Rust

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
async fn create_group(client: &Client, name: &str, id: &str) -> Result<(), Error>
 {
    client
        .create_auto_scaling_group()
        .auto_scaling_group_name(name)
        .instance_id(id)
        .min_size(1)
        .max_size(5)
        .send()
        .await?;

    println!("Created AutoScaling group");

    Ok(())
}
```

- Pour plus de détails sur l'API, voir CreateAutoScalingGroupla section de référence de l'API AWS SDK for Rust.

Pour des exemples que vous pouvez utiliser lors de la création de groupes d'instances mixtes, consultez les ressources suivantes.

- AWS SDK pour .NET

- AWS SDK pour Go
- AWS SDK pour JavaScript
- AWS SDK pour PHP V3
- AWS SDK pour Python
- AWS SDK pour Ruby V3

Mettre à jour un groupe Auto Scaling à l'aide d'un AWS SDK

Les exemples de code suivants montrent comment utiliser`UpdateAutoScalingGroup`.

.NET

AWS SDK for .NET

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
    /// <summary>
    /// Update the capacity of an Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Auto Scaling group.</param>
    /// <param name="launchTemplateName">The name of the EC2 launch template.</
param>
    /// <param name="maxSize">The maximum number of instances that can be
    /// created for the Auto Scaling group.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> UpdateAutoScalingGroupAsync(
        string groupName,
        string launchTemplateName,
        int maxSize)
    {
        var templateSpecification = new LaunchTemplateSpecification
        {
            LaunchTemplateName = launchTemplateName,
        };

        var groupRequest = new UpdateAutoScalingGroupRequest
```

```
        {
            MaxSize = maxSize,
            AutoScalingGroupName = groupName,
            LaunchTemplate = templateSpecification,
        };

        var response = await
_amazonAutoScaling.UpdateAutoScalingGroupAsync(groupRequest);
        if (response.HttpStatusCode == System.Net.HttpStatusCode.OK)
        {
            Console.WriteLine($"You successfully updated the Auto Scaling group
{groupName}.");
            return true;
        }
        else
        {
            return false;
        }
    }
```

- Pour plus de détails sur l'API, reportez-vous UpdateAutoScalingGroupà la section Référence des AWS SDK for .NET API.

C++

SDK pour C++

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);
```

```
        Aws::AutoScaling::Model::UpdateAutoScalingGroupRequest request;
        request.SetAutoScalingGroupName(groupName);
        request.SetMaxSize(3);

        Aws::AutoScaling::Model::UpdateAutoScalingGroupOutcome outcome =
                autoScalingClient.UpdateAutoScalingGroup(request);

        if (!outcome.IsSuccess()) {
            std::cerr << "Error with AutoScaling::UpdateAutoScalingGroup. "
                      << outcome.GetError().GetMessage()
                      << std::endl;

        }
```

- Pour plus de détails sur l'API, reportez-vous UpdateAutoScalingGroupà la section Référence des AWS SDK for C++ API.

CLI

AWS CLI

Exemple 1 : Pour mettre à jour les limites de taille d'un groupe Auto Scaling

Cet exemple met à jour le groupe Auto Scaling spécifié avec une taille minimale de 2 et une taille maximale de 10.

```
aws autoscaling update-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --min-size 2 \
    --max-size 10
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, consultez la section Définition des limites de capacité pour votre groupe Auto Scaling dans le guide de l'utilisateur d'Amazon EC2 Auto Scaling.

Exemple 2 : pour ajouter des contrôles de santé d'Elastic Load Balancing et spécifier les zones de disponibilité et les sous-réseaux à utiliser

Cet exemple met à jour le groupe Auto Scaling spécifié pour ajouter les tests de santé d'Elastic Load Balancing. Cette commande met également à jour la valeur de `--vpc-`

`zone-identifier` avec une liste d'identifiants de sous-réseaux dans plusieurs zones de disponibilité.

```
aws autoscaling update-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --health-check-type ELB \
    --health-check-grace-period 600 \
    --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, consultez Elastic Load Balancing et Amazon EC2 Auto Scaling dans le manuel Guide l'utilisateur Amazon EC2 Auto Scaling.

Exemple 3 : pour mettre à jour le groupe de placement et la politique de résiliation

Cet exemple met à jour le groupe de placement et la politique de résiliation à utiliser.

```
aws autoscaling update-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --placement-group my-placement-group \
    --termination-policies "OldestInstance"
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, voir Groupes Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.

Exemple 4 : Pour utiliser la dernière version du modèle de lancement

Cet exemple met à jour le groupe Auto Scaling spécifié pour utiliser la dernière version du modèle de lancement spécifié.

```
aws autoscaling update-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateId=lt-1234567890abcde12,Version='$Latest'
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, consultez Modèles de lancement dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.

Exemple 5 : Pour utiliser une version spécifique du modèle de lancement

Cet exemple met à jour le groupe Auto Scaling spécifié pour utiliser une version spécifique d'un modèle de lancement au lieu de la version la plus récente ou de la version par défaut.

```
aws autoscaling update-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateName=my-template-for-auto-scaling,Version='2'
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, consultez Modèles de lancement dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.

Exemple 6 : pour définir une politique d'instances mixtes et activer le rééquilibrage des capacités

Cet exemple met à jour le groupe Auto Scaling spécifié afin d'utiliser une politique d'instances mixtes et d'activer le rééquilibrage des capacités. Cette structure vous permet de spécifier des groupes dotés de capacités ponctuelles et à la demande et d'utiliser différents modèles de lancement pour différentes architectures.

```
aws autoscaling update-auto-scaling-group \
    --cli-input-json file://~/config.json
```

Contenu de `config.json` :

```
{
    "AutoScalingGroupName": "my-asg",
    "CapacityRebalance": true,
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template-for-x86",
                "Version": "$Latest"
            },
            "Overrides": [
                {
                    "InstanceType": "c6g.large",
                    "LaunchTemplateSpecification": {
                        "LaunchTemplateName": "my-launch-template-for-arm",
                        "Version": "$Latest"
```

```
                    }
                },
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandPercentageAboveBaseCapacity": 50,
            "SpotAllocationStrategy": "capacity-optimized"
        }
    }
}
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, consultez Groupes Auto Scaling avec types d'instance et options d'achat multiples dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.

- Pour plus de détails sur l'API, reportez-vous UpdateAutoScalingGroupà la section Référence des AWS CLI commandes.

Java

SDK pour Java 2.x

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
   public static void updateAutoScalingGroup(AutoScalingClient
autoScalingClient, String groupName,
            String launchTemplateName) {
        try {
            AutoScalingWaiter waiter = autoScalingClient.waiter();
            LaunchTemplateSpecification templateSpecification =
LaunchTemplateSpecification.builder()
```

```
                    .launchTemplateName(launchTemplateName)
                    .build();

            UpdateAutoScalingGroupRequest groupRequest =
UpdateAutoScalingGroupRequest.builder()
                    .maxSize(3)
                    .autoScalingGroupName(groupName)
                    .launchTemplate(templateSpecification)
                    .build();

            autoScalingClient.updateAutoScalingGroup(groupRequest);
            DescribeAutoScalingGroupsRequest groupsRequest =
DescribeAutoScalingGroupsRequest.builder()
                    .autoScalingGroupNames(groupName)
                    .build();

            WaiterResponse<DescribeAutoScalingGroupsResponse> waiterResponse =
waiter
                    .waitUntilGroupInService(groupsRequest);
            waiterResponse.matched().response().ifPresent(System.out::println);
            System.out.println("You successfully updated the auto scaling group
" + groupName);

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
```

- Pour plus de détails sur l'API, reportez-vous UpdateAutoScalingGroupà la section Référence des AWS SDK for Java 2.x API.

Kotlin

SDK pour Kotlin

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
suspend fun updateAutoScalingGroup(
    groupName: String,
    launchTemplateNameVal: String,
    serviceLinkedRoleARNVal: String
) {
    val templateSpecification =
        LaunchTemplateSpecification {
            launchTemplateName = launchTemplateNameVal
        }

    val groupRequest =
        UpdateAutoScalingGroupRequest {
            maxSize = 3
            serviceLinkedRoleArn = serviceLinkedRoleARNVal
            autoScalingGroupName = groupName
            launchTemplate = templateSpecification
        }

    val groupsRequestWaiter =
        DescribeAutoScalingGroupsRequest {
            autoScalingGroupNames = listOf(groupName)
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.updateAutoScalingGroup(groupRequest)
        autoScalingClient.waitUntilGroupExists(groupsRequestWaiter)
        println("You successfully updated the Auto Scaling group  $groupName")
    }
}
```

- Pour plus de détails sur l'API, reportez-vous UpdateAutoScalingGroupà la section AWS SDK pour la référence de l'API Kotlin.

PHP

Kit SDK pour PHP

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
public function updateAutoScalingGroup($autoScalingGroupName, $args)
{
    if (array_key_exists('MaxSize', $args)) {
        $maxSize = ['MaxSize' => $args['MaxSize']];
    } else {
        $maxSize = [];
    }
    if (array_key_exists('MinSize', $args)) {
        $minSize = ['MinSize' => $args['MinSize']];
    } else {
        $minSize = [];
    }
    $parameters = ['AutoScalingGroupName' => $autoScalingGroupName];
    $parameters = array_merge($parameters, $minSize, $maxSize);
    return $this->autoScalingClient->updateAutoScalingGroup($parameters);
}
```

- Pour plus de détails sur l'API, reportez-vous UpdateAutoScalingGroupà la section Référence des AWS SDK for PHP API.

PowerShell

Outils pour PowerShell

Exemple 1 : Cet exemple met à jour la taille minimale et maximale du groupe Auto Scaling spécifié.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -MaxSize 5 -MinSize 1
```

Exemple 2 : Cet exemple met à jour la période de recharge par défaut du groupe Auto Scaling spécifié.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -DefaultCooldown 10
```

Exemple 3 : Cet exemple met à jour les zones de disponibilité du groupe Auto Scaling spécifié.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -AvailabilityZone @("us-
west-2a", "us-west-2b")
```

Exemple 4 : Cet exemple met à jour le groupe Auto Scaling spécifié pour utiliser les contrôles de santé d'Elastic Load Balancing.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -HealthCheckType ELB -
HealthCheckGracePeriod 60
```

- Pour plus de détails sur l'API, reportez-vous UpdateAutoScalingGroupà la section Référence des AWS Tools for PowerShell applets de commande.

Python

SDK pour Python (Boto3)

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
class AutoScalingWrapper:
    """Encapsulates Amazon EC2 Auto Scaling actions."""

    def __init__(self, autoscaling_client):
        """
        :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client.
        """
        self.autoscaling_client = autoscaling_client


    def update_group(self, group_name, **kwargs):
```

```
        """
        Updates an Auto Scaling group.

        :param group_name: The name of the group to update.
        :param kwargs: Keyword arguments to pass through to the service.
        """
        try:
            self.autoscaling_client.update_auto_scaling_group(
                AutoScalingGroupName=group_name, **kwargs
            )
        except ClientError as err:
            logger.error(
                "Couldn't update group %s. Here's why: %s: %s",
                group_name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
```

- Pour plus de détails sur l'API, consultez UpdateAutoScalingGroupe AWS manuel de référence de l'API SDK for Python (Boto3).

Rust

SDK pour Rust

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
async fn update_group(client: &Client, name: &str, size: i32) -> Result<(),
 Error> {
    client
        .update_auto_scaling_group()
        .auto_scaling_group_name(name)
        .max_size(size)
        .send()
```

```
        .await?;

    println!("Updated AutoScaling group");

    Ok(())
}
```

- Pour plus de détails sur l'API, voir UpdateAutoScalingGroupla section de référence de l'API AWS SDK for Rust.

Décrire un groupe Auto Scaling à l'aide d'un AWS SDK

Les exemples de code suivants montrent comment utiliser`DescribeAutoScalingGroups`.

.NET

AWS SDK for .NET

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
    /// <summary>
    /// Get data about the instances in an Amazon EC2 Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Amazon EC2 Auto Scaling group.</
param>
    /// <returns>A list of Amazon EC2 Auto Scaling details.</returns>
    public async Task<List<AutoScalingInstanceDetails>>
 DescribeAutoScalingInstancesAsync(
        string groupName)
    {
        var groups = await DescribeAutoScalingGroupsAsync(groupName);
        var instanceIds = new List<string>();
        groups!.ForEach(group =>
        {
            if (group.AutoScalingGroupName == groupName)
```

```
            {
                group.Instances.ForEach(instance =>
                {
                    instanceIds.Add(instance.InstanceId);
                });
            }
        });

        var scalingGroupsRequest = new DescribeAutoScalingInstancesRequest
        {
            MaxRecords = 10,
            InstanceIds = instanceIds,
        };

        var response = await
_amazonAutoScaling.DescribeAutoScalingInstancesAsync(scalingGroupsRequest);
        var instanceDetails = response.AutoScalingInstances;

        return instanceDetails;
    }
```

- Pour plus de détails sur l'API, reportez-vous DescribeAutoScalingGroupsà la section Référence des AWS SDK for .NET API.

C++

SDK pour C++

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);
```

```
    Aws::AutoScaling::Model::DescribeAutoScalingGroupsRequest request;
    Aws::Vector<Aws::String> groupNames;
    groupNames.push_back(groupName);
    request.SetAutoScalingGroupNames(groupNames);

    Aws::AutoScaling::Model::DescribeAutoScalingGroupsOutcome outcome =
            client.DescribeAutoScalingGroups(request);

    if (outcome.IsSuccess()) {
        autoScalingGroup = outcome.GetResult().GetAutoScalingGroups();
    }
    else {
        std::cerr << "Error with AutoScaling::DescribeAutoScalingGroups. "
                  << outcome.GetError().GetMessage()
                  << std::endl;
    }
```

- Pour plus de détails sur l'API, reportez-vous DescribeAutoScalingGroupsà la section Référence des AWS SDK for C++ API.

CLI

AWS CLI

Exemple 1 : Pour décrire le groupe Auto Scaling spécifié

Cet exemple décrit le groupe Auto Scaling spécifié.

```
aws autoscaling describe-auto-scaling-groups \
    --auto-scaling-group-name my-asg
```

Sortie :

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:930d940e-891e-4781-
a11a-7b0acd480f03:autoScalingGroupName/my-asg",
```

```
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-1234567890abcde12"
            },
            "MinSize": 0,
            "MaxSize": 1,
            "DesiredCapacity": 1,
            "DefaultCooldown": 300,
            "AvailabilityZones": [
                "us-west-2a",
                "us-west-2b",
                "us-west-2c"
            ],
            "LoadBalancerNames": [],
            "TargetGroupARNs": [],
            "HealthCheckType": "EC2",
            "HealthCheckGracePeriod": 0,
            "Instances": [
                {
                    "InstanceId": "i-06905f55584de02da",
                    "InstanceType": "t2.micro",
                    "AvailabilityZone": "us-west-2a",
                    "HealthStatus": "Healthy",
                    "LifecycleState": "InService",
                    "ProtectedFromScaleIn": false,
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
                        "LaunchTemplateId": "lt-1234567890abcde12"
                    }
                }
            ],
            "CreatedTime": "2023-10-28T02:39:22.152Z",
            "SuspendedProcesses": [],
            "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-
c934b782",
            "EnabledMetrics": [],
            "Tags": [],
            "TerminationPolicies": [
                "Default"
            ],
            "NewInstancesProtectedFromScaleIn": false,
            "ServiceLinkedRoleARN":"arn",
```

```
            "TrafficSources": []
        }
    ]
}
```

Exemple 2 : Pour décrire les 100 premiers groupes Auto Scaling spécifiés

Cet exemple décrit les groupes Auto Scaling spécifiés. Il vous permet de spécifier jusqu'à 100 noms de groupes.

```
aws autoscaling describe-auto-scaling-groups \
    --max-items 100 \
    --auto-scaling-group-name "group1" "group2" "group3" "group4"
```

Voir l'exemple 1 pour un exemple de sortie.

Exemple 3 : Pour décrire un groupe Auto Scaling dans la région spécifiée

Cet exemple décrit les groupes Auto Scaling dans la région spécifiée, jusqu'à un maximum de 75 groupes.

```
aws autoscaling describe-auto-scaling-groups \
    --max-items 75 \
    --region us-east-1
```

Voir l'exemple 1 pour un exemple de sortie.

Exemple 4 : Pour décrire le nombre spécifié de groupes Auto Scaling

Pour renvoyer un nombre spécifique de groupes Auto Scaling, utilisez l'`--max-items`option.

```
aws autoscaling describe-auto-scaling-groups \
    --max-items 1
```

Voir l'exemple 1 pour un exemple de sortie.

Si la sortie inclut un `NextToken` champ, il existe d'autres groupes. Pour obtenir les groupes supplémentaires, utilisez la valeur de ce champ avec l'`--starting-token`option lors d'un appel suivant comme suit.

```
aws autoscaling describe-auto-scaling-groups \
    --starting-token Z3M3LMPEXAMPLE
```

Voir l'exemple 1 pour un exemple de sortie.

Exemple 5 : Pour décrire les groupes Auto Scaling qui utilisent des configurations de lancement

Cet exemple utilise l'`--query`option pour décrire les groupes Auto Scaling qui utilisent des configurations de lancement.

```
aws autoscaling describe-auto-scaling-groups \
    --query 'AutoScalingGroups[?LaunchConfigurationName!=`null`]'
```

Sortie :

```
[
    {
        "AutoScalingGroupName": "my-asg",
        "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:930d940e-891e-4781-
a11a-7b0acd480f03:autoScalingGroupName/my-asg",
        "LaunchConfigurationName": "my-lc",
        "MinSize": 0,
        "MaxSize": 1,
        "DesiredCapacity": 1,
        "DefaultCooldown": 300,
        "AvailabilityZones": [
            "us-west-2a",
            "us-west-2b",
            "us-west-2c"
        ],
        "LoadBalancerNames": [],
        "TargetGroupARNs": [],
        "HealthCheckType": "EC2",
        "HealthCheckGracePeriod": 0,
        "Instances": [
            {
                "InstanceId": "i-088c57934a6449037",
                "InstanceType": "t2.micro",
                "AvailabilityZone": "us-west-2c",
                "HealthStatus": "Healthy",
                "LifecycleState": "InService",
                "LaunchConfigurationName": "my-lc",
                "ProtectedFromScaleIn": false
            }
```

```
        ],
        "CreatedTime": "2023-10-28T02:39:22.152Z",
        "SuspendedProcesses": [],
        "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
        "EnabledMetrics": [],
        "Tags": [],
        "TerminationPolicies": [
            "Default"
        ],
        "NewInstancesProtectedFromScaleIn": false,
        "ServiceLinkedRoleARN":"arn",
        "TrafficSources": []
    }
]
```

Pour plus d'informations, consultez la section Filtrer la sortie de la AWS CLI dans le guide de l'utilisateur de l'interface de ligne de AWS commande.

- Pour plus de détails sur l'API, reportez-vous DescribeAutoScalingGroupsà la section Référence des AWS CLI commandes.

Java

SDK pour Java 2.x

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.autoscaling.AutoScalingClient;
import software.amazon.awssdk.services.autoscaling.model.AutoScalingException;
import software.amazon.awssdk.services.autoscaling.model.AutoScalingGroup;
import
 software.amazon.awssdk.services.autoscaling.model.DescribeAutoScalingGroupsResponse;
import
 software.amazon.awssdk.services.autoscaling.model.DescribeAutoScalingGroupsRequest;
import software.amazon.awssdk.services.autoscaling.model.Instance;
import java.util.List;
```

```
/**
 * Before running this SDK for Java (v2) code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 */
public class DescribeAutoScalingInstances {
    public static void main(String[] args) {
        final String usage = """

                Usage:
                    <groupName>

                Where:
                    groupName - The name of the Auto Scaling group.
                """;

        if (args.length != 1) {
            System.out.println(usage);
            System.exit(1);
        }

        String groupName = args[0];
        AutoScalingClient autoScalingClient = AutoScalingClient.builder()
                .region(Region.US_EAST_1)
                .build();

        String instanceId = getAutoScaling(autoScalingClient, groupName);
        System.out.println(instanceId);
        autoScalingClient.close();
    }

    public static String getAutoScaling(AutoScalingClient autoScalingClient,
 String groupName) {
        try {
            String instanceId = "";
            DescribeAutoScalingGroupsRequest scalingGroupsRequest =
 DescribeAutoScalingGroupsRequest.builder()
                    .autoScalingGroupNames(groupName)
                    .build();
```

```
            DescribeAutoScalingGroupsResponse response = autoScalingClient
                    .describeAutoScalingGroups(scalingGroupsRequest);
            List<AutoScalingGroup> groups = response.autoScalingGroups();
            for (AutoScalingGroup group : groups) {
                System.out.println("The group name is " +
 group.autoScalingGroupName());
                System.out.println("The group ARN is " +
 group.autoScalingGroupARN());

                List<Instance> instances = group.instances();
                for (Instance instance : instances) {
                    instanceId = instance.instanceId();
                }
            }
            return instanceId;
        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
        return "";
    }
}
```

- Pour plus de détails sur l'API, reportez-vous DescribeAutoScalingGroupsà la section Référence des AWS SDK for Java 2.x API.

Kotlin

SDK pour Kotlin

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
suspend fun getAutoScalingGroups(groupName: String) {
    val scalingGroupsRequest =
        DescribeAutoScalingGroupsRequest {
            autoScalingGroupNames = listOf(groupName)
```

```
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        val response =
 autoScalingClient.describeAutoScalingGroups(scalingGroupsRequest)
        response.autoScalingGroups?.forEach { group ->
            println("The group name is ${group.autoScalingGroupName}")
            println("The group ARN is ${group.autoScalingGroupArn}")
            group.instances?.forEach { instance ->
                println("The instance id is ${instance.instanceId}")
                println("The lifecycle state is " + instance.lifecycleState)
            }
        }
    }
}
```

- Pour plus de détails sur l'API, reportez-vous DescribeAutoScalingGroupsà la section AWS SDK pour la référence de l'API Kotlin.

PHP

Kit SDK pour PHP

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
    public function describeAutoScalingGroups($autoScalingGroupNames)
    {
        return $this->autoScalingClient->describeAutoScalingGroups([
            'AutoScalingGroupNames' => $autoScalingGroupNames
        ]);
    }
```

- Pour plus de détails sur l'API, reportez-vous DescribeAutoScalingGroupsà la section Référence des AWS SDK for PHP API.

PowerShell

Outils pour PowerShell

Exemple 1 : Cet exemple répertorie les noms de vos groupes Auto Scaling.

```
Get-ASAutoScalingGroup | format-table -property AutoScalingGroupName
```

Sortie :

```
AutoScalingGroupName
--------------------
my-asg-1
my-asg-2
my-asg-3
my-asg-4
my-asg-5
my-asg-6
```

Exemple 2 : Cet exemple décrit le groupe Auto Scaling spécifié.

```
Get-ASAutoScalingGroup -AutoScalingGroupName my-asg-1
```

Sortie :

```
AutoScalingGroupARN     : arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:930d940e-891e-4781-a11a-7b0acd480
                          f03:autoScalingGroupName/my-asg-1
AutoScalingGroupName    : my-asg-1
AvailabilityZones       : {us-west-2b, us-west-2a}
CreatedTime             : 3/1/2015 9:05:31 AM
DefaultCooldown         : 300
DesiredCapacity         : 2
EnabledMetrics          : {}
HealthCheckGracePeriod  : 300
HealthCheckType         : EC2
Instances               : {my-lc}
LaunchConfigurationName : my-lc
LoadBalancerNames       : {}
MaxSize                 : 0
MinSize                 : 0
PlacementGroup          :
```

```
Status                  :
SuspendedProcesses      : {}
Tags                    : {}
TerminationPolicies     : {Default}
VPCZoneIdentifier       : subnet-e4f33493,subnet-5264e837
```

Exemple 3 : Cet exemple décrit les deux groupes Auto Scaling spécifiés.

```
Get-ASAutoScalingGroup -AutoScalingGroupName @("my-asg-1", "my-asg-2")
```

Exemple 4 : Cet exemple décrit les instances Auto Scaling pour le groupe Auto Scaling spécifié.

```
(Get-ASAutoScalingGroup -AutoScalingGroupName my-asg-1).Instances
```

Exemple 5 : Cet exemple décrit tous vos groupes Auto Scaling.

```
Get-ASAutoScalingGroup
```

Exemple 6 : Cet exemple décrit tous vos groupes Auto Scaling, par lots de 10.

```
$nextToken = $null
do {
  Get-ASAutoScalingGroup -NextToken $nextToken -MaxRecord 10
  $nextToken = $AWSHistory.LastServiceResponse.NextToken
} while ($nextToken -ne $null)
```

Exemple 7 : Cet LaunchTemplate exemple décrit le groupe Auto Scaling spécifié. Cet exemple suppose que les « Options d'achat d'instance » sont définies sur « Adhérer au modèle de lancement ». Si cette option est définie sur « Combiner les options d'achat et les types d'instances », elle est accessible LaunchTemplate à l'aide de « MixedInstances Policy ». LaunchTemplate« propriété.

```
(Get-ASAutoScalingGroup -AutoScalingGroupName my-ag-1).LaunchTemplate
```

Sortie :

```
LaunchTemplateId     LaunchTemplateName   Version
----------------     ------------------   -------
```

```
lt-06095fd619cb40371 test-launch-template $Default
```

- Pour plus de détails sur l'API, reportez-vous DescribeAutoScalingGroupsà la section Référence des AWS Tools for PowerShell applets de commande.

Python

SDK pour Python (Boto3)

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
class AutoScalingWrapper:
    """Encapsulates Amazon EC2 Auto Scaling actions."""

    def __init__(self, autoscaling_client):
        """
        :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client.
        """
        self.autoscaling_client = autoscaling_client


    def describe_group(self, group_name):
        """
        Gets information about an Auto Scaling group.

        :param group_name: The name of the group to look up.
        :return: Information about the group, if found.
        """
        try:
            response = self.autoscaling_client.describe_auto_scaling_groups(
                AutoScalingGroupNames=[group_name]
            )
        except ClientError as err:
            logger.error(
                "Couldn't describe group %s. Here's why: %s: %s",
                group_name,
                err.response["Error"]["Code"],
```

```
                err.response["Error"]["Message"],
            )
            raise
        else:
            groups = response.get("AutoScalingGroups", [])
            return groups[0] if len(groups) > 0 else None
```

- Pour plus de détails sur l'API, consultez DescribeAutoScalingGroupsle AWS manuel de référence de l'API SDK for Python (Boto3).

Rust

SDK pour Rust

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
async fn list_groups(client: &Client) -> Result<(), Error> {
    let resp = client.describe_auto_scaling_groups().send().await?;

    println!("Groups:");

    let groups = resp.auto_scaling_groups();

    for group in groups {
        println!(
            "Name:  {}",
            group.auto_scaling_group_name().unwrap_or("Unknown")
        );
        println!(
            "Arn:   {}",
            group.auto_scaling_group_arn().unwrap_or("unknown"),
        );
        println!("Zones: {:?}", group.availability_zones(),);
        println!();
    }
```

```
    println!("Found {} group(s)", groups.len());

    Ok(())
}
```

- Pour plus de détails sur l'API, voir DescribeAutoScalingGroupsla section de référence de l'API AWS SDK for Rust.

Supprimer un groupe Auto Scaling à l'aide d'un AWS SDK

Les exemples de code suivants montrent comment utiliserDeleteAutoScalingGroup.

.NET

AWS SDK for .NET

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

Mettez à jour la taille minimale d'un groupe Auto Scaling à zéro, mettez fin à toutes les instances du groupe et supprimez le groupe.

```
    /// <summary>
    /// Try to terminate an instance by its Id.
    /// </summary>
    /// <param name="instanceId">The Id of the instance to terminate.</param>
    /// <returns>Async task.</returns>
    public async Task TryTerminateInstanceById(string instanceId)
    {
        var stopping = false;
        Console.WriteLine($"Stopping {instanceId}...");
        while (!stopping)
        {
            try
            {
```

```
                await
_amazonAutoScaling.TerminateInstanceInAutoScalingGroupAsync(
                    new TerminateInstanceInAutoScalingGroupRequest()
                    {
                        InstanceId = instanceId,
                        ShouldDecrementDesiredCapacity = false
                    });
                stopping = true;
            }
            catch (ScalingActivityInProgressException)
            {
                Console.WriteLine($"Scaling activity in progress for
{instanceId}. Waiting...");
                Thread.Sleep(10000);
            }
        }
    }

    /// <summary>
    /// Tries to delete the EC2 Auto Scaling group. If the group is in use or in
progress,
    /// waits and retries until the group is successfully deleted.
    /// </summary>
    /// <param name="groupName">The name of the group to try to delete.</param>
    /// <returns>Async task.</returns>
    public async Task TryDeleteGroupByName(string groupName)
    {
        var stopped = false;
        while (!stopped)
        {
            try
            {
                await _amazonAutoScaling.DeleteAutoScalingGroupAsync(
                    new DeleteAutoScalingGroupRequest()
                    {
                        AutoScalingGroupName = groupName
                    });
                stopped = true;
            }
            catch (Exception e)
                when ((e is ScalingActivityInProgressException)
                      || (e is Amazon.AutoScaling.Model.ResourceInUseException))
            {
```

```
                Console.WriteLine($"Some instances are still running.
Waiting...");
                Thread.Sleep(10000);
            }
        }
    }

    /// <summary>
    /// Terminate instances and delete the Auto Scaling group by name.
    /// </summary>
    /// <param name="groupName">The name of the group to delete.</param>
    /// <returns>Async task.</returns>
    public async Task TerminateAndDeleteAutoScalingGroupWithName(string
groupName)
    {
        var describeGroupsResponse = await
_amazonAutoScaling.DescribeAutoScalingGroupsAsync(
            new DescribeAutoScalingGroupsRequest()
            {
                AutoScalingGroupNames = new List<string>() { groupName }
            });
        if (describeGroupsResponse.AutoScalingGroups.Any())
        {
            // Update the size to 0.
            await _amazonAutoScaling.UpdateAutoScalingGroupAsync(
                new UpdateAutoScalingGroupRequest()
                {
                    AutoScalingGroupName = groupName,
                    MinSize = 0
                });
            var group = describeGroupsResponse.AutoScalingGroups[0];
            foreach (var instance in group.Instances)
            {
                await TryTerminateInstanceById(instance.InstanceId);
            }

            await TryDeleteGroupByName(groupName);
        }
        else
        {
            Console.WriteLine($"No groups found with name {groupName}.");
        }
    }
```

```
    /// <summary>
    /// Delete an Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Amazon EC2 Auto Scaling group.</
param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> DeleteAutoScalingGroupAsync(
        string groupName)
    {
        var deleteAutoScalingGroupRequest = new DeleteAutoScalingGroupRequest
        {
            AutoScalingGroupName = groupName,
            ForceDelete = true,
        };

        var response = await
 _amazonAutoScaling.DeleteAutoScalingGroupAsync(deleteAutoScalingGroupRequest);
        if (response.HttpStatusCode == System.Net.HttpStatusCode.OK)
        {
            Console.WriteLine($"You successfully deleted {groupName}");
            return true;
        }

        Console.WriteLine($"Couldn't delete {groupName}.");
        return false;
    }
```

- Pour plus de détails sur l'API, reportez-vous DeleteAutoScalingGroupà la section Référence des AWS SDK for .NET API.

C++

SDK pour C++

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

            Aws::AutoScaling::Model::DeleteAutoScalingGroupRequest request;
            request.SetAutoScalingGroupName(groupName);

            Aws::AutoScaling::Model::DeleteAutoScalingGroupOutcome outcome =
                    autoScalingClient.DeleteAutoScalingGroup(request);

            if (outcome.IsSuccess()) {
                std::cout << "Auto Scaling group '" << groupName << "' was
deleted."
                          << std::endl;
            }
            else {
                std::cerr << "Error with AutoScaling::DeleteAutoScalingGroup. "
                          << outcome.GetError().GetMessage()
                          << std::endl;
                result = false;
            }
        }
```

- Pour plus de détails sur l'API, reportez-vous DeleteAutoScalingGroupà la section Référence des AWS SDK for C++ API.

CLI

AWS CLI

Exemple 1 : pour supprimer le groupe Auto Scaling spécifié

Cet exemple supprime le groupe Auto Scaling spécifié.

```
aws autoscaling delete-auto-scaling-group \
    --auto-scaling-group-name my-asg
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, consultez Supprimer votre infrastructure Auto Scaling dans le guide de l'utilisateur d'Amazon EC2 Auto Scaling.

Exemple 2 : Pour forcer la suppression du groupe Auto Scaling spécifié

Pour supprimer le groupe Auto Scaling sans attendre que les instances du groupe se terminent, utilisez l'`--force-delete`option.

```
aws autoscaling delete-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --force-delete
```

Cette commande ne produit aucun résultat.

Pour plus d'informations, consultez Supprimer votre infrastructure Auto Scaling dans le guide de l'utilisateur d'Amazon EC2 Auto Scaling.

- Pour plus de détails sur l'API, reportez-vous DeleteAutoScalingGroupà la section Référence des AWS CLI commandes.

Java

SDK pour Java 2.x

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.autoscaling.AutoScalingClient;
import software.amazon.awssdk.services.autoscaling.model.AutoScalingException;
import
 software.amazon.awssdk.services.autoscaling.model.DeleteAutoScalingGroupRequest;

/**
 * Before running this SDK for Java (v2) code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 */
public class DeleteAutoScalingGroup {
    public static void main(String[] args) {
        final String usage = """

                Usage:
                    <groupName>

                Where:
                    groupName - The name of the Auto Scaling group.
                """;

        if (args.length != 1) {
            System.out.println(usage);
            System.exit(1);
        }

        String groupName = args[0];
        AutoScalingClient autoScalingClient = AutoScalingClient.builder()
                .region(Region.US_EAST_1)
                .build();

        deleteAutoScalingGroup(autoScalingClient, groupName);
        autoScalingClient.close();
    }

    public static void deleteAutoScalingGroup(AutoScalingClient
 autoScalingClient, String groupName) {
        try {
```

```
            DeleteAutoScalingGroupRequest deleteAutoScalingGroupRequest =
 DeleteAutoScalingGroupRequest.builder()
                    .autoScalingGroupName(groupName)
                    .forceDelete(true)
                    .build();


 autoScalingClient.deleteAutoScalingGroup(deleteAutoScalingGroupRequest);
            System.out.println("You successfully deleted " + groupName);

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
}
```

- Pour plus de détails sur l'API, reportez-vous DeleteAutoScalingGroupà la section Référence des AWS SDK for Java 2.x API.

Kotlin

SDK pour Kotlin

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
suspend fun deleteSpecificAutoScalingGroup(groupName: String) {
    val deleteAutoScalingGroupRequest =
        DeleteAutoScalingGroupRequest {
            autoScalingGroupName = groupName
            forceDelete = true
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.deleteAutoScalingGroup(deleteAutoScalingGroupRequest)
        println("You successfully deleted $groupName")
```

```
    }
}
```

- Pour plus de détails sur l'API, reportez-vous DeleteAutoScalingGroupà la section AWS SDK pour la référence de l'API Kotlin.

PHP

Kit SDK pour PHP

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
    public function deleteAutoScalingGroup($autoScalingGroupName)
    {
        return $this->autoScalingClient->deleteAutoScalingGroup([
            'AutoScalingGroupName' => $autoScalingGroupName,
            'ForceDelete' => true,
        ]);
    }
```

- Pour plus de détails sur l'API, reportez-vous DeleteAutoScalingGroupà la section Référence des AWS SDK for PHP API.

PowerShell

Outils pour PowerShell

Exemple 1 : Cet exemple supprime le groupe Auto Scaling spécifié s'il ne possède aucune instance en cours d'exécution. Vous êtes invité à confirmer avant que l'opération ne se poursuive.

```
Remove-ASAutoScalingGroup -AutoScalingGroupName my-asg
```

Sortie :

```
Confirm
Are you sure you want to perform this action?
Performing operation "Remove-ASAutoScalingGroup (DeleteAutoScalingGroup)" on
 Target "my-asg".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemple 2 : Si vous spécifiez le paramètre Force, aucune confirmation ne vous est demandée avant le début de l'opération.

```
Remove-ASAutoScalingGroup -AutoScalingGroupName my-asg -Force
```

Exemple 3 : Cet exemple supprime le groupe Auto Scaling spécifié et met fin à toutes les instances en cours d'exécution qu'il contient.

```
Remove-ASAutoScalingGroup -AutoScalingGroupName my-asg -ForceDelete $true -Force
```

- Pour plus de détails sur l'API, reportez-vous DeleteAutoScalingGroupà la section Référence des AWS Tools for PowerShell applets de commande.

Python

SDK pour Python (Boto3)

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

Mettez à jour la taille minimale d’un groupe Auto Scaling à zéro, mettez fin à toutes les instances du groupe et supprimez le groupe.

```
class AutoScaler:
    """
    Encapsulates Amazon EC2 Auto Scaling and EC2 management actions.
    """
```

```
    def __init__(
        self,
        resource_prefix,
        inst_type,
        ami_param,
        autoscaling_client,
        ec2_client,
        ssm_client,
        iam_client,
    ):
        """
        :param resource_prefix: The prefix for naming AWS resources that are
created by this class.
        :param inst_type: The type of EC2 instance to create, such as t3.micro.
        :param ami_param: The Systems Manager parameter used to look up the AMI
that is
                          created.
        :param autoscaling_client: A Boto3 EC2 Auto Scaling client.
        :param ec2_client: A Boto3 EC2 client.
        :param ssm_client: A Boto3 Systems Manager client.
        :param iam_client: A Boto3 IAM client.
        """
        self.inst_type = inst_type
        self.ami_param = ami_param
        self.autoscaling_client = autoscaling_client
        self.ec2_client = ec2_client
        self.ssm_client = ssm_client
        self.iam_client = iam_client
        self.launch_template_name = f"{resource_prefix}-template"
        self.group_name = f"{resource_prefix}-group"
        self.instance_policy_name = f"{resource_prefix}-pol"
        self.instance_role_name = f"{resource_prefix}-role"
        self.instance_profile_name = f"{resource_prefix}-prof"
        self.bad_creds_policy_name = f"{resource_prefix}-bc-pol"
        self.bad_creds_role_name = f"{resource_prefix}-bc-role"
        self.bad_creds_profile_name = f"{resource_prefix}-bc-prof"
        self.key_pair_name = f"{resource_prefix}-key-pair"


    def _try_terminate_instance(self, inst_id):
        stopping = False
        log.info(f"Stopping {inst_id}.")
        while not stopping:
            try:
```

```
                self.autoscaling_client.terminate_instance_in_auto_scaling_group(
                    InstanceId=inst_id, ShouldDecrementDesiredCapacity=True
                )
                stopping = True
            except ClientError as err:
                if err.response["Error"]["Code"] == "ScalingActivityInProgress":
                    log.info("Scaling activity in progress for %s. Waiting...",
inst_id)
                    time.sleep(10)
                else:
                    raise AutoScalerError(f"Couldn't stop instance {inst_id}:
{err}.")

    def _try_delete_group(self):
        """
        Tries to delete the EC2 Auto Scaling group. If the group is in use or in
progress,
        the function waits and retries until the group is successfully deleted.
        """
        stopped = False
        while not stopped:
            try:
                self.autoscaling_client.delete_auto_scaling_group(
                    AutoScalingGroupName=self.group_name
                )
                stopped = True
                log.info("Deleted EC2 Auto Scaling group %s.", self.group_name)
            except ClientError as err:
                if (
                    err.response["Error"]["Code"] == "ResourceInUse"
                    or err.response["Error"]["Code"] ==
"ScalingActivityInProgress"
                ):
                    log.info(
                        "Some instances are still running. Waiting for them to
stop..."
                    )
                    time.sleep(10)
                else:
                    raise AutoScalerError(
                        f"Couldn't delete group {self.group_name}: {err}."
                    )

    def delete_group(self):
```

```
        """
        Terminates all instances in the group, deletes the EC2 Auto Scaling
 group.
        """
        try:
            response = self.autoscaling_client.describe_auto_scaling_groups(
                AutoScalingGroupNames=[self.group_name]
            )
            groups = response.get("AutoScalingGroups", [])
            if len(groups) > 0:
                self.autoscaling_client.update_auto_scaling_group(
                    AutoScalingGroupName=self.group_name, MinSize=0
                )
                instance_ids = [inst["InstanceId"] for inst in groups[0]
["Instances"]]
                for inst_id in instance_ids:
                    self._try_terminate_instance(inst_id)
                self._try_delete_group()
            else:
                log.info("No groups found named %s, nothing to do.",
 self.group_name)
        except ClientError as err:
            raise AutoScalerError(f"Couldn't delete group {self.group_name}:
 {err}.")
```

- Pour plus de détails sur l'API, consultez DeleteAutoScalingGrouple AWS manuel de référence de l'API SDK for Python (Boto3).

Rust

SDK pour Rust

> **Note**
>
> Il y en a plus sur GitHub. Trouvez l'exemple complet et découvrez comment le configurer et l'exécuter dans le référentiel d'exemples de code AWS.

```
async fn delete_group(client: &Client, name: &str, force: bool) -> Result<(),
 Error> {
    client
        .delete_auto_scaling_group()
        .auto_scaling_group_name(name)
        .set_force_delete(if force { Some(true) } else { None })
        .send()
        .await?;

    println!("Deleted Auto Scaling group");

    Ok(())
}
```

- Pour plus de détails sur l'API, voir DeleteAutoScalingGroupla section de référence de l'API AWS SDK for Rust.

Recycler les instances de votre groupe Auto Scaling

Amazon EC2 Auto Scaling propose des fonctionnalités qui vous permettent de remplacer les instances Amazon EC2 de votre groupe Auto Scaling après avoir effectué des mises à jour, telles que l'ajout d'un nouveau modèle de lancement par une nouvelle Amazon Machine Image (AMI) ou l'ajout de nouveaux types d'instances. Il vous permet également de rationaliser les mises à jour en vous donnant la possibilité de les inclure dans la même opération que celle qui remplace les instances.

Cette section contient des informations qui vous aideront à effectuer les opérations suivantes :

- Lancer une actualisation d'instance pour remplacer des instances de votre groupe Auto Scaling
- Déclarer des mises à jour spécifiques décrivant une configuration souhaitée et mettre à jour le groupe Auto Scaling en fonction de cette configuration
- Ignorer le remplacement des instances déjà mises à jour
- Utilisez les points de contrôle pour mettre à jour les instances par étapes et effectuez des vérifications sur vos instances à des points spécifiques.
- Recevoir des notifications par e-mail lorsqu'un point de contrôle est atteint
- Utilisez une restauration pour restaurer le groupe Auto Scaling à la configuration qu'il utilisait précédemment.
- Annulation automatique si l'actualisation de l'instance échoue pour une raison ou une autre ou si l'une des CloudWatch alarmes Amazon que vous spécifiez passe à l'`ALARM`état initial.
- Limiter la durée de vie des instances pour assurer la cohérence des versions logicielles et des configurations d'instance au sein du groupe Auto Scaling.

Table des matières

Utiliser une actualisation d'instance pour mettre à jour les instances d'un groupe Auto Scaling

Vous pouvez utiliser une actualisation d'instance pour mettre à jour les instances de votre groupe Auto Scaling. Cette fonctionnalité peut être utile lorsqu'une modification de configuration nécessite

le remplacement d'instances, en particulier si votre groupe Auto Scaling contient un grand nombre d'instances.

Parmi les situations dans lesquelles l'actualisation d'une instance peut être utile, citons :

- Déploiement d'une nouvelle Amazon Machine Image (AMI) ou d'un nouveau script de données utilisateur au sein de votre groupe Auto Scaling. Vous pouvez créer un nouveau modèle de lancement avec les modifications, puis utiliser une actualisation d'instance pour déployer les mises à jour immédiatement.
- Migration de vos instances vers de nouveaux types d'instances pour tirer parti des dernières améliorations et optimisations.
- Faire passer vos groupes Auto Scaling d'une configuration de lancement à un modèle de lancement. Vous pouvez copier vos configurations de lancement dans des modèles de lancement, puis utiliser une actualisation d'instance pour mettre à jour vos instances avec les nouveaux modèles. Pour en savoir plus sur la migration vers des modèles de lancement, consultez Migrez vos groupes Auto Scaling pour lancer des modèles.

Table des matières

Fonctionnement d'une actualisation d'instance

Cette rubrique décrit le fonctionnement d'une actualisation d'instance et présente les concepts clés que vous devez comprendre pour l'utiliser efficacement.

Table des matières

- Période de grâce de surveillance de l'état
- Compatibilité des types d'instance
- Limites

Comment ça marche

Pour actualiser les instances d'un groupe Auto Scaling, vous pouvez définir une nouvelle configuration contenant la dernière version de votre application et toutes les autres mises à jour que vous souhaitez apporter. Lancez ensuite une actualisation de l'instance pour remplacer les instances existantes par de nouvelles en fonction de cette configuration.

Pour actualiser une instance :

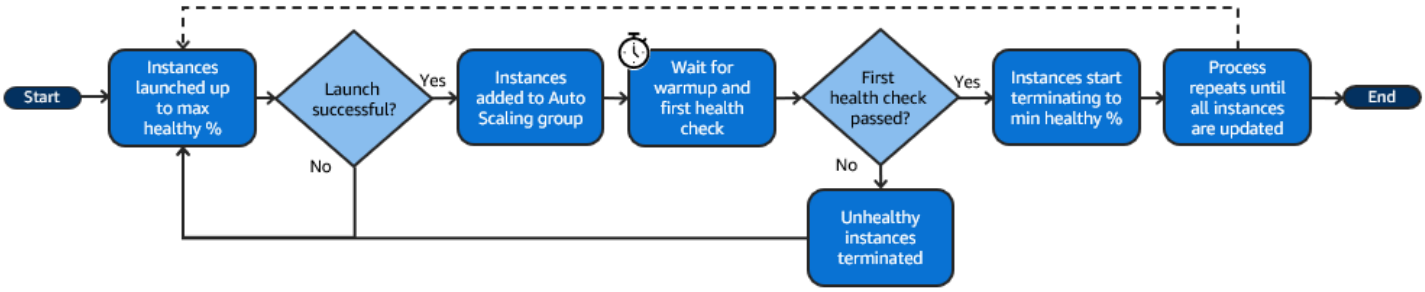
1. Créez un nouveau modèle de lancement ou mettez à jour le modèle existant avec les modifications de configuration souhaitées, telles qu'une nouvelle Amazon Machine Image (AMI). Pour plus d'informations, consultez Créer un modèle de lancement pour un groupe Auto Scaling.
2. Démarrez l'actualisation de l'instance à l'aide de la console Amazon EC2 Auto Scaling AWS CLI, ou du SDK :
   - Spécifiez le nouveau modèle de lancement ou la version du modèle de lancement que vous avez créé. Cela sera utilisé pour lancer de nouvelles instances.
   - Définissez les pourcentages de santé minimum et maximum préférés. Cela permet de contrôler le nombre d'instances remplacées simultanément et de savoir si de nouvelles instances sont lancées avant de mettre fin aux anciennes.
   - Configurez tous les paramètres facultatifs, tels que :
     - Points de contrôle : interrompez l'actualisation de l'instance après un certain pourcentage de remplacements pour vérifier la progression.
     - Ignorer la correspondance : comparez les anciennes instances à la nouvelle configuration et ne remplacez que celles qui ne correspondent pas. Lorsque vous lancez une actualisation d'instance depuis la console, la mise en correspondance par défaut est activée par défaut.
     - Types d'instances multiples : appliquez une politique d'instances mixtes nouvelle ou mise à jour dans le cadre de la configuration souhaitée.

Une fois l'actualisation de l'instance lancée, Amazon EC2 Auto Scaling va :

- Remplacez les instances par lots en fonction des pourcentages sains minimum et maximum.

- Lancez d'abord les nouvelles instances avant de mettre fin aux anciennes si le pourcentage de santé minimum est défini sur 100 %. Cela garantit que la capacité souhaitée est maintenue à tout moment.
- Vérifiez l'état de santé des instances et donnez-leur le temps de se réchauffer avant que d'autres instances ne soient remplacées.
- Arrêtez et remplacez les instances qui s'avèrent défectueuses.
- Mettez automatiquement à jour les paramètres du groupe Auto Scaling avec les nouvelles modifications de configuration une fois l'actualisation de l'instance réussie.
- Remplacez `InService` les instances avant les instances qui se trouvent dans un pool chaud.

L'organigramme suivant illustre le comportement du lancement avant la fin lorsque vous définissez le pourcentage de santé minimum à 100 %.



> **Note**
>
> Les pourcentages de santé minimum et maximum pour une actualisation d'instance doivent uniquement être spécifiés si vous n'avez pas défini de politique de maintenance d'instance ou si vous devez remplacer la politique existante. Pour plus d'informations, consultez Politiques de maintenance des instances.
>
> De même, vous devez uniquement spécifier la période de préchauffage de l'instance pour une actualisation d'instance si vous n'avez pas activé le préchauffage par défaut ou si vous devez remplacer le préchauffage par défaut. Pour plus d'informations, consultez Définir la préparation par défaut d'instance d'un groupe Auto Scaling.

Concepts de base

Avant de commencer, familiarisez-vous avec les concepts de base de l'actualisation d'instance :

Pourcentage minimal d’intégrité

Le pourcentage d'intégrité minimum est le pourcentage de la capacité souhaitée pour rester en service, saine et prête à être utilisée lors de l'actualisation d'une instance afin que l'actualisation puisse se poursuivre. Par exemple, si le pourcentage minimal d’intégrité est de 90 %, et le pourcentage maximal d’intégrité est de 100 %, 10 % est le pourcentage de capacité qui sera résilié et remplacé. Si les nouvelles instances échouent aux surveillances de l'état, Amazon EC2 Auto Scaling les résilie et les remplace. Si l’actualisation d’instance ne peut lancer aucune instance saine, elle échouera, laissant les 90 % restants du groupe intacts. Si les nouvelles instances restent saines et terminent leur période de préchauffage, Amazon EC2 Auto Scaling peut continuer à remplacer d'autres instances.

Une actualisation d'instance peut remplacer les instances une par une, plusieurs à la fois ou toutes à la fois. Pour remplacer une instance à la fois, définissez le pourcentage minimal et maximal d’instances saines sur 100 %. Cela modifie le comportement d’une actualisation d’instance à lancer avant toute résiliation, ce qui empêche la capacité du groupe de tomber en dessous de 100 % de la capacité souhaitée. Pour remplacer toutes les instances à la fois, définissez le pourcentage minimal d’instances saines sur 0 %.

Pourcentage maximal d’intégrité

Le pourcentage maximal d’intégrité est le pourcentage de la capacité souhaitée que votre groupe Auto Scaling peut atteindre lors du remplacement d’instances. La différence entre le minimum et le maximum ne peut pas être supérieure à 100. Une plage étendue augmente le nombre d’instances qui peuvent être remplacées en même temps.

Préparation d’instance

La préparation d’instance correspond à la période qui sépare le moment où l’état de la nouvelle instance est remplacé par `InService` au moment où elle est considérée comme ayant fini son initialisation. Lors d'une actualisation d'instance, si les instances réussissent les surveillances de l'état, Amazon EC2 Auto Scaling ne passe pas immédiatement au remplacement de l'instance suivante après avoir déterminé qu'une instance nouvellement lancée est saine. Il attend la période de préchauffage avant de passer au remplacement de l'instance suivante. Cela peut être utile lorsque votre application a encore besoin d’un certain temps d’initialisation avant de répondre aux demandes.

La préparation d’instance fonctionne de la même manière que la préparation d’instance par défaut. Par conséquent, les mêmes considérations de mise à l’échelle s’appliquent. Pour plus d’informations, consultez Définir la préparation par défaut d'instance d'un groupe Auto Scaling.

Configuration souhaitée

La configuration souhaitée désigne la nouvelle configuration que vous souhaitez qu'Amazon EC2 Auto Scaling déploie dans votre groupe Auto Scaling. Par exemple, vous pouvez spécifier un nouveau modèle de lancement et de nouveaux types d'instance pour vos instances. Lors d'une actualisation d'instance, Amazon EC2 Auto Scaling met à jour le groupe Auto Scaling en fonction de la configuration souhaitée. Si un événement de montée en puissance se produit lors d'une actualisation d'instance, Amazon EC2 Auto Scaling lance de nouvelles instances avec la configuration souhaitée au lieu d'utiliser les paramètres actuels du groupe. Une fois l'actualisation d'instance réussie, Amazon EC2 Auto Scaling met à jour les paramètres du groupe Auto Scaling pour refléter la nouvelle configuration souhaitée que vous avez spécifiée dans le cadre de l'actualisation d'instance.

Ignorer la correspondance

La fonction Ignorer la correspondance indique à Amazon EC2 Auto Scaling d'ignorer les instances qui disposent déjà de vos dernières mises à jour. De cette façon, vous ne remplacez pas plus d'instances que nécessaire. Cela est utile quand vous souhaitez vous assurer que votre groupe Auto Scaling utilise une version particulière de votre modèle de lancement et ne remplace que les instances qui utilisent une version différente.

Points de contrôle

Un point de contrôle désigne un moment où l'actualisation d'instance s'interrompt pour une durée déterminée. Une actualisation d'instance peut contenir plusieurs points de contrôle. Amazon EC2 Auto Scaling émet des événements pour chaque point de contrôle. Par conséquent, vous pouvez ajouter une EventBridge règle pour envoyer les événements à une cible, telle qu'Amazon SNS, afin qu'elle soit avertie lorsqu'un point de contrôle est atteint. Une fois qu'un point de contrôle est atteint, vous avez la possibilité de vérifier votre déploiement. Si des problèmes sont identifiés, vous pouvez annuler l'actualisation d'instance ou la restaurer. La possibilité de déployer les mises à jour par phases est un avantage clé des points de contrôle. Si vous n'utilisez pas de points de contrôle, des remplacements sont effectués en continu.

Pour en savoir plus sur tous les paramètres par défaut que vous pouvez configurer lors du lancement d’une actualisation d’instance, consultez Comprendre les valeurs par défaut d'une actualisation d'instance.

Période de grâce de surveillance de l'état

Amazon EC2 Auto Scaling détermine si une instance est saine en fonction du statut des surveillances de l'état que votre groupe Auto Scaling utilise. Pour plus d'informations, consultez Surveillance de l'état des instances dans un groupe Auto Scaling.

Pour vous assurer que ces surveillances de l'état commencent le plus rapidement possible, ne définissez pas une période de grâce de la surveillance de l'état du groupe trop élevée, mais suffisamment élevée pour que vos surveillances de l'état Elastic Load Balancing déterminent si une cible est disponible pour traiter les demandes. Pour plus d'informations, consultez Définir la période de grâce de la surveillance de l'état pour un groupe Auto Scaling.

Compatibilité des types d'instance

Avant de modifier votre type d'instance, il est conseillé de vérifier qu'il fonctionne avec votre modèle de lancement. Cela confirme la compatibilité avec l'AMI que vous avez spécifiée. Par exemple, disons que vous avez lancé vos instances d'origine à partir d'une AMI paravirtuelle (PV), mais que vous souhaitez passer à un type d'instance de génération actuelle qui n'est pris en charge que par une AMI de machine virtuelle matérielle (HVM). Dans ce cas, vous devez utiliser une AMI HVM dans votre modèle de lancement.

Pour confirmer la compatibilité du type d'instance sans lancer d'instances, utilisez la commande run-instances avec l'option `--dry-run`, comme indiqué dans l'exemple suivant.

```
aws ec2 run-instances --launch-template LaunchTemplateName=my-template,Version='1' --
dry-run
```

Pour plus d'informations sur la manière dont la compatibilité est déterminée, consultez la section Compatibilité pour la modification du type d'instance dans le guide de l'utilisateur Amazon EC2.

Limites

- Durée totale : la durée maximale pendant laquelle une actualisation d'instance peut continuer à remplacer activement des instances est de 14 jours.
- Différence de comportement spécifique aux groupes pondérés : si un groupe d'instances mixtes est configuré avec un poids d'instance supérieur ou égal à la capacité souhaitée du groupe, Amazon EC2 Auto Scaling peut remplacer toutes les instances `InService` à la fois. Pour éviter cette situation, suivez les recommandations de la rubrique Configurer un groupe Auto Scaling pour

utiliser les poids d'instance. Précisez une capacité souhaitée supérieure à votre poids maximal lorsque vous utilisez des poids avec votre groupe Auto Scaling.

- Délai d'expiration d'une heure : lorsqu'une actualisation d'instance ne peut pas continuer à effectuer des remplacements parce qu'elle attend des instances en veille ou protégées contre la mise à l'échelle horizontale, ou que les nouvelles instances échouent aux surveillances de l'état, Amazon EC2 Auto Scaling réessaie encore pendant une heure. Un message de statut est également fourni pour vous aider à résoudre le problème. Si le problème persiste au bout d'une heure, l'opération échoue. L'objectif est de lui donner le temps de récupérer en cas de problème temporaire.
- Déploiement de code via les données utilisateur : Skip matching ne vérifie pas les modifications de code déployées à partir d'un script de données utilisateur. Si vous utilisez les données utilisateur pour extraire du nouveau code et installer ces mises à jour sur de nouvelles instances, nous vous recommandons de désactiver la mise en correspondance afin de vous assurer que toutes les instances reçoivent votre dernier code, même sans mise à jour de la version du modèle de lancement.
- Restriction de mise à jour : si vous tentez de mettre à jour le modèle de lancement, la configuration de lancement ou la politique d'instances mixtes d'un groupe Auto Scaling alors qu'une actualisation d'instance avec la configuration souhaitée est active, la demande échouera avec l'erreur de validation suivante : `An active instance refresh with a desired configuration exists. All configuration options derived from the desired configuration are not available for update while the instance refresh is active.`

Comprendre les valeurs par défaut d'une actualisation d'instance

Avant de commencer une actualisation d'instance, vous pouvez personnaliser les diverses préférences qui affectent l'actualisation d'instance. Certaines préférences par défaut sont différentes selon que vous utilisez la console ou la ligne de commande (AWS CLI ou le AWS SDK).

Le tableau suivant répertorie les valeurs par défaut des paramètres d'actualisation d'instance.

Paramètre	AWS CLI ou AWS SDK	Console Amazon EC2 Auto Scaling
CloudWatch alarme	Désactivé (null)	Désactivées
Restauration automatique	Désactivé (`false`)	Désactivées

Paramètre	AWS CLI ou AWS SDK	Console Amazon EC2 Auto Scaling
Points de contrôle	Désactivé (`false`)	Désactivées
Délai de contrôle	1 heure (`3600` secondes)	1 heure
Préparation d'instance	La préparation d'instance par défaut, si elle est définie, ou la période de grâce de la surveillance de l'état dans le cas contraire.	La préparation d'instance par défaut, si elle est définie, ou la période de grâce de la surveillance de l'état dans le cas contraire.
Pourcentage maximal d'intégrité	Varie en fonction de votre politique de maintenance des instances. En l'absence de politique de maintenance des instances, la valeur par défaut est de 100 % (null).	Varie en fonction de votre politique de maintenance des instances. En l'absence de politique de maintenance des instances, la valeur par défaut est de 100 % (null).
Pourcentage minimal d'intégrité	Varie en fonction de votre politique de maintenance des instances. En l'absence de politique de maintenance des instances, la valeur par défaut est de 90 %.	Varie en fonction de votre politique de maintenance des instances. En l'absence de politique de maintenance des instances, la valeur par défaut est de 90 %.
Instances protégées contre la mise à l'échelle horizontale	Attente	Ignorer
Ignorer la correspondance	Désactivé (`false`)	Activées
Instances en veille	Attente	Ignorer

Voici une description de chaque paramètre :

CloudWatch alarme (**`AlarmSpecification`**)

La spécification de l' CloudWatch alarme. CloudWatch les alarmes peuvent être utilisées pour identifier tout problème et faire échouer le fonctionnement si une alarme passe à l'`ALARM`état. Pour plus d’informations, consultez Lancer une actualisation d'instance avec restauration automatique.

Restauration automatique (**`AutoRollback`**)

Contrôle si Amazon EC2 Auto Scaling restaure ou non la configuration précédente du groupe Auto Scaling en cas d’échec de l’actualisation d’instance. Pour plus d’informations, consultez Annuler les modifications avec une restauration.

Points de contrôle (**`CheckpointPercentages`**)

Contrôle si Amazon EC2 Auto Scaling remplace les instances par phases. Cela est utile si vous devez effectuer des vérifications sur vos instances avant de les remplacer toutes. Pour plus d’informations, consultez Ajouter des points de contrôle à une actualisation d'instance.

Délai de contrôle (**`CheckpointDelay`**)

Délai d'attente, en secondes, après avoir atteint un point de contrôle avant de poursuivre. Pour plus d’informations, consultez Ajouter des points de contrôle à une actualisation d'instance.

Préparation d'instance (**`InstanceWarmup`**)

Période, en secondes, pendant laquelle Amazon EC2 Auto Scaling attend qu’une nouvelle instance soit considérée comme ayant fini son initialisation avant de passer au remplacement de l’instance suivante. Si vous avez déjà défini correctement une préparation d'instance par défaut pour le groupe Auto Scaling, vous n'avez pas besoin de modifier la préparation d'instance (sauf si vous voulez remplacer la valeur par défaut). Pour plus d’informations, consultez Définir la préparation par défaut d'instance d'un groupe Auto Scaling.

Pourcentage maximal d’intégrité (**`MaxHealthyPercentage`**)

Le pourcentage de la capacité souhaitée du groupe Auto Scaling que votre groupe peut atteindre lors du remplacement d’instances.

Pourcentage minimal d'instances saines (**`MinHealthyPercentage`**)

Pourcentage de la capacité souhaitée du groupe Auto Scaling qui doit être en service, en bon état et prête à être utilisée avant de pouvoir poursuivre l’opération.

Instances protégées contre la mise à l'échelle horizontale (**`ScaleInProtectedInstances`**)

Contrôle ce que fait Amazon EC2 Auto Scaling si des instances protégées contre la mise à l'échelle horizontale sont détectées. Pour plus d'informations sur ces instances, consultez Utiliser la protection de la taille d'instance.

Amazon EC2 Auto Scaling fournit les options suivantes :

- Replace (`Refresh`) — Remplace les instances protégées contre la mise à l'échelle.
- Ignorer (`Ignore`) — Ignore les instances protégées contre le dimensionnement et continue de remplacer les instances qui ne le sont pas.
- Attendre (`Wait`) : attend une heure pour que vous supprimiez la protection scale-in. Si vous ne le faites pas, l'actualisation d'instance échoue.

Ignorer la correspondance (**`SkipMatching`**)

Contrôle si Amazon EC2 Auto Scaling ignore le remplacement des instances qui correspondent à la configuration souhaitée. Si aucune configuration souhaitée n'est spécifiée, il ignore le remplacement des instances dont le modèle de lancement et les types d'instance sont identiques à ceux utilisés par le groupe Auto Scaling avant le lancement de l'actualisation d'instance. Pour plus d'informations, consultez Utiliser une actualisation d'instance avec la fonction Ignorer la correspondance.

Instances en veille (**`StandbyInstances`**)

Contrôle ce que fait Amazon EC2 Auto Scaling si des instances sont à l'état `Standby`. Pour plus d'informations sur ces instances, consultez Supprimer temporairement des instances du groupe Auto Scaling.

Amazon EC2 Auto Scaling fournit les options suivantes :

- Terminate (`Terminate`) — Met fin aux instances présentes dans`Standby`.
- Ignorer (`Ignore`) — Ignore les instances présentes dans l'état `Standby` et continue de remplacer les instances présentes dans `InService` cet état.
- Attendre (`Wait`) : attend une heure avant que vous remettiez les instances en service. Si vous ne le faites pas, l'actualisation d'instance échoue.

Lancer une actualisation d’instance

> **Important**
>
> Vous pouvez restaurer une actualisation d'instance en cours afin d'annuler toutes les modifications. Pour que cela fonctionne, le groupe Auto Scaling doit remplir les conditions préalables à l'utilisation des restaurations avant de lancer l'actualisation d'instance. Pour plus d’informations, consultez Annuler les modifications avec une restauration.

Les procédures suivantes vous aident à démarrer une actualisation d'instance à l'aide du AWS Management Console ou AWS CLI.

Lancer une actualisation d'instance (console)

Si c'est la première fois que vous démarrez une actualisation d'instance, l'utilisation de la console vous aidera à comprendre les fonctions et les options disponibles.

Lancer une actualisation d'instance dans la console (procédure de base)

Utilisez la procédure suivante si vous n'avez pas encore défini de politique d'instances mixtes pour votre groupe Auto Scaling. Si vous avez déjà défini une politique d'instances mixtes, consultez Lancer une actualisation d'instance dans la console (groupe d'instances mixtes) pour démarrer une actualisation d'instance.

Pour lancer une actualisation d’instance

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre dans la partie inférieure de la page Groupes Auto Scaling.
3. Dans l’onglet Actualisation d’instance, dans Actualisation d’instance active, sélectionnez Démarrer l’actualisation d’instance.
4. Pour les paramètres de disponibilité, procédez comme suit :

   a. Pour la méthode de remplacement des instances :

- Si vous n’avez pas défini de politique de maintenance des instances pour le groupe Auto Scaling, le paramètre par défaut pour la méthode de remplacement d’instance est Résilier et lancer. Il s’agit de l’ancien comportement par défaut d’une actualisation d’instance.
- Si vous définissez une politique de maintenance des instances sur le groupe Auto Scaling, elle fournit des valeurs par défaut pour la méthode de remplacement d’instance. Pour annuler la politique de maintenance des instances, choisissez Remplacer. Le remplacement s’applique uniquement à l’actualisation d’instance en cours. La prochaine fois que vous lancerez une actualisation d’instance, ces valeurs seront rétablies selon les valeurs par défaut de la politique de maintenance des instances.

La procédure suivante explique comment mettre à jour la méthode de remplacement d’instance.

i. Choisissez l’une des méthodes de remplacement d’instance suivantes :

   - Lancer avant toute résiliation : une nouvelle instance doit d’abord être mise en service avant qu’une instance existante puisse être résiliée. Il s’agit d’un bon choix pour les applications qui privilégient la disponibilité plutôt que les économies de coûts.
   - Résilier et lancer : les nouvelles instances sont mises en service en même temps que les instances existantes sont résiliées. Il s’agit d’un bon choix pour les applications qui privilégient les économies de coûts plutôt que la disponibilité. C’est également un bon choix pour les applications qui ne doivent pas libérer plus de capacité que ce qui est actuellement disponible.
   - Comportement personnalisé : cette option vous permet de définir une plage minimale et maximale personnalisée pour la quantité de capacité que vous souhaitez avoir à disposition lors du remplacement d’instances. Cela peut vous aider à trouver le juste équilibre entre le coût et la disponibilité.

ii. Pour Définir un pourcentage d’intégrité, saisissez des valeurs pour l’un ou les deux champs suivants. Les champs d’activation varient en fonction de l’option que vous choisissez pour la méthode de remplacement d’instance.

   - Min : définit le pourcentage minimal d’intégrité requis pour procéder à l’actualisation des instances.
   - Max : définit le pourcentage maximal d’intégrité possible lors de l’actualisation des instances.

iii. Développez la section Afficher la capacité temporaire estimée pendant les remplacements en fonction de la taille actuelle du groupe pour confirmer comment les valeurs de Min et Max s'appliquent à votre groupe. Les valeurs exactes utilisées dépendent de la valeur de capacité souhaitée, qui changera si le groupe est mis à l'échelle.

iv. Développez la section Définir un comportement de repli pour les tailles de remplacement non valides, puis choisissez de déroger au pourcentage maximal valide afin de prioriser la disponibilité ou de déroger au pourcentage minimal valide.

Il n'est pas recommandé de conserver l'option par défaut Déroger au pourcentage minimal valide pour les très petits groupes. Lorsque le groupe Auto Scaling ne contient qu'une seule instance, le lancement d'une actualisation d'instance peut provoquer une panne.

Cette étape configure le comportement de secours si vous utilisez un groupe Auto Scaling qui n'a pas encore de politique de maintenance des instances. Cette option n'est pas disponible et n'apparaît pas lorsque votre groupe dispose d'une politique de maintenance des instances. Cette option n'est également disponible que pour la méthode de remplacement Résilier et lancer. Les autres méthodes de remplacement dérogeront au pourcentage maximal valide afin de donner la priorité à la disponibilité.

b. Pour Préparation d'instance, saisissez le nombre de secondes à partir du moment où l'état d'une nouvelle instance change en `InService` au moment où elle finit l'initialisation. Amazon EC2 Auto Scaling attend ce laps de temps avant de passer au remplacement de l'instance suivante.

Lors de la préparation, une instance nouvellement lancée n'est pas non plus prise en compte dans les métriques d'instance agrégées du groupe Auto Scaling (telles que `CPUUtilization`, `NetworkIn` et `NetworkOut`). Si vous avez ajouté des politiques de mise à l'échelle au groupe Auto Scaling, les activités de mise à l'échelle s'exécutent en parallèle. Si vous définissez un intervalle long pour la période de préchauffage de l'actualisation des instances, les instances nouvellement lancées mettent plus de temps à apparaître dans les métriques. Par conséquent, une période de préchauffage adéquate empêche Amazon EC2 Auto Scaling de s'adapter à des données métriques périmées.

Si vous avez déjà défini correctement une préparation d'instance par défaut pour le groupe Auto Scaling, vous n'avez pas besoin de modifier la préparation d'instance. Toutefois, si vous souhaitez remplacer la valeur par défaut, vous pouvez définir une valeur pour cette

option. Pour plus d'informations sur la définition de la préparation d'instance par défaut, consultez Définir la préparation par défaut d'instance d'un groupe Auto Scaling.

5. Sous Actualiser les paramètres, procédez comme suit :

   a. (Facultatif) Pour Points de contrôle, sélectionnez Activer les points de contrôle pour remplacer les instances à l'aide d'une approche progressive ou par phase d'une actualisation d'instance. Cela donne plus de temps pour la vérification entre les séries de remplacements. Si vous choisissez de ne pas activer les points de contrôle, les instances sont remplacées en une seule opération quasi continue.

      Si vous activez les points de contrôle, consultez Activer les points de contrôle (console) pour connaître les étapes supplémentaires.

   b. Activez ou désactivez Ignorer la correspondance :

      - Pour ignorer le remplacement des instances qui correspondent déjà à votre modèle de lancement, laissez la case Activer la fonction Ignorer la correspondance cochée.
      - Si vous désactivez la fonction Ignorer la correspondance en décochant cette case, toutes les instances peuvent être remplacées.

      Lorsque vous activez la fonction Ignorer la correspondance, vous pouvez définir un nouveau modèle de lancement ou une nouvelle version du modèle de lancement. Vous pouvez le faire dans la section Configuration souhaitée de la page Démarrer l'actualisation d'instance.

      > **Note**
      >
      > Pour utiliser la fonctionnalité Ignorer la correspondance afin de mettre à jour un groupe Auto Scaling qui utilise actuellement une configuration de lancement, vous devez sélectionner un modèle de lancement dans Configuration souhaitée. La correspondance de saut avec une configuration de lancement n'est pas prise en charge.

   c. Pour les instances en veille, choisissez Ignorer, Résilier ou Attendre. Cela détermine ce qui se passe si des instances sont à l'état `Standby`. Pour plus d'informations, consultez Supprimer temporairement des instances du groupe Auto Scaling.

      Si vous choisissez Attendre, vous devez prendre des mesures supplémentaires pour remettre ces instances en service. Si vous ne le faites pas, l'actualisation d'instance

remplace toutes les instances `InService` et attend pendant une heure. Ensuite, s'il reste des instances `Standby`, l'actualisation d'instance échoue. Pour éviter cette situation, choisissez à la place Ignorer ou Résilier ces instances.

d. Pour les instances protégées contre la mise à l'échelle horizontale, choisissez Ignorer, Remplacer ou Attendre. Cela détermine ce qui se passe si des instances protégées contre la mise à l'échelle horizontale sont trouvées. Pour plus d'informations, consultez Utiliser la protection de la taille d'instance.

   Si vous choisissez Attendre, vous devez prendre des mesures supplémentaires pour supprimer la protection contre la mise à l'échelle horizontale de ces instances. Si vous ne le faites pas, l'actualisation d'instance remplace toutes les instances non protégées et attend pendant une heure. Ensuite, s'il reste des instances protégées contre la mise à l'échelle horizontale, l'actualisation d'instance échoue. Pour éviter cette situation, choisissez à la place Ignorer ou Remplacer ces instances.

6. (Facultatif) Pour les CloudWatch CloudWatch alarmes, choisissez Activer les alarmes, puis choisissez une ou plusieurs alarmes. CloudWatch les alarmes peuvent être utilisées pour identifier tout problème et faire échouer le fonctionnement si une alarme passe à l'`ALARM`état. Pour plus d'informations, consultez Lancer une actualisation d'instance avec restauration automatique.

7. (Facultatif) Développez la section Configuration souhaitée pour spécifier les mises à jour que vous souhaitez apporter à votre groupe Auto Scaling.

   Pour cette étape, vous pouvez choisir d'utiliser la syntaxe JSON ou YAML pour modifier les valeurs des paramètres au lieu de faire des sélections dans l'interface de la console. Pour ce faire, sélectionnez Utiliser l'éditeur de code au lieu de Utiliser l'interface de la console. La procédure suivante explique comment effectuer des sélections à l'aide de l'interface de la console.

   a. Pour Mettre à jour le modèle de lancement :

      - Si vous n'avez pas créé de nouveau modèle de lancement ou de nouvelle version de modèle de lancement pour votre groupe Auto Scaling, ne cochez pas cette case.
      - Si vous avez créé un nouveau modèle de lancement ou une nouvelle version du modèle de lancement, cochez cette case. Lorsque vous sélectionnez cette option, Amazon EC2 Auto Scaling affiche le modèle de lancement actuel et la version actuelle du modèle de lancement. Il répertorie également toutes les autres versions disponibles. Choisissez le modèle de lancement, puis la version.

Après avoir choisi une version, vous pouvez voir les informations de version. Il s'agit de la version du modèle de lancement qui sera utilisée lors du remplacement d'instances dans le cadre d'une actualisation d'instance. Si l'actualisation de l'instance réussit, cette version du modèle de lancement sera également utilisée chaque fois que de nouvelles instances seront lancées, par exemple, lorsque le groupe évolue.

b. Pour Choose a set of instance types and purchase options to override the instance type in the launch template (Choisir un ensemble de types d'instances et d'options d'achat pour remplacer le type d'instance dans le modèle de lancement) :

- Ne cochez pas cette case si vous voulez utiliser le type d'instance et l'option d'achat que vous avez spécifiés dans votre modèle de lancement.
- Cochez cette case si vous souhaitez remplacer le type d'instance dans le modèle de lancement ou exécuter des instances Spot. Vous pouvez soit ajouter manuellement chaque type d'instance, soit choisir un type d'instance principal et une option de recommandation qui récupère pour vous tous les types d'instance correspondants supplémentaires. Si vous envisagez de lancer des instances Spot, nous vous recommandons d'ajouter plusieurs types d'instance différents. Amazon EC2 Auto Scaling peut ainsi lancer un autre type d'instance si la capacité d'instance est insuffisante dans les zones de disponibilité que vous avez choisies. Pour plus d'informations, consultez Groupes Auto Scaling combinant plusieurs types d'instances et options d'achat.

  > **Warning**
  >
  > N'utilisez pas d'instances Spot avec des applications qui ne peuvent pas gérer une interruption d'instances Spot. Des interruptions peuvent se produire si le service Amazon EC2 Spot doit récupérer de la capacité.

Si vous cochez cette case, assurez-vous que le modèle de lancement ne demande pas déjà des instances Spot. Vous ne pouvez pas utiliser un modèle de lancement qui demande aux instances Spot de créer un groupe Auto Scaling qui utilise plusieurs types d'instances et lance des instances Spot et à la demande.

> **Note**
>
> Pour configurer ces options sur un groupe Auto Scaling qui utilise actuellement une configuration de lancement, vous devez sélectionner un modèle de lancement dans Update launch template (Mettre à jour le modèle de lancement). Le remplacement du type d'instance dans votre configuration de lancement n'est pas pris en charge.

8. (Facultatif) Dans Paramètres de restauration, choisissez Activer la restauration automatique pour restaurer automatiquement l'actualisation d'instance en cas d'échec.

   Ce paramètre ne peut être activé que lorsque le groupe Auto Scaling remplit les conditions préalables à l'utilisation des restaurations.

   Pour plus d'informations, consultez Annuler les modifications avec une restauration.

9. Passez en revue toutes vos sélections pour vous assurer que tout est correctement configuré.

   À ce stade, il est conseillé de vérifier que les différences entre les modifications actuelles et proposées n'affecteront pas votre application de manière inattendue ou indésirable. Pour vérifier que votre type d'instance est compatible avec votre modèle de lancement, consultez Compatibilité des types d'instance.

10. Lorsque vos sélections d'actualisation d'instance vous conviennent, sélectionnez Démarrer l'actualisation de l'instance.

Lancer une actualisation d'instance dans la console (groupe d'instances mixtes)

Utilisez la procédure suivante si vous avez créé un groupe Auto Scaling avec une politique d'instances mixtes. Si vous n'avez pas encore défini de politique d'instances mixtes pour votre groupe, consultez Lancer une actualisation d'instance dans la console (procédure de base) pour démarrer une actualisation d'instance.

Pour lancer une actualisation d'instance

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre dans la partie inférieure de la page Groupes Auto Scaling.

3. Dans l’onglet Actualisation d’instance, dans Actualisation d’instance active, sélectionnez Démarrer l’actualisation d’instance.
4. Pour les paramètres de disponibilité, procédez comme suit :

   a. Pour la méthode de remplacement des instances :

      - Si vous n’avez pas défini de politique de maintenance des instances pour le groupe Auto Scaling, le paramètre par défaut pour la méthode de remplacement d’instance est Résilier et lancer. Il s’agit de l’ancien comportement par défaut d’une actualisation d’instance.
      - Si vous définissez une politique de maintenance des instances sur le groupe Auto Scaling, elle fournit des valeurs par défaut pour la méthode de remplacement d’instance. Pour annuler la politique de maintenance des instances, choisissez Remplacer. Le remplacement s’applique uniquement à l’actualisation d’instance en cours. La prochaine fois que vous lancerez une actualisation d’instance, ces valeurs seront rétablies selon les valeurs par défaut de la politique de maintenance des instances.

      La procédure suivante explique comment mettre à jour la méthode de remplacement d’instance.

      i. Choisissez l’une des méthodes de remplacement d’instance suivantes :

         - Lancer avant toute résiliation : une nouvelle instance doit d’abord être mise en service avant qu’une instance existante puisse être résiliée. Il s’agit d’un bon choix pour les applications qui privilégient la disponibilité plutôt que les économies de coûts.
         - Résilier et lancer : les nouvelles instances sont mises en service en même temps que les instances existantes sont résiliées. Il s’agit d’un bon choix pour les applications qui privilégient les économies de coûts plutôt que la disponibilité. C’est également un bon choix pour les applications qui ne doivent pas libérer plus de capacité que ce qui est actuellement disponible.
         - Comportement personnalisé : cette option vous permet de définir une plage minimale et maximale personnalisée pour la quantité de capacité que vous souhaitez avoir à disposition lors du remplacement d’instances. Cela peut vous aider à trouver le juste équilibre entre le coût et la disponibilité.

      ii. Pour Définir un pourcentage d’intégrité, saisissez des valeurs pour l’un ou les deux champs suivants. Les champs d’activation varient en fonction de l’option que vous choisissez pour la méthode de remplacement d’instance.

- Min : définit le pourcentage minimal d’intégrité requis pour procéder à l’actualisation des instances.
- Max : définit le pourcentage maximal d’intégrité possible lors de l’actualisation des instances.

iii. Développez la section Afficher la capacité temporaire estimée pendant les remplacements en fonction de la taille actuelle du groupe pour confirmer comment les valeurs de Min et Max s’appliquent à votre groupe. Les valeurs exactes utilisées dépendent de la valeur de capacité souhaitée, qui changera si le groupe est mis à l’échelle.

iv. Développez la section Définir un comportement de repli pour les tailles de remplacement non valides, puis choisissez de déroger au pourcentage maximal valide afin de prioriser la disponibilité ou de déroger au pourcentage minimal valide.

Il n’est pas recommandé de conserver l’option par défaut Déroger au pourcentage minimal valide pour les très petits groupes. Lorsque le groupe Auto Scaling ne contient qu’une seule instance, le lancement d’une actualisation d’instance peut provoquer une panne.

Cette étape configure le comportement de secours si vous utilisez un groupe Auto Scaling qui n’a pas encore de politique de maintenance des instances. Cette option n’est pas disponible et n’apparaît pas lorsque votre groupe dispose d’une politique de maintenance des instances. Cette option n’est également disponible que pour la méthode de remplacement Résilier et lancer. Les autres méthodes de remplacement dérogeront au pourcentage maximal valide afin de donner la priorité à la disponibilité.

b. Pour Préparation d’instance, saisissez le nombre de secondes à partir du moment où l’état d’une nouvelle instance change en `InService` au moment où elle finit l’initialisation. Amazon EC2 Auto Scaling attend ce laps de temps avant de passer au remplacement de l’instance suivante.

Lors de la préparation, une instance nouvellement lancée n'est pas non plus prise en compte dans les métriques d'instance agrégées du groupe Auto Scaling (telles que `CPUUtilization`, `NetworkIn` et `NetworkOut`). Si vous avez ajouté des politiques de mise à l’échelle au groupe Auto Scaling, les activités de mise à l’échelle s’exécutent en parallèle. Si vous définissez un intervalle long pour la période de préchauffage de l'actualisation des instances, les instances nouvellement lancées mettent plus de temps

à apparaître dans les métriques. Par conséquent, une période de préchauffage adéquate empêche Amazon EC2 Auto Scaling de s'adapter à des données métriques périmées.

Si vous avez déjà défini correctement une préparation d'instance par défaut pour le groupe Auto Scaling, vous n'avez pas besoin de modifier la préparation d'instance. Toutefois, si vous souhaitez remplacer la valeur par défaut, vous pouvez définir une valeur pour cette option. Pour plus d'informations sur la définition de la préparation d'instance par défaut, consultez Définir la préparation par défaut d'instance d'un groupe Auto Scaling.

5. Sous Actualiser les paramètres, procédez comme suit :

   a. (Facultatif) Pour Points de contrôle, sélectionnez Activer les points de contrôle pour remplacer les instances à l'aide d'une approche progressive ou par phase d'une actualisation d'instance. Cela donne plus de temps pour la vérification entre les séries de remplacements. Si vous choisissez de ne pas activer les points de contrôle, les instances sont remplacées en une seule opération quasi continue.

      Si vous activez les points de contrôle, consultez Activer les points de contrôle (console) pour connaître les étapes supplémentaires.

   b. Activez ou désactivez Ignorer la correspondance :

      - Pour ignorer le remplacement des instances qui correspondent déjà à votre modèle de lancement et n'importe quel remplacement de type d'instance, laissez la case Activer la fonction Ignorer la correspondance cochée.
      - Si vous choisissez de désactiver la fonction Ignorer la correspondance en décochant cette case, toutes les instances peuvent être remplacées.

      Lorsque vous activez la fonction Ignorer la correspondance, vous pouvez définir un nouveau modèle de lancement ou une nouvelle version du modèle de lancement. Vous pouvez le faire dans la section Configuration souhaitée de la page Démarrer l'actualisation d'instance. Vous pouvez également mettre à jour vos remplacements de type d'instance dans Desired configuration (Configuration souhaitée).

   c. Pour les instances en veille, choisissez Ignorer, Résilier ou Attendre. Cela détermine ce qui se passe si des instances sont à l'état `Standby`. Pour plus d'informations, consultez Supprimer temporairement des instances du groupe Auto Scaling.

      Si vous choisissez Attendre, vous devez prendre des mesures supplémentaires pour remettre ces instances en service. Si vous ne le faites pas, l'actualisation d'instance

remplace toutes les instances `InService` et attend une heure. Ensuite, s'il reste des instances `Standby`, l'actualisation d'instance échoue. Pour éviter cette situation, choisissez à la place Ignorer ou Résilier ces instances.

d. Pour les instances protégées contre la mise à l'échelle horizontale, choisissez Ignorer, Remplacer ou Attendre. Cela détermine ce qui se passe si des instances protégées contre la mise à l'échelle horizontale sont trouvées. Pour plus d'informations, consultez Utiliser la protection de la taille d'instance.

   Si vous choisissez Attendre, vous devez prendre des mesures supplémentaires pour supprimer la protection contre la mise à l'échelle horizontale de ces instances. Si vous ne le faites pas, l'actualisation d'instance remplace toutes les instances non protégées et attend pendant une heure. Ensuite, s'il reste des instances protégées contre la mise à l'échelle horizontale, l'actualisation d'instance échoue. Pour éviter cette situation, choisissez à la place Ignorer ou Remplacer ces instances.

6. (Facultatif) Pour les CloudWatch CloudWatch alarmes, choisissez Activer les alarmes, puis choisissez une ou plusieurs alarmes. CloudWatch les alarmes peuvent être utilisées pour identifier tout problème et faire échouer le fonctionnement si une alarme passe à l'`ALARM`état. Pour plus d'informations, consultez Lancer une actualisation d'instance avec restauration automatique.

7. Dans la section Desired configuration (Configuration de cluster), effectuez les opérations suivantes.

   Pour cette étape, vous pouvez choisir d'utiliser la syntaxe JSON ou YAML pour modifier les valeurs des paramètres au lieu de faire des sélections dans l'interface de la console. Pour ce faire, sélectionnez Utiliser l'éditeur de code au lieu de Utiliser l'interface de la console. La procédure suivante explique comment effectuer des sélections à l'aide de l'interface de la console.

   a. Pour Mettre à jour le modèle de lancement :

      - Si vous n'avez pas créé de nouveau modèle de lancement ou de nouvelle version de modèle de lancement pour votre groupe Auto Scaling, ne cochez pas cette case.
      - Si vous avez créé un nouveau modèle de lancement ou une nouvelle version du modèle de lancement, cochez cette case. Lorsque vous sélectionnez cette option, Amazon EC2 Auto Scaling affiche le modèle de lancement actuel et la version actuelle du modèle de lancement. Il répertorie également toutes les autres versions disponibles. Choisissez le modèle de lancement, puis la version.

Après avoir choisi une version, vous pouvez voir les informations de version. Il s'agit de la version du modèle de lancement qui sera utilisée lors du remplacement d'instances dans le cadre d'une actualisation d'instance. Si l'actualisation de l'instance réussit, cette version du modèle de lancement sera également utilisée chaque fois que de nouvelles instances seront lancées, par exemple, lorsque le groupe évolue.

b. Pour Use these settings to override the instance type and purchase option defined in the launch template (Utiliser ces paramètres pour remplacer le type d'instance et l'option d'achat définis dans le modèle de lancement) :

Par défaut, cette case est cochée. Amazon EC2 Auto Scaling remplit chaque paramètre avec la valeur actuellement définie dans la politique d'instances mixtes pour le groupe Auto Scaling. Ne mettez à jour que les valeurs des paramètres que vous souhaitez modifier. Pour obtenir des conseils sur ces paramètres, consultez Groupes Auto Scaling combinant plusieurs types d'instances et options d'achat.

> ⚠ Warning
>
> Nous vous recommandons de ne pas décocher cette case. Décochez-la uniquement si vous souhaitez arrêter d'utiliser une politique d'instances mixtes. Une fois l'actualisation de l'instance réussie, Amazon EC2 Auto Scaling met à jour votre groupe pour qu'il corresponde à la Desired configuration (Configuration souhaitée). S'il n'inclut plus de politique d'instances mixtes, Amazon EC2 Auto Scaling résilie progressivement toutes les instances Spot en cours d'exécution et les remplace par des instances à la demande. Ou, si votre modèle de lancement demande des instances Spot, alors Amazon EC2 Auto Scaling résilie progressivement toutes les instances à la demande en cours d'exécution et les remplace par des instances Spot.

8. (Facultatif) Pour Paramètres de restauration, choisissez Activer la restauration automatique pour restaurer automatiquement l'actualisation d'instance si elle échoue pour quelque raison que ce soit.

   Ce paramètre ne peut être activé que lorsque le groupe Auto Scaling remplit les conditions préalables à l'utilisation des restaurations.

   Pour plus d'informations, consultez Annuler les modifications avec une restauration.

9. Passez en revue toutes vos sélections pour vous assurer que tout est correctement configuré.

À ce stade, il est conseillé de vérifier que les différences entre les modifications actuelles et proposées n'affecteront pas votre application de manière inattendue ou indésirable. Pour vérifier que votre type d'instance est compatible avec votre modèle de lancement, consultez Compatibilité des types d'instance.

Lorsque vos sélections d'actualisation d'instance vous conviennent, sélectionnez Démarrer l'actualisation de l'instance.

Lancer une actualisation d'instance (AWS CLI)

Pour lancer une actualisation d'instance

Utilisez la commande start-instance-refresh pour lancer une actualisation d'instance à partir de l'interface AWS CLI. Vous pouvez spécifier les préférences que vous souhaitez modifier dans un fichier de configuration JSON. Lorsque vous référencez le fichier de configuration, indiquez le chemin d'accès et le nom du fichier comme indiqué dans l'exemple suivant.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Contenu de `config.json` :

```
{
    "AutoScalingGroupName": "my-asg",
    "Preferences": {
      "InstanceWarmup": 60,
      "MinHealthyPercentage": 50,
      "AutoRollback": true,
      "ScaleInProtectedInstances": Ignore,
      "StandbyInstances": Terminate
    }
}
```

Si les préférences ne sont pas fournies, les valeurs par défaut sont utilisées. Pour plus d'informations, consultez Comprendre les valeurs par défaut d'une actualisation d'instance.

Exemple de sortie :

```
{
    "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"
```

```
}
```

Surveiller l'actualisation d'une instance

Vous pouvez surveiller une actualisation d'instance en cours ou consulter l'état des actualisations d'instance passées au cours des six dernières semaines à l'aide du AWS Management Console or. AWS CLI

Surveiller et vérifier l'état de l'actualisation d'une instance

Pour surveiller et vérifier l'état de l'actualisation d'une instance, appliquez l'une des méthodes suivantes :

Console

> **Tip**
>
> Dans cette procédure, les colonnes nommées doivent déjà être affichées. Pour afficher les colonnes masquées ou modifier le nombre de lignes affichées, cliquez sur l'icône représentant un engrenage dans le coin supérieur droit de la section pour ouvrir le mode des préférences. Mettez à jour les paramètres selon vos besoins et choisissez Confirmer.

Pour surveiller et vérifier l'état de l'actualisation d'une instance (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Dans l'onglet Instance refresh (Actualisation de l'instance), sous Instance refresh history (Historique d'actualisation de l'instance), vous pouvez déterminer l'état de votre demande en consultant la colonne Statut (Status). L'opération passe en `Pending` statut lors de son initialisation. Cet état devrait ensuite rapidement passer à `InProgress`. Lorsque toutes les instances sont mises à jour, l'état devient `Successful`.
4. Vous pouvez également suivre le succès ou l'échec des activités en cours en consultant les activités de dimensionnement du groupe. Dans l'onglet Activity (Activité), sous Activity history (Historique des activités), lorsque l'actualisation de l'instance démarre, vous voyez des entrées lorsque les instances sont mises hors service, et un autre ensemble

d'entrées lorsque les instances sont lancées. Si vous avez de nombreuses activités de dimensionnement, vous pouvez en voir davantage en cliquant sur l'icône **>** en haut de l'historique des activités. Pour plus d'informations sur la résolution des problèmes susceptibles d'entraîner l'échec des activités, consultez[Résoudre les problèmes d'Amazon EC2 Auto Scaling](#).

5. (Facultatif) Dans l'onglet Gestion des instances, sous Instances, vous pouvez suivre la progression d'instances spécifiques selon vos besoins.

AWS CLI

Pour surveiller et vérifier l'état de l'actualisation d'une instance (AWS CLI)

Utilisez la commande describe-instance-refreshes suivante.

```
aws autoscaling describe-instance-refreshes --auto-scaling-group-name my-asg
```

Voici un exemple de sortie.

Les actualisations des instances sont ordonnées par heure de début. Les actualisations d'instances toujours en cours sont décrites en premier.

```
{
  "InstanceRefreshes":[
    {
      "InstanceRefreshId":"08b91cf7-8fa6-48af-b6a6-d227f40f1b9b",
      "AutoScalingGroupName":"my-asg",
      "Status":"InProgress",
      "StatusReason":"Waiting for instances to warm up before continuing. For
 example: i-0645704820a8e83ff is warming up.",
      "StartTime":"2023-11-24T16:46:52+00:00",
      "PercentageComplete":50,
      "InstancesToUpdate":0,
      "Preferences":{
        "MaxHealthyPercentage":120,
        "MinHealthyPercentage":90,
        "InstanceWarmup":60,
        "SkipMatching":false,
        "AutoRollback":true,
        "ScaleInProtectedInstances":"Ignore",
        "StandbyInstances":"Ignore"
      }
```

```
    },
    {
      "InstanceRefreshId":"0e151305-1e57-4a32-a256-1fd14157c5ec",
      "AutoScalingGroupName":"my-asg",
      "Status":"Successful",
      "StartTime":"2023-11-22T13:53:37+00:00",
      "EndTime":"2023-11-22T13:59:45+00:00",
      "PercentageComplete":100,
      "InstancesToUpdate":0,
      "Preferences":{
        "MaxHealthyPercentage":120,
        "MinHealthyPercentage":90,
        "InstanceWarmup":60,
        "SkipMatching":false,
        "AutoRollback":true,
        "ScaleInProtectedInstances":"Ignore",
        "StandbyInstances":"Ignore"
      }
    }
  ]
}
```


Vous pouvez également suivre le succès ou l'échec des activités en cours en consultant les activités de dimensionnement du groupe. Les activités de dimensionnement vous permettent également d'obtenir plus de détails afin de résoudre les problèmes liés à l'actualisation d'une instance. Pour plus d’informations, consultez Résoudre les problèmes d'Amazon EC2 Auto Scaling.

États d'actualisation d'instance

Lorsque vous lancez une actualisation d'instance, celle-ci passe à l'état En attente. Il passe de En attente à InProgressjusqu'à ce qu'il atteigne Successful, Echec RollbackSuccessful, Annulé ou RollbackFailed.

Une actualisation d'instance peut avoir les états suivants :

État	Description
En suspens	La demande a été créée, mais l'actualisation d'instance n'a pas démarré.
InProgress	Une actualisation d'instance est en cours.

État	Description
Réussite	Une actualisation d'instance s'est terminée avec succès.
Échec	L'actualisation d'instance a échoué. Vous pouvez résoudre les problèmes en consultant le motif de cet état et les activités de mise à l'échelle.
Annulation	Une actualisation d'instance en cours est en cours d'annulation.
Annulée	L'actualisation d'instance est annulée.
RollbackInProgrès	Une actualisation d'instance est en cours de restauration.
RollbackFailed	La restauration a échoué. Vous pouvez résoudre les problèmes en consultant le motif de cet état et les activités de mise à l'échelle.
RollbackSuccessful	La restauration s'est terminée avec succès.

Annuler une actualisation d'instance

Vous pouvez annuler une actualisation d'instance qui est toujours en cours. Vous ne pouvez pas l'annuler une fois qu'elle est terminée.

L'annulation d'une actualisation d'instance ne restaure pas les instances qui ont déjà été remplacées. Pour restaurer les modifications apportées à vos instances, effectuez plutôt une restauration. Pour plus d'informations, consultez Annuler les modifications avec une restauration.

Rubriques

- Annuler une actualisation d'instance (console)
- Annuler une actualisation d'instance (AWS CLI)

Annuler une actualisation d'instance (console)

Pour annuler une actualisation d'instance

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.

2. Cochez la case située en regard du groupe Auto Scaling.
3. Sous l'onglet Actualisation de l'instance, dans Actualisation d'instance active, sélectionnez Annuler l'actualisation d'instance, Annuler.
4. Lorsque vous êtes invité à confirmer l'opération, choisissez Confirmer.

L'état de l'actualisation d'instance est défini sur Annulation. Une fois l'annulation terminée, l'état de l'actualisation d'instance est défini sur Annulé.

Annuler une actualisation d'instance (AWS CLI)

Pour annuler une actualisation d’instance

Utilisez la commande cancel-instance-refresh du AWS CLI et indiquez le nom du groupe Auto Scaling.

```
aws autoscaling cancel-instance-refresh --auto-scaling-group-name my-asg
```

Exemple de sortie :

```
{
    "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"
}
```

Annuler les modifications avec une restauration

Vous pouvez restaurer une actualisation d'instance qui est toujours en cours. Vous ne pouvez pas la restaurer une fois qu'elle est terminée. Toutefois, vous pouvez à nouveau mettre à jour votre groupe Auto Scaling en lançant une nouvelle actualisation d'instance.

Lors de la restauration, Amazon EC2 Auto Scaling remplace les instances déployées jusqu'à présent. Les nouvelles instances correspondent à la configuration que vous avez enregistrée pour la dernière fois dans le groupe Auto Scaling avant de commencer l'actualisation d'instance.

Amazon EC2 Auto Scaling fournit les méthodes de restauration suivantes :

- Restauration manuelle : vous lancez une restauration manuellement pour inverser ce qui a été déployé jusqu'au point de restauration.

- Annulation automatique : Amazon EC2 Auto Scaling annule automatiquement ce qui a été déployé si l'actualisation de l'instance échoue pour une raison ou une CloudWatch autre ou si l'une des alarmes que vous spécifiez passe à l'état normal. `ALARM`

Table des matières

Considérations

Les considérations suivantes s'appliquent lorsque vous utilisez une restauration :

- L'option de restauration n'est disponible que si vous spécifiez la configuration souhaitée lors du démarrage de l'actualisation d'une instance.
- Vous ne pouvez revenir à une version précédente d'un modèle de lancement que s'il s'agit d'une version numérotée spécifique. L'option de restauration n'est pas disponible si le groupe Auto Scaling est configuré pour utiliser la version du modèle de lancement `$Latest` ou `$Default`.
- Vous ne pouvez pas non plus revenir à un modèle de lancement configuré pour utiliser un alias d'AMI depuis le AWS Systems Manager Parameter Store.
- La configuration que vous avez enregistrée pour la dernière fois dans le groupe Auto Scaling doit être stable. Si elle n'est pas stable, le flux de travail de restauration s'exécutera toujours, mais il finira par échouer. Tant que vous n'aurez pas résolu le problème, le groupe Auto Scaling risque de se trouver à un état d'échec qui ne lui permet plus de lancer les instances avec succès. Cela peut avoir un impact sur la disponibilité du service ou de l'application.

Lancer manuellement une restauration

Console

Pour lancer manuellement une restauration d'une actualisation d'instance (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

3. Dans l'onglet Actualisation d'instance, dans Actualisation d'instance active, sélectionnez Actions, Lancer une restauration.
4. Lorsque vous êtes invité à confirmer l'opération, choisissez Confirmer.

AWS CLI

Pour lancer manuellement une restauration d'une actualisation d'instance (AWS CLI)

Utilisez la commande rollback-instance-refresh à partir de l' AWS CLI et indiquez le nom du groupe Auto Scaling.

```
aws autoscaling rollback-instance-refresh --auto-scaling-group-name my-asg
```

Exemple de sortie :

```
{
    "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"
}
```

> **Tip**
>
> Si cette commande génère une erreur, assurez-vous d'avoir mis à jour AWS CLI localement la dernière version.

Lancer une actualisation d'instance avec restauration automatique

À l'aide de la fonction d'annulation automatique, vous pouvez annuler automatiquement l'actualisation de l'instance en cas d'échec, par exemple en cas d'erreur ou en cas de passage à l'`ALARM`état CloudWatch d'une alarme Amazon spécifiée.

Si vous activez la restauration automatique et que des erreurs se produisent lors du remplacement des instances, l'actualisation des instances tente de terminer tous les remplacements pendant une heure avant d'échouer et de restaurer les instances. Ces erreurs sont généralement causées par des facteurs tels que l'échec du lancement d'EC2, des surveillances de l'état mal configurées, le fait de ne pas ignorer ou d'autoriser la résiliation d'instances dans l'état `Standby` ou protégées contre la mise à l'échelle horizontale.

La spécification CloudWatch des alarmes est facultative. Pour définir une alarme, vous devez d'abord la créer. Vous pouvez créer des alarmes de métrique et des alarmes composites. Pour plus d'informations sur la création de l'alarme, consultez le guide de CloudWatch l'utilisateur Amazon. En utilisant les métriques de Elastic Load Balancing par exemple, si vous utilisez un Application Load Balancer, vous pouvez utiliser les métriques `HTTPCode_ELB_5XX_Count` et `HTTPCode_ELB_4XX_Count`.

Considérations

- Si vous spécifiez une CloudWatch alarme mais que vous n'activez pas la restauration automatique et que l'état de l'alarme passe à`ALARM`, l'actualisation de l'instance échoue sans restauration.
- Vous pouvez choisir un maximum de 10 alarmes lorsque vous lancez une actualisation d'instance.
- Lorsque vous choisissez une CloudWatch alarme, celle-ci doit être dans un état compatible. Si l'état d'alarme est `INSUFFICIENT_DATA` ou `ALARM`, vous recevez un message d'erreur lorsque vous essayez de démarrer l'actualisation de l'instance.
- Lorsque vous créez une alarme à utiliser par Amazon EC2 Auto Scaling, l'alarme doit indiquer comment traiter les points de données manquants. Si une métrique manque fréquemment des points de données par conception, l'état de l'alarme est `INSUFFICIENT_DATA` pendant ces périodes. Dans ce cas, Amazon EC2 Auto Scaling ne peut pas remplacer les instances tant que de nouveaux points de données ne sont pas trouvés. Pour forcer l'alarme à maintenir l'état précédent `ALARM` ou `OK`, vous pouvez choisir d'ignorer les données manquantes. Pour plus d'informations, consultez la section Configuration de la manière dont les alarmes traitent les données manquantes dans le guide de CloudWatch l'utilisateur Amazon.

Console

Pour lancer une actualisation d'instance avec restauration automatique (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.
3. Dans l'onglet Actualisation d'instance, dans Actualisation d'instance active, sélectionnez Démarrer l'actualisation d'instance.
4. Suivez la procédure Lancer une actualisation d'instance (console) et configurez vos paramètres d'actualisation d'instance selon vos besoins.

5. (Facultatif) Sous Actualiser les paramètres, pour les CloudWatch CloudWatch alarmes, choisissez Activer les alarmes, puis choisissez une ou plusieurs alarmes pour identifier les problèmes éventuels et faire échouer l'opération si une alarme passe à l'`ALARM`état indiqué.
6. Dans Paramètres de restauration, choisissez Activer la restauration automatique pour restaurer automatiquement une actualisation d'instance ayant échoué à la configuration que vous avez enregistrée pour la dernière fois dans le groupe Auto Scaling avant de commencer l'actualisation d'instance.
7. Passez en revue vos sélections, puis choisissez Démarrer l'actualisation de l'instance.

AWS CLI

Pour lancer une actualisation d'instance avec restauration automatique (AWS CLI)

Utilisez la commande start-instance-refresh et spécifiez `true` pour l'option `AutoRollback`en `Preferences`.

L'exemple suivant montre comment démarrer une actualisation d'instance qui sera automatiquement restaurée en cas d'échec. Remplacez les valeurs de paramètre *`italicized`* par vos propres valeurs.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Contenu de `config.json`.

```
{
    "AutoScalingGroupName": "my-asg",
    "DesiredConfiguration": {
      "LaunchTemplate": {
          "LaunchTemplateName": "my-launch-template",
          "Version": "1"
       }
    },
    "Preferences": {
      "AutoRollback": true
    }
}
```

Sinon, pour revenir automatiquement en arrière lorsque l'actualisation de l'instance échoue ou lorsqu'une CloudWatch alarme spécifiée est active`ALARM`, spécifiez

l'`AlarmSpecification`option dans le `Preferences` et fournissez le nom de l'alarme, comme dans l'exemple suivant. Remplacez les valeurs de paramètre *`italicized`* par vos propres valeurs.

```
{
    "AutoScalingGroupName": "my-asg",
    "DesiredConfiguration": {
      "LaunchTemplate": {
          "LaunchTemplateName": "my-launch-template",
          "Version": "1"
       }
    },
    "Preferences": {
      "AutoRollback": true,
      "AlarmSpecification": { "Alarms": [ "my-alarm" ] }
    }
}
```

Si elle aboutit, la commande renvoie un résultat semblable au suivant :

```
{
  "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"
}
```

> **Tip**
>
> Si cette commande génère une erreur, assurez-vous d'avoir mis à jour AWS CLI localement la dernière version.

Utiliser une actualisation d'instance avec la fonction Ignorer la correspondance

La fonction Ignorer la correspondance indique à Amazon EC2 Auto Scaling d'ignorer les instances qui disposent déjà de vos dernières mises à jour. De cette façon, vous ne remplacez pas plus d'instances que nécessaire. Cela est utile quand vous souhaitez vous assurer que votre groupe Auto Scaling utilise une version particulière de votre modèle de lancement et ne remplace que les instances qui utilisent une version différente.

Les considérations suivantes s'appliquent à la fonction Ignorer la correspondance :

- Si vous lancez une actualisation d'instance à la fois avec la fonction Ignorer la correspondance et la configuration souhaitée, Amazon EC2 Auto Scaling vérifie si des instances correspondent à la configuration souhaitée. Ensuite, il remplace uniquement les instances qui ne correspondent pas à la configuration souhaitée. Une fois l'actualisation d'instance réussie, Amazon EC2 Auto Scaling met à jour le groupe pour qu'il corresponde à la configuration souhaitée.
- Si vous lancez une actualisation d'instance avec la fonction Ignorer la correspondance, mais que vous ne spécifiez pas la configuration souhaitée, Amazon EC2 Auto Scaling vérifie si des instances correspondent à la configuration que vous avez enregistrée pour la dernière fois dans le groupe Auto Scaling. Ensuite, il remplace uniquement les instances qui ne correspondent pas à la configuration que vous avez enregistrée la dernière fois.
- Vous pouvez utiliser la fonction Ignorer la correspondance avec un nouveau modèle de lancement, une nouvelle version du modèle de lancement ou un ensemble de types d'instance. Si vous activez la fonction Ignorer la correspondance, mais qu'aucune d'entre elles ne change, le rafraîchissement des instances réussira immédiatement sans remplacer aucune instance. Si vous avez apporté d'autres modifications à la configuration souhaitée (à votre politique d'allocation des instances Spot, par exemple), Amazon EC2 Auto Scaling attend la réussite de l'actualisation d'instance. Il met ensuite à jour les paramètres du groupe Auto Scaling pour refléter la nouvelle configuration souhaitée.
- Vous ne pouvez pas utiliser la fonction Ignorer la correspondance avec une nouvelle configuration de lancement.
- Lorsque vous lancez une actualisation d'instance et que vous fournissez la configuration souhaitée, Amazon EC2 Auto Scaling garantit que toutes les instances utilisent la configuration souhaitée. Par conséquent, lorsque vous spécifiez l'une `$Default` ou `$Latest` l'autre version souhaitée pour votre modèle de lancement, puis que vous créez une nouvelle version du modèle de lancement alors qu'une actualisation d'instance est en cours, toutes les instances déjà remplacées seront remplacées à nouveau.
- Skip matching ne permet pas de savoir si un script de données utilisateur du modèle de lancement extraira le code mis à jour et l'installera sur les nouvelles instances. Par conséquent, il se peut que le remplacement des instances sur lesquelles un code obsolète soit installé ne soit pas remplacé. Dans ce cas, vous devez désactiver la mise en correspondance des erreurs pour vous assurer que toutes les instances reçoivent votre dernier code, même si la version du modèle de lancement n'est pas mise à jour.

Cette section contient des AWS CLI instructions pour démarrer une actualisation d'instance en activant la mise en correspondance des sauts. Pour des instructions sur l'utilisation de la console, consultez Lancer une actualisation d'instance (console).

Ignorer la correspondance (procédure de base)

Suivez les étapes décrites dans cette section AWS CLI pour effectuer les opérations suivantes :

- Créez le modèle de lancement que vous souhaitez appliquer à vos instances.
- Lancez une actualisation d'instance pour appliquer votre modèle de lancement à votre groupe Auto Scaling. Si vous n'activez pas la fonction Ignorer la correspondance, toutes les instances seront remplacées. Cela est vrai même si le modèle de lancement utilisé pour provisionner l'instance est le même que celui que vous avez spécifié pour la configuration souhaitée.

Pour utiliser la fonction Ignorer la correspondance avec un nouveau modèle de lancement

1. Utilisez la commande create-launch-template pour créer un nouveau modèle de lancement pour votre groupe Auto Scaling. Incluez l'option `--launch-template-data` et l'entrée JSON qui définit les détails des instances créées pour votre groupe Auto Scaling.

   Par exemple, utilisez la commande suivante pour créer un modèle de lancement de base avec l'ID d'AMI *ami-0123456789abcdef0* et le type d'instance *t2.micro*.

   ```
   aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling
    --version-description version1 \
     --launch-template-data
    '{"ImageId":"ami-0123456789abcdef0","InstanceType":"t2.micro"}'
   ```

   Si elle aboutit, la commande renvoie un résultat semblable au suivant :

   ```
   {
      "LaunchTemplate": {
        "LaunchTemplateId": "lt-068f72b729example",
        "LaunchTemplateName": "my-template-for-auto-scaling",
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
        "CreateTime": "2023-01-30T18:16:06.000Z",
        "DefaultVersionNumber": 1,
        "LatestVersionNumber": 1
     }
   }
   ```

Pour plus d'informations, consultez Exemples de création et de gestion de modèles de lancement à l'aide du AWS CLI.

2. Utilisez la commande start-instance-refresh pour lancer le flux de travail de remplacement d'instance et appliquer votre nouveau modèle de lancement avec l'ID *lt-068f72b729example*. Le modèle de lancement étant nouveau, il ne comporte qu'une seule version. Cela signifie que la version 1 du modèle de lancement est la cible de cette actualisation d'instance. Si un événement de montée en puissance se produit lors de l'actualisation d'instance et si Amazon EC2 Auto Scaling provisionne de nouvelles instances à l'aide de la version 1 de ce modèle de lancement, elles ne seront pas remplacées. Une fois l'opération terminée avec succès, le nouveau modèle de lancement est appliqué à votre groupe Auto Scaling.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Contenu de `config.json`.

```
{
    "AutoScalingGroupName": "my-asg",
    "DesiredConfiguration": {
      "LaunchTemplate": {
          "LaunchTemplateId": "lt-068f72b729example",
          "Version": "$Default"
       }
    },
    "Preferences": {
      "SkipMatching": true
    }
}
```

Si elle aboutit, la commande renvoie un résultat semblable au suivant :

```
{
  "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"
}
```

Fonction Ignorer la correspondance (groupe d'instances mixtes)

Si vous avez un groupe Auto Scaling doté d'une politique d'instances mixtes, suivez les étapes décrites dans cette section AWS CLI pour démarrer une actualisation d'instance en sautant la correspondance. Vous avez les options suivantes :

- Fournissez un nouveau modèle de lancement à appliquer à tous les types d'instances spécifiés dans la politique.
- Fournissez un ensemble actualisé de types d'instances avec ou sans modification du modèle de lancement dans la politique. Par exemple, il se peut que vous souhaitiez migrer les types d'instance indésirables. Vous utiliseriez le modèle de lancement tel quel, sans modifier l'AMI, les groupes de sécurité ou les autres spécificités des instances remplacées.

Suivez les étapes décrites dans l'une des sections suivantes, selon l'option qui répond à vos besoins.

Pour utiliser la fonction Ignorer la correspondance avec un nouveau modèle de lancement

1. Utilisez la commande create-launch-template pour créer un nouveau modèle de lancement pour votre groupe Auto Scaling. Incluez l'option `--launch-template-data` et l'entrée JSON qui définit les détails des instances créées pour votre groupe Auto Scaling.

   Par exemple, utilisez la commande suivante pour créer un modèle de lancement avec l'ID d'AMI *ami-0123456789abcdef0*.

   ```
   aws ec2 create-launch-template --launch-template-name my-new-template --version-
   description version1 \
     --launch-template-data '{"ImageId":"ami-0123456789abcdef0"}'
   ```

   Si elle aboutit, la commande renvoie un résultat semblable au suivant :

   ```
   {
      "LaunchTemplate": {
        "LaunchTemplateId": "lt-04d5cc9b88example",
        "LaunchTemplateName": "my-new-template",
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
        "CreateTime": "2023-01-31T15:56:02.000Z",
        "DefaultVersionNumber": 1,
        "LatestVersionNumber": 1
     }
   ```

```
}
```

Pour plus d'informations, consultez Exemples de création et de gestion de modèles de lancement à l'aide du AWS CLI.

2. Pour afficher la politique d'instances mixtes existante de votre groupe Auto Scaling, exécutez la commande describe-auto-scaling-groups. Vous aurez besoin de ces informations à l'étape suivante, lorsque vous lancerez l'actualisation d'instance.

   L'exemple de commande suivant renvoie la politique d'instances mixtes configurée pour le groupe Auto Scaling nommé *my-asg*.

   ```
   aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
   ```

   Si elle aboutit, la commande renvoie un résultat semblable au suivant :

   ```
   {
     "AutoScalingGroups":[
       {
         "AutoScalingGroupName":"my-asg",
         "AutoScalingGroupARN":"arn",
         "MixedInstancesPolicy":{
           "LaunchTemplate":{
             "LaunchTemplateSpecification":{
               "LaunchTemplateId":"lt-073693ed27example",
               "LaunchTemplateName":"my-old-template",
               "Version":"$Default"
             },
             "Overrides":[
               {
                 "InstanceType":"c5.large"
               },
               {
                 "InstanceType":"c5a.large"
               },
               {
                 "InstanceType":"m5.large"
               },
               {
                 "InstanceType":"m5a.large"
               }
             ]
   ```

```
        },
        "InstancesDistribution":{
          "OnDemandAllocationStrategy":"prioritized",
          "OnDemandBaseCapacity":1,
          "OnDemandPercentageAboveBaseCapacity":50,
          "SpotAllocationStrategy":"price-capacity-optimized"
        }
      },
      "MinSize":1,
      "MaxSize":5,
      "DesiredCapacity":4,
      ...
    }
  ]
}
```

3. Utilisez la commande start-instance-refresh pour lancer le flux de travail de remplacement d'instance et appliquer votre nouveau modèle de lancement avec l'ID *lt-04d5cc9b88example*. Le modèle de lancement étant nouveau, il ne comporte qu'une seule version. Cela signifie que la version 1 du modèle de lancement est la cible de cette actualisation d'instance. Si un événement de montée en puissance se produit lors de l'actualisation d'instance et si Amazon EC2 Auto Scaling provisionne de nouvelles instances à l'aide de la version 1 de ce modèle de lancement, elles ne seront pas remplacées. Une fois l'opération terminée avec succès, la politique d'instances mixtes mise à jour est appliquée à votre groupe Auto Scaling.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Contenu de `config.json`.

```
{
  "AutoScalingGroupName":"my-asg",
  "DesiredConfiguration":{
    "MixedInstancesPolicy":{
      "LaunchTemplate":{
        "LaunchTemplateSpecification":{
          "LaunchTemplateId":"lt-04d5cc9b88example",
          "Version":"$Default"
        },
        "Overrides":[
          {
            "InstanceType":"c5.large"
```

```
          },
          {
            "InstanceType":"c5a.large"
          },
          {
            "InstanceType":"m5.large"
          },
          {
            "InstanceType":"m5a.large"
          }
        ]
      },
      "InstancesDistribution":{
        "OnDemandAllocationStrategy":"prioritized",
        "OnDemandBaseCapacity":1,
        "OnDemandPercentageAboveBaseCapacity":50,
        "SpotAllocationStrategy":"price-capacity-optimized"
        }
      }
    }
  },
  "Preferences":{
    "SkipMatching":true
  }
}
```

Si elle aboutit, la commande renvoie un résultat semblable au suivant :

```
{
  "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"
}
```

Dans cette procédure suivante, vous allez fournir un ensemble actualisé de types d'instances sans modifier le modèle de lancement.

Pour utiliser la fonction Ignorer la correspondance avec un ensemble actualisé de types d'instances

1. Pour afficher la politique d'instances mixtes existante de votre groupe Auto Scaling, exécutez la commande describe-auto-scaling-groups. Vous aurez besoin de ces informations à l'étape suivante, lorsque vous lancerez l'actualisation d'instance.

L'exemple de commande suivant renvoie la politique d'instances mixtes configurée pour le groupe Auto Scaling nommé *my-asg*.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Si elle aboutit, la commande renvoie un résultat semblable au suivant :

```
{
  "AutoScalingGroups":[
    {
      "AutoScalingGroupName":"my-asg",
      "AutoScalingGroupARN":"arn",
      "MixedInstancesPolicy":{
        "LaunchTemplate":{
          "LaunchTemplateSpecification":{
            "LaunchTemplateId":"lt-073693ed27example",
            "LaunchTemplateName":"my-template-for-auto-scaling",
            "Version":"$Default"
          },
          "Overrides":[
            {
              "InstanceType":"c5.large"
            },
            {
              "InstanceType":"c5a.large"
            },
            {
              "InstanceType":"m5.large"
            },
            {
              "InstanceType":"m5a.large"
            }
          ]
        },
        "InstancesDistribution":{
          "OnDemandAllocationStrategy":"prioritized",
          "OnDemandBaseCapacity":1,
          "OnDemandPercentageAboveBaseCapacity":50,
          "SpotAllocationStrategy":"price-capacity-optimized"
        }
      },
      "MinSize":1,
```

```
      "MaxSize":5,
      "DesiredCapacity":4,
      ...
    }
  ]
}
```

2. Utilisez la commande start-instance-refresh pour lancer le flux de travail de remplacement d'instance et appliquer vos mises à jour. Si vous souhaitez remplacer les instances qui utilisent des types d'instance spécifiques, la configuration souhaitée doit spécifier la politique d'instances mixtes comprenant uniquement les types d'instance que vous souhaitez. Vous pouvez choisir d'ajouter de nouveaux types d'instance à leur place.

   L'exemple de commande suivant lance une actualisation d'instance sans le type d'instance indésirable *m5a.large*. Lorsqu'un type d'instance de votre groupe ne correspond pas à l'un des trois types d'instances restants, les instances sont remplacées. (Notez qu'une actualisation d'instance ne choisit pas les types d'instance à partir desquels les nouvelles instances doivent être approvisionnées ; ce choix revient aux stratégies d'allocation.) Une fois l'opération terminée avec succès, la politique d'instances mixtes mise à jour est appliquée à votre groupe Auto Scaling.

   ```
   aws autoscaling start-instance-refresh --cli-input-json file://config.json
   ```

   Contenu de `config.json`

   ```
   {
     "AutoScalingGroupName":"my-asg",
     "DesiredConfiguration":{
       "MixedInstancesPolicy":{
         "LlaunchTemplate":{
           "LaunchTemplateSpecification":{
             "LaunchTemplateId":"lt-073693ed27example",
             "Version":"$Default"
           },
           "Overrides":[
             {
               "InstanceType":"c5.large"
             },
             {
               "InstanceType":"c5a.large"
             },
   ```

```
            {
              "InstanceType":"m5.large"
            }
          ]
        },
        "InstancesDistribution":{
          "OnDemandAllocationStrategy":"prioritized",
          "OnDemandBaseCapacity":1,
          "OnDemandPercentageAboveBaseCapacity":50,
          "SpotAllocationStrategy":"price-capacity-optimized"
          }
        }
      }
    },
    "Preferences":{
      "SkipMatching":true
    }
}
```

Ajouter des points de contrôle à une actualisation d'instance

Lorsque vous utilisez une actualisation d'instance, vous pouvez choisir de remplacer les instances par phases, afin de pouvoir effectuer des vérifications sur vos instances au fur et à mesure. Pour effectuer un remplacement par phases, vous devez ajouter des points de contrôle qui permettront de mettre l'actualisation d'instance en pause. L'utilisation de points de contrôle vous permet de mieux contrôler la façon dont vous choisissez de mettre à jour votre groupe Auto Scaling. Cela vous aide à vous assurer que votre application fonctionnera de manière fiable et prévisible.

Table des matières

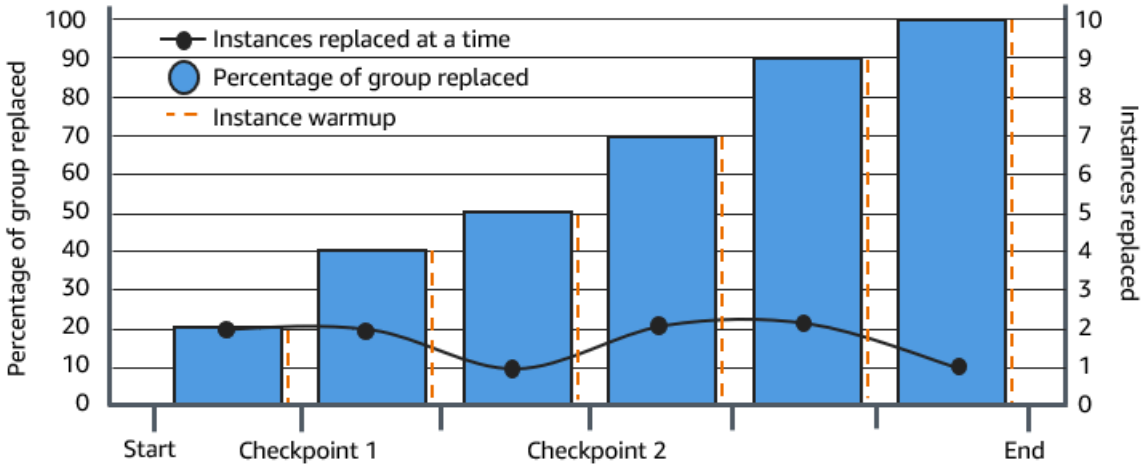
Comment ça marche

Lorsque vous lancez une actualisation d'instance, vous spécifiez les points de contrôle sous forme de pourcentages du nombre total d'instances du groupe Auto Scaling. Ces points de contrôle indiquent

le pourcentage minimum d'instances du groupe Auto Scaling qui doivent être de nouvelles instances avant que le point de contrôle soit considéré comme atteint. Par exemple, si vos points de contrôle le sont`[20, 50, 100]`, le premier point de contrôle est atteint lorsque 20 % des instances sont nouvelles, le second lorsque 50 % sont nouvelles et le dernier point de contrôle lorsque toutes les instances sont nouvelles.

Amazon EC2 Auto Scaling accélère le remplacement des instances afin de respecter les pourcentages de points de contrôle spécifiés tout en maintenant le pourcentage de santé minimum du groupe. Pour atteindre un pourcentage de point de contrôle, Amazon EC2 Auto Scaling remplace parfois un nombre de points de contrôle inférieur, mais jamais plus que ce que permet le pourcentage minimal valide.

Prenons l'exemple du groupe Auto Scaling suivant qui compte 10 instances. Les pourcentages de points de contrôle sont `[20,50,100]`, le pourcentage minimal valide est de 80 % et le pourcentage maximal valide est de 100 %. Pour maintenir le pourcentage minimal valide, seulement deux instances peuvent être remplacées à la fois. Le graphique suivant résume le processus de remplacement des instances avant qu'un point de contrôle ne soit atteint.



Dans l'exemple ci-dessus, il existe une période de préchauffage pour chaque nouvelle instance qui démarre. Vous pouvez également avoir un hook de cycle de vie qui met une instance en attente, puis exécute une action personnalisée lors de son lancement ou de sa résiliation.

Amazon EC2 Auto Scaling émet des événements pour chaque point de contrôle, à l'exception du point de contrôle complet à 100 %. Vous pouvez ajouter une EventBridge règle pour envoyer les événements à une cible telle qu'Amazon SNS. De cette façon, vous recevez une notification lorsque vous pouvez effectuer les vérifications requises. Pour plus d'informations, consultez Créez des EventBridge règles pour les événements d'actualisation, par exemple.

Considérations

Lorsque vous utilisez des points de contrôle, gardez à l'esprit les considérations suivantes :

- Étant donné que les points de contrôle sont basés sur des pourcentages, le nombre d'instances à remplacer change en fonction de la taille du groupe. Lorsqu'une activité de montée en puissance se produit et que la taille du groupe augmente, une opération en cours peut à nouveau atteindre un point de contrôle. Dans ce cas, Amazon EC2 Auto Scaling envoie une autre notification et applique à nouveau le délai d'attente spécifié entre les points de contrôle avant de continuer.
- Dans certaines circonstances, un point de contrôle peut être ignoré. Par exemple, supposons que votre groupe Auto Scaling contienne deux instances et que les pourcentages associés à vos points de contrôle soient de `[10,40,100]`. Une fois la première instance remplacée, Amazon EC2 Auto Scaling calcule que 50 % du groupe a été remplacé. Étant donné que 50 % est un pourcentage supérieur à ceux des deux premiers points de contrôle, il ignore le premier point de contrôle (`10`) et envoie une notification pour le deuxième point de contrôle (`40`).
- L'annulation de l'opération empêche tout remplacement ultérieur. Si vous annulez l'opération ou qu'elle échoue avant d'atteindre le dernier point de contrôle, l'ancienne configuration des instances qui ont déjà été remplacées n'est pas restaurée.
- Dans le cas d'une actualisation partielle, lorsque vous relancez l'opération, Amazon EC2 Auto Scaling ne redémarre pas à partir du dernier point de contrôle et ne s'arrête pas lorsque seules les instances antérieures sont remplacées. Cela dit, il cible d'abord les instances antérieures à remplacer avant de cibler les nouvelles.
- Le pourcentage réel d'achèvement peut être supérieur au pourcentage pour ce point de contrôle lorsque le pourcentage du point de contrôle est trop faible par rapport au nombre d'instances du groupe. Supposons, par exemple, que le pourcentage du point de contrôle soit de 20 % et que le groupe compte quatre instances. Si Amazon EC2 Auto Scaling remplace l'une des quatre instances, le pourcentage remplacé réel (25 %) sera supérieur au pourcentage du point de contrôle (20 %).
- Une fois qu'un point de contrôle est atteint, le pourcentage global d'achèvement affiché n'est mis à jour qu'une fois le préchauffage des instances terminé. Par exemple, vos pourcentages de points de contrôle correspondent `[20,50]` à un délai de 15 minutes et à un pourcentage de santé minimum de 80 %. Votre groupe Auto Scaling compte 10 instances et effectue les remplacements suivants :
  - 0:00 : deux instances antérieures sont remplacées par des nouvelles.
  - 0:10 : deux nouvelles instances finissent leur préparation.

- 0:25 : deux instances antérieures sont remplacées par des nouvelles. (Seulement deux instances sont remplacées pour maintenir le pourcentage minimal valide.)
- 0:35 : deux nouvelles instances finissent leur préparation.
- 0:35 : une instance antérieure est remplacée par une nouvelle.
- 0:45 : une nouvelle instance finit sa préparation.

À 0:35, l'opération cesse de lancer de nouvelles instances. Le pourcentage d'achèvement ne reflète pas encore avec précision le nombre de remplacements terminés (50 %), car la nouvelle instance n'a pas encore terminé sa préparation. Une fois que la nouvelle instance a terminé sa période de préchauffage à 0h45, le pourcentage d'achèvement indique 50 %.

Activer les points de contrôle (console)

Vous pouvez activer des points de contrôle avant de démarrer une actualisation d'instance pour remplacer des instances à l'aide d'une approche progressive ou par phases. Cela donne plus de temps pour la vérification.

Pour lancer une actualisation d'instance qui utilise des points de contrôle

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre dans la partie inférieure de la page Groupes Auto Scaling.
3. Dans l'onglet Instance refresh (Actualisation d'instance), dans Active instance refresh (Actualisation d'instance active), sélectionnez Start instance refresh (Démarrer l'actualisation d'instance).
4. Sur la page Start instance refresh (Lancer une actualisation d'instance), choisissez les valeurs à attribuer dans les champs Minimum healthy percentage (Pourcentage minimal d'instances saines) et Instance warmup (Préparation d'instance).
5. Cochez la case Enable checkpoints (Activer les points de contrôle).

   Vous accédez alors à une zone dans laquelle vous pouvez définir le seuil, en pourcentage, du premier point de contrôle.

6. Pour Proceed until ____ % of the group is refreshed (Continuer jusqu'à ce que ____ % du groupe soit actualisé), saisissez un nombre (entre 1 et 100). Cela définit le pourcentage pour le premier point de contrôle.
7. Pour ajouter un autre point de contrôle, choisissez Add checkpoint (Ajouter un point de contrôle), puis définissez le pourcentage à associer à celui-ci.
8. Pour spécifier le délai à l'issue duquel Amazon EC2 Auto Scaling pourra reprendre l'actualisation d'instance après avoir atteint un point de contrôle, mettez à jour les champs du paramètre Patienter **`1 hour`** entre des points de contrôle. Le temps peut être exprimé en heures, en minutes ou en secondes.
9. Lorsque vous en avez terminé avec les sélections d'actualisation d'instance, sélectionnez Démarrer l'actualisation de l'instance.

Activer les points de contrôle (AWS CLI)

Pour démarrer une actualisation d'instance avec des points de contrôle activés à l'aide de AWS CLI, vous avez besoin d'un fichier de configuration qui définit les paramètres suivants :

- `CheckpointPercentages` : spécifie des valeurs de seuil pour le pourcentage d'instances à remplacer. Ces valeurs de seuil fournissent les points de contrôle. Lorsque le pourcentage d'instances remplacées et prêtes atteint l'un des seuils spécifiés, l'opération attend la fin du délai spécifié. Vous devez spécifier ce délai (en secondes) dans `CheckpointDelay`. Une fois le délai spécifié écoulé, l'actualisation d'instance reprend jusqu'à ce qu'elle atteigne le point de contrôle suivant (le cas échéant).
- `CheckpointDelay` : spécifie le délai, en secondes, à l'issue duquel l'actualisation d'instance pourra reprendre après avoir atteint un point de contrôle. Choisissez une période qui offre suffisamment de temps pour effectuer vos vérifications.

La dernière valeur affichée dans le tableau `CheckpointPercentages` décrit le pourcentage du groupe Auto Scaling qui doit être remplacé avec succès. L'opération affiche `Successful` une fois que ce pourcentage a été remplacé avec succès et que chaque instance est considérée comme ayant terminé son initialisation.

Pour créer plusieurs points de contrôle

Pour créer plusieurs points de contrôle, utilisez l'exemple de commande start-instance-refresh suivant. Cet exemple illustre la configuration d'une actualisation d'instance qui actualise initialement

1 % du groupe Auto Scaling. Après avoir attendu 10 minutes, il actualise ensuite les 19 % suivants et attend encore 10 minutes. Enfin, il rafraîchit le reste du groupe avant de conclure l'opération.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Contenu de `config.json` :

```
{
    "AutoScalingGroupName": "my-asg",
    "Preferences": {
      "InstanceWarmup": 60,
      "MinHealthyPercentage": 80,
      "CheckpointPercentages": [1,20,100],
      "CheckpointDelay": 600
    }
}
```

Pour créer un point de contrôle unique

Pour créer un point de contrôle unique, utilisez l'exemple de commande start-instance-refresh suivant. Cet exemple illustre la configuration d'une actualisation d'instance qui actualise initialement 20 % du groupe Auto Scaling. Après avoir attendu 10 minutes, il actualise ensuite le reste du groupe avant de conclure l'opération.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Contenu de `config.json` :

```
{
    "AutoScalingGroupName": "my-asg",
    "Preferences": {
      "InstanceWarmup": 60,
      "MinHealthyPercentage": 80,
      "CheckpointPercentages": [20,100],
      "CheckpointDelay": 600
    }
}
```

Pour actualiser partiellement le groupe Auto Scaling

Pour remplacer uniquement une partie de votre groupe Auto Scaling, puis arrêter complètement, utilisez l'exemple de commande start-instance-refresh suivant. Cet exemple illustre la configuration d'une actualisation d'instance qui actualise initialement 1 % du groupe Auto Scaling. Après avoir attendu 10 minutes, il actualise ensuite les 19 % suivants avant de conclure l'opération.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Contenu de `config.json` :

```
{
    "AutoScalingGroupName": "my-asg",
    "Preferences": {
      "InstanceWarmup": 60,
      "MinHealthyPercentage": 80,
      "CheckpointPercentages": [1,20],
      "CheckpointDelay": 600
    }
}
```

Remplacer des instances Auto Scaling en fonction de la durée de vie maximale de l'instance

La durée de vie maximale de l'instance spécifie la durée maximale (en secondes) pendant laquelle une instance peut être en service avant d'être résiliée et remplacée. Il arrive couramment que vous deviez remplacer vos instances selon un calendrier en raison de politiques de sécurité internes ou de contrôles de conformité externes.

Vous devez spécifier une valeur d'au moins 86 400 secondes (un jour). Pour effacer une valeur précédemment définie, spécifiez une nouvelle valeur de 0. Ce paramètre s'applique à toutes les instances actuelles et futures de votre groupe Auto Scaling.

Table des matières

Considérations

Les points suivants doivent être pris en compte lors de l'utilisation de cette fonctionnalité :

- Chaque fois qu'une instance antérieure est remplacée et qu'une nouvelle instance est lancée, la nouvelle utilise le modèle de lancement ou la configuration de lancement actuellement associée au groupe Auto Scaling. Si votre modèle de lancement ou votre configuration de lancement spécifie l'ID Amazon Machine Image (AMI) d'une autre version de votre application, cette version de votre application sera déployée automatiquement.
- Si la durée de vie maximale des instances est trop faible, les instances peuvent être remplacées plus rapidement que prévu. Amazon EC2 Auto Scaling remplace généralement les instances une par une, avec une pause entre les remplacements. Toutefois, si la durée de vie maximale des instances spécifiée ne laisse pas suffisamment de temps pour remplacer chaque instance individuellement, Amazon EC2 Auto Scaling doit remplacer plusieurs instances à la fois. Plusieurs instances peuvent être remplacées à la fois, jusqu'à 10 % de la capacité actuelle de votre groupe Auto Scaling. Pour éviter de remplacer un trop grand nombre d'instances à la fois, définissez une durée de vie maximale des instances plus longue ou utilisez la protection évolutive des instances pour empêcher temporairement la mise hors service d'instances individuelles. Pour plus d'informations, consultez Utiliser la protection de la taille d'instance.
- Amazon EC2 Auto Scaling crée une nouvelle activité de mise à l'échelle pour résilier l'instance, puis la résilie. Pendant que l'instance est résiliée, une autre activité de mise à l'échelle lance une nouvelle instance. Vous pouvez modifier ce comportement pour lancer avant toute résiliation en utilisant une politique de maintenance des instances. Pour plus d'informations, consultez Politiques de maintenance des instances.

Définir la durée de vie maximale de l'instance

Lorsque vous créez un groupe Auto Scaling dans la console, vous ne pouvez pas définir la durée de vie maximale d'une instance. Cependant, après la création du groupe, vous pouvez modifier celui-ci pour définir la durée de vie maximale de l'instance.

Pour définir la durée de vie maximale d'une instance pour un groupe (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

Un panneau fractionné s'ouvre dans la partie inférieure de la page Groupes Auto Scaling avec des informations sur le groupe que vous avez sélectionné.

3. Sous l'onglet Détails, choisissez Configurations avancées, Modifier.
4. Pour Maximum instance lifetime (Durée de vie maximale de l'instance), saisissez le nombre maximal de secondes pendant lesquelles une instance peut être en service.
5. Choisissez Mettre à jour.

L'onglet Activity (Activité), sous Activity history (Historique des activités), vous permet de voir l'historique de remplacement des instances du groupe.

Pour définir la durée de vie maximale d'une instance pour un groupe (AWS CLI)

Vous pouvez également utiliser le AWS CLI pour définir la durée de vie maximale des instances pour les groupes Auto Scaling nouveaux ou existants.

Pour les nouveaux groupes Auto Scaling, utilisez la commande create-auto-scaling-group.

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

Voici un exemple de fichier `config.json` qui montre une durée de vie maximale de l'instance de `2592000` secondes (30 jours).

```
{
    "AutoScalingGroupName": "my-asg",
    "LaunchTemplate": {
        "LaunchTemplateName": "my-launch-template",
        "Version": "$Default"
    },
    "MinSize": 1,
    "MaxSize": 5,
    "MaxInstanceLifetime": 2592000,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
    "Tags": []
}
```

Pour les groupes Auto Scaling existants, utilisez la commande update-auto-scaling-group.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-existing-asg --
max-instance-lifetime 2592000
```

Pour vérifier la durée de vie maximale d'instance d'un groupe Auto Scaling

Utilisez la commande describe-auto-scaling-groups.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Limites

- La durée de vie maximale n'est pas garantie d'être exacte pour chaque instance : les instances ne sont pas garanties d'être remplacées uniquement à la fin de leur durée maximale. Dans certains cas, Amazon EC2 Auto Scaling peut avoir besoin de lancer le remplacement des instances juste après que vous ayez mis à jour le paramètre de durée de vie maximale de l'instance. La raison de ce comportement est d'éviter de remplacer toutes les instances en même temps.
- Protection évolutive des instances respectée : Amazon EC2 Auto Scaling fournit une protection évolutive des instances pour vous aider à contrôler les instances auxquelles elle peut mettre fin. Lorsque cette protection est activée sur une instance, Amazon EC2 Auto Scaling ne met pas fin à l'instance même si elle a atteint sa durée de vie maximale.
- Instances résiliées avant le lancement : lorsqu'il n'y a qu'une seule instance dans le groupe Auto Scaling, la fonctionnalité de durée de vie maximale des instances peut provoquer une panne car Amazon EC2 Auto Scaling résilie une instance avant d'en lancer une nouvelle. Pour modifier ce comportement afin de lancer avant toute résiliation, consultez Politiques de maintenance des instances.

Mettre la taille de votre groupe Auto Scaling à l'échelle

La mise à l'échelle est la capacité à augmenter ou à diminuer la capacité de calcul d'une application. La mise à l'échelle commence par un événement ou une action qui demande à un groupe Auto Scaling de lancer ou de résilier les instances Amazon EC2.

Amazon EC2 Auto Scaling fournit plusieurs moyens d'ajuster la mise à l'échelle pour mieux répondre aux besoins des applications. Par conséquent, il est important que vous compreniez bien l'application. Gardez les considérations suivantes à l'esprit :

- Quel rôle Amazon EC2 Auto Scaling doit jouer dans l'architecture de l'application ? Il est fréquent de considérer la scalabilité automatique comme un moyen d'augmenter ou de diminuer la capacité, mais il est également utile pour maintenir un nombre stable de serveurs.
- Quelles sont les contraintes de coût importantes pour vous ? Amazon EC2 Auto Scaling utilisant des instances EC2, vous ne payez que pour les ressources que vous utilisez. Le fait de connaître les contraintes de coût vous aide à décider quand vous allez mettre à l'échelle les applications et dans quelle mesure.
- Quelles sont les métriques importantes pour votre application ? Amazon CloudWatch prend en charge un certain nombre de mesures différentes que vous pouvez utiliser avec votre groupe Auto Scaling.

Table des matières

Choisissez votre méthode de mise à l’échelle

Amazon EC2 Auto Scaling vous propose plusieurs moyens de mettre à l'échelle votre groupe Auto Scaling.

Maintenez un nombre fixe d'instances

La valeur par défaut d’un groupe Auto Scaling est de ne pas être associé à des politiques de mise à l’échelle ou à des actions planifiées, ce qui lui permet de conserver une taille fixe. Après avoir créé votre groupe Auto Scaling, celui-ci démarre en lançant suffisamment d’instances pour atteindre la capacité souhaitée. Si aucune condition de mise à l’échelle n’est associée au groupe, celui-ci continue à maintenir la capacité souhaitée même si une instance devient défectueuse. Amazon EC2 Auto Scaling surveille l’état de chaque instance de votre groupe Auto Scaling. Lorsqu’il détecte qu’une instance devient défectueuse, il la remplace par une nouvelle instance. Vous pouvez lire une description plus détaillée de ce processus dans Surveillance de l’état des instances dans un groupe Auto Scaling.

Mise à l'échelle manuelle

La mise à l’échelle manuelle est le moyen le plus basique de mettre à l’échelle votre groupe Auto Scaling. Vous pouvez soit mettre à jour la capacité souhaitée du groupe Auto Scaling, soit mettre fin à des instances du groupe Auto Scaling. Pour plus d’informations, consultez Mise à l'échelle manuelle pour Amazon EC2 Auto Scaling.

Mise à l'échelle selon un calendrier

Le dimensionnement par calendrier signifie que les actions de dimensionnement sont effectuées automatiquement en fonction de la date et de l'heure. Il arrive que vous sachiez exactement quand vous aurez besoin d'augmenter ou de diminuer le nombre d'instances dans le groupe, simplement parce que ce besoin résulte d'un calendrier prévisible. Pour plus d’informations, consultez Mise à l'échelle planifiée pour Amazon EC2 Auto Scaling.

Évoluez dynamiquement en fonction de la demande

Une méthode plus avancée de mise à l’échelle de vos ressources, à l’aide de la mise à l’échelle dynamique, vous permet de définir une politique de mise à l’échelle qui redimensionne dynamiquement votre groupe Auto Scaling pour répondre aux changements de demande. Par exemple, vous avez une application Web qui s'exécute actuellement sur deux instances et vous souhaitez que l'utilisation de l'UC du groupe Auto Scaling reste à environ 50 % lorsque la charge sur l'application change. Cette méthode est utile pour effectuer une mise à l'échelle au fur et à mesure

que le trafic change, lorsque vous ne savez pas quand le trafic va changer. Vous pouvez configurer des politiques de mise à l'échelle pour qu'elles répondent à votre place. Il existe plusieurs types de politiques (ou une combinaison des deux) que vous pouvez utiliser pour évoluer en fonction de l'évolution du trafic. Pour plus d'informations, consultez Mise à l'échelle dynamique pour Amazon EC2 Auto Scaling.

Évoluez de manière proactive

Vous pouvez également combiner la mise à l'échelle prédictive et la mise à l'échelle dynamique (les approches proactive et réactive, respectivement) pour une adaptation plus rapide de votre capacité EC2. Utilisez la mise à l'échelle prédictive pour augmenter le nombre d'instances EC2 dans votre groupe Auto Scaling en anticipant les tendances quotidiennes et hebdomadaires en matière de flux de trafic. Pour plus d'informations, consultez Mise à l'échelle prédictive pour Amazon EC2 Auto Scaling.

Définissez des limites de mise à l'échelle pour votre groupe Auto Scaling

Les limites de mise à l'échelle représentent les tailles minimale et maximale de groupe que vous désirez pour votre groupe Auto Scaling. Vous définissez des limites séparément pour la taille minimale et la taille maximale.

La capacité souhaitée du groupe peut être ajustée à un nombre compris dans la plage des limites de tailles minimale et maximale. Cette capacité souhaitée doit être supérieure ou égale à la taille minimale du groupe et inférieure ou égale à la taille maximale du groupe.

- Capacité souhaitée : fait référence à la capacité initiale du groupe Auto Scaling à sa création. Le groupe Auto Scaling s'efforce de conserver la capacité souhaitée. Il démarre en lançant le nombre d'instances spécifiées pour la capacité souhaitée et conserve ce nombre d'instances aussi longtemps qu'il n'existe aucune politique de mise à l'échelle ou d'actions planifiées associées au groupe Auto Scaling.
- Capacité minimale : représente la taille minimale du groupe. Une fois des politiques de mise à l'échelle définies, la capacité du groupe ne peut être réduite en dessous de la limite de taille minimale.
- Capacité maximale : représente la taille maximale du groupe. Une fois des politiques de mise à l'échelle définies, la capacité du groupe ne peut être augmentée au-dessus de la limite de taille maximale.

Les limites de tailles minimale et maximale s'appliquent également aux scénarios suivants :

- en cas de mise à l'échelle manuelle de votre groupe Auto Scaling à travers une mise à jour de sa capacité souhaitée.
- Lorsque des actions planifiées sont exécutées, la capacité souhaitée est mise à jour. Si une action planifiée s'exécute sans aucune indication des nouvelles limites de taille minimale et maximale du groupe, les limites de taille minimale et maximale actuelles du groupe s'appliqueront.

Le groupe Auto Scaling s'efforce toujours de conserver la capacité souhaitée. Si une instance se ferme de façon inattendue (en raison d'une interruption de l'instance spot, d'un échec de la surveillance de l'état de santé ou d'une action humaine par exemple), le groupe lancera automatiquement une nouvelle instance afin de maintenir la capacité souhaitée.

Pour gérer ces paramètres dans la console

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Dans le volet de navigation, sous Auto Scaling, choisissez Auto Scaling Groups (Groupes Auto Scaling).
3. Dans la page des groupes Auto Scaling, cochez la case située en regard de votre groupe Auto Scaling.

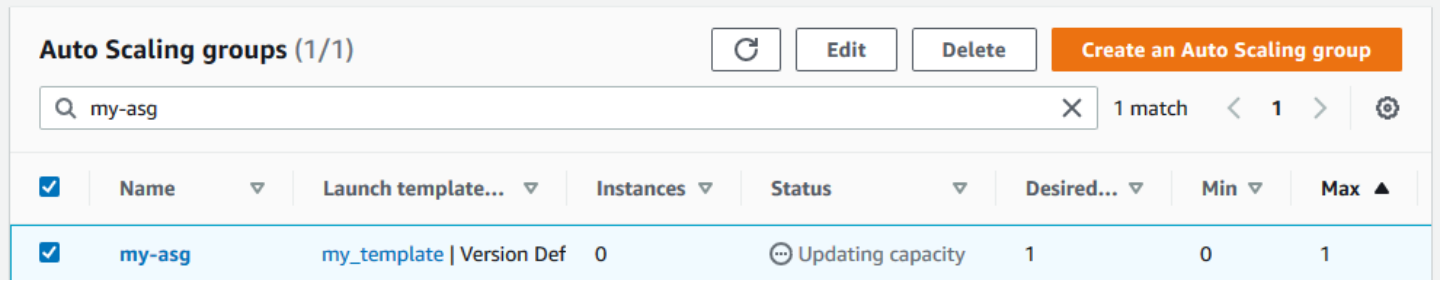
   Un volet fractionné s’ouvre en bas de la page.
4. Sous l’onglet Détails dans le volet inférieur, affichez ou modifiez les paramètres actuels pour la capacité souhaitée, minimale et maximale. Pour plus d’informations, consultez Changer la capacité souhaitée d’un groupe Auto Scaling existant.

Au-dessus du volet Détails, vous trouverez des informations telles que le nombre actuel d’instances dans le groupe Auto Scaling, la capacité souhaitée, minimale et maximale, ainsi qu’une colonne d’état. Si le groupe Auto Scaling utilise des poids d'instance, vous pouvez également trouver le nombre d'unités de capacité ayant contribué à la capacité souhaitée.

Pour ajouter ou supprimer des colonnes de la liste, choisissez l'icône Paramètres en haut de la page. Ensuite, pour Attributs de groupes Auto Scaling, activez ou désactivez chaque colonne, puis choisissez Confirmer.

Pour vérifier la taille de votre groupe Auto Scaling une fois les modifications effectuées

La colonne Instances affiche le nombre d'instances en cours d'exécution. Pendant le lancement ou la fin d'une instance, la colonne Status (État) affiche l'état Mise à jour de la capacité, comme indiqué dans l'image suivante.



Attendez quelques minutes, puis actualisez la vue pour voir le dernier état. Une fois la mise à l'échelle terminée, la colonne Instances affichera une nouvelle valeur.

Vous pouvez voir le nombre d'instances et l'état des instances en cours d'exécution dans l'onglet Gestion des instances sous Instances.

Définir la préparation par défaut d'instance d'un groupe Auto Scaling

CloudWatch collecte et agrège les données d'utilisation, telles que le processeur et les E/S réseau, sur vos instances Auto Scaling. Vous utilisez ces métriques pour créer des politiques de mise à l'échelle qui ajustent le nombre d'instances de votre groupe Auto Scaling lorsque la valeur de la métrique sélectionnée augmente et diminue.

Vous pouvez spécifier le délai après qu'une instance a atteint l'`InService`état dans lequel elle attend avant de fournir des données d'utilisation aux métriques agrégées. Cette durée spécifiée est appelée préchauffage de l'instance par défaut. Cela permet d'éviter que le dimensionnement dynamique ne soit affecté par les métriques relatives à des instances individuelles qui ne gèrent pas encore le trafic applicatif et qui sont susceptibles de connaître une utilisation temporairement élevée des ressources de calcul.

Pour optimiser les performances de vos politiques de suivi des cibles et de dimensionnement par étapes, nous vous recommandons vivement d'activer et de configurer le préchauffage de l'instance par défaut. Il n'est ni activé ni configuré par défaut.

Lorsque vous activez le préchauffage d'instance par défaut, n'oubliez pas que si votre groupe Auto Scaling est configuré pour utiliser une politique de maintenance des instances, ou si vous utilisez

une actualisation d'instance pour remplacer des instances, vous pouvez empêcher que les instances ne soient prises en compte dans le calcul du pourcentage de santé minimum avant la fin de leur initialisation.

Table des matières

Considérations sur les performances de la mise à l’échelle

Il est utile pour la plupart des applications d'avoir un temps de préchauffage d'instance par défaut qui s'applique à toutes les fonctionnalités, plutôt que des temps de préchauffage différents pour les différentes fonctionnalités. Par exemple, si vous ne définissez pas de préchauffage d'instance par défaut, la fonction d'actualisation de l'instance utilise le délai de grâce du bilan de santé comme temps de préchauffage par défaut. Si vous avez des politiques de suivi des cibles et de dimensionnement des étapes, elles utilisent la valeur définie pour le temps de recharge par défaut comme temps de préchauffage par défaut. Si vous avez des politiques de dimensionnement prédictif, elles n'ont pas de temps de préchauffage par défaut.

Pendant le préchauffage des instances, vos politiques de dimensionnement dynamique ne sont redimensionnées que si la valeur métrique des instances qui ne s'échauffent pas est supérieure au seuil d'alarme de la politique (ou à l'utilisation cible d'une politique de dimensionnement de suivi des cibles). Si la demande diminue, le dimensionnement dynamique devient plus prudent afin de protéger la disponibilité de votre application. Cela bloque les activités de mise à l'échelle pour une mise à l'échelle dynamique jusqu'à ce que les nouvelles instances aient fini de s'échauffer.

Lors de la mise à l'échelle, Amazon EC2 Auto Scaling prend en compte les instances en phase d'échauffement dans le cadre de la capacité du groupe lorsqu'il décide du nombre d'instances à ajouter au groupe. Par conséquent, plusieurs violations d'alarme nécessitant l'ajout d'une capacité similaire entraînent une seule activité de mise à l'échelle. L'objectif est de continuer à évoluer, sans le faire de manière excessive.

Si le préchauffage de l'instance par défaut n'est pas activé, le temps qu'une instance attend avant d'envoyer des métriques CloudWatch et de les compter dans la capacité actuelle varie d'une instance à l'autre. Il est donc possible que vos politiques de dimensionnement fonctionnent de manière imprévisible par rapport à la charge de travail réelle qui se produit.

Prenons l'exemple d'une application avec un schéma de on-and-off charge de travail récurrent. Une politique de mise à l'échelle prédictive est utilisée pour prendre des décisions récurrentes quant à l'augmentation du nombre d'instances. Comme il n'existe pas de temps de préchauffage par défaut pour les politiques de dimensionnement prédictif, les instances commencent immédiatement à contribuer aux métriques agrégées. Si ces instances utilisent davantage de ressources au démarrage, l'ajout d'instances peut entraîner une hausse des métriques agrégées. En fonction du temps nécessaire à la stabilisation de l'utilisation, cela peut avoir un impact sur les politiques de mise à l'échelle dynamique utilisant ces métriques. Si le seuil d'alarme supérieur d'une politique de mise à l'échelle dynamique est dépassé, la taille du groupe augmente à nouveau. Pendant la préparation des nouvelles instances, les activités de mise à l'échelle horizontale seront bloquées.

Choisissez le temps de préchauffage de l'instance par défaut

Pour définir la préparation d'instance par défaut, il est essentiel de déterminer le temps dont vos instances ont besoin pour terminer leur initialisation et pour que la consommation de ressources se stabilise une fois qu'elles ont atteint l'état `InService`. Lorsque vous choisissez le temps de préchauffage de l'instance, essayez de maintenir un équilibre optimal entre la collecte de données d'utilisation pour le trafic légitime et la minimisation de la collecte de données associée aux pics d'utilisation temporaires au démarrage.

Supposons qu'un groupe Auto Scaling soit associé à un équilibreur de charge Elastic Load Balancing. Lorsque le lancement des nouvelles instances est terminé, elles sont enregistrées dans l'équilibreur de charge avant d'entrer dans l'état `InService`. Une fois que les instances passent à l'état `InService`, la consommation de ressources peut encore connaître des pics temporaires et avoir besoin de temps pour se stabiliser. Par exemple, la consommation de ressources pour un serveur d'applications qui doit télécharger et mettre en cache des ressources volumineuses prend plus de temps à stabiliser qu'un serveur Web léger sans ressources volumineuses à télécharger. La préparation d'instance fournit le délai nécessaire à la stabilisation de la consommation de ressources.

> **Important**
>
> Si vous n'êtes pas sûr du temps dont vous avez besoin pour le préchauffage, vous pouvez commencer par 300 secondes. Puis diminuez-le ou augmentez-le progressivement

jusqu'à obtenir les meilleures performances de mise à l'échelle pour votre application. Vous devrez peut-être le faire plusieurs fois pour bien faire les choses. Sinon, si vous avez des politiques de dimensionnement dotées de leur propre temps de préchauffage (`EstimatedInstanceWarmup`), vous pouvez utiliser cette valeur pour commencer. Pour plus d'informations, consultez Trouvez des politiques de dimensionnement avec un temps de préchauffage de l'instance défini au préalable.

Vous pourriez également envisager d'utiliser des hooks de cycle de vie pour les cas d'utilisation dans lesquels vous avez des tâches de configuration ou des scripts à exécuter au démarrage. Les hooks de cycle de vie peuvent retarder la mise en service des instances jusqu'à ce que leur initialisation soit terminée. Ils sont particulièrement utiles si vous disposez de scripts d'amorçage qui nécessitent un certain temps. Si vous ajoutez un hook de cycle de vie, vous pouvez réduire la valeur de la préparation d'instance par défaut. Pour plus d'informations sur l'utilisation des hooks de cycle de vie, consultez Hooks de cycle de vie Amazon EC2 Auto Scaling.

Activer la préparation d'instance par défaut pour un groupe

Vous pouvez activer la préparation d'instance par défaut au moment de la création d'un groupe Auto Scaling. Vous pouvez également l'activer pour les groupes existants.

En activant la fonction de préchauffage de l'instance par défaut, vous n'avez plus besoin de spécifier des valeurs pour les paramètres de préchauffage pour les fonctionnalités suivantes :

- Actualisation d'instance
- Mise à l'échelle de suivi de cible
- Mise à l'échelle par étapes

Console

Pour activer la préparation d'instance par défaut pour un nouveau groupe (console)

Lorsque vous créez le groupe Auto Scaling, sur la page Configure advanced options (Configurer des options avancées), sous Additional settings (Paramètres supplémentaires), sélectionnez l'option Enable default instance warmup (Activer la préparation d'instance par défaut). Choisissez le temps de préchauffage dont vous avez besoin pour votre application.

AWS CLI

Pour activer la préparation d'instance par défaut pour un nouveau groupe (AWS CLI)

Pour activer la préparation d'instance par défaut pour un groupe Auto Scaling, ajoutez l'option `--default-instance-warmup` et spécifiez une valeur, en secondes, comprise entre 0 et 3 600. Une fois qu'elle est activée, une valeur de `-1` va désactiver ce réglage.

Procédez comme suit : la commande create-auto-scaling-group crée un groupe Auto Scaling portant le nom *mon-asg* et active la préparation d'instance par défaut avec une valeur de *120* secondes.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg --
default-instance-warmup 120 ...
```

> **Tip**
>
> Si cette commande génère une erreur, assurez-vous d'avoir mis à jour AWS CLI localement la dernière version.

Console

Pour activer la préparation d'instance par défaut pour un groupe existant (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Dans la barre de navigation située en haut de l'écran, choisissez l' Région AWS dans laquelle vous avez créé votre groupe Auto Scaling.
3. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
4. Sous l'onglet Détails, choisissez Configurations avancées, Modifier.
5. Pour le préchauffage de l'instance par défaut, choisissez le temps de préchauffage dont vous avez besoin pour votre application.
6. Choisissez Mettre à jour.

AWS CLI

Pour activer la préparation d'instance par défaut pour un groupe existant (AWS CLI)

L'exemple suivant utilise la commande update-auto-scaling-group pour activer la préparation d'instance par défaut avec une valeur de *120* secondes pour un groupe Auto Scaling existant portant le nom de *my-asg*.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg --
default-instance-warmup 120
```

> **Tip**
>
> Si cette commande génère une erreur, assurez-vous d'avoir mis à jour AWS CLI localement la dernière version.

Vérifier la préparation d'instance par défaut pour un groupe

Pour vérifier la préparation d'instance par défaut d'un groupe Auto Scaling (AWS CLI)

Utilisez la commande describe-auto-scaling-groups suivante. Remplacez *my-asg* par le nom de votre groupe Auto Scaling.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Voici un exemple de réponse.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            ...
            "DefaultInstanceWarmup": 120
        }
    ]
}
```

Trouvez des politiques de dimensionnement avec un temps de préchauffage de l'instance défini au préalable

Pour déterminer si vos politiques ont leur propre temps de préchauffage pour`EstimatedInstanceWarmup`, exécutez la commande describe-policies suivante à l'aide de. AWS CLI Remplacez *my-asg* par le nom de votre groupe Auto Scaling.

```
aws autoscaling describe-policies --auto-scaling-group-name my-asg
  --query 'ScalingPolicies[?EstimatedInstanceWarmup!=`null`]'
```

Voici un exemple de sortie.

```
[
  {
    "AutoScalingGroupName":"my-asg",
    "PolicyName":"cpu50-target-tracking-scaling-policy",
    "PolicyARN":"arn",
    "PolicyType":"TargetTrackingScaling",
    "StepAdjustments":[],
    "EstimatedInstanceWarmup":120,
    "Alarms":[{
        "AlarmARN": "arn:aws:cloudwatch:us-west-2:123456789012:alarm:TargetTracking-my-
asg-AlarmHigh-fc0e4183-23ac-497e-9992-691c9980c38e",
        "AlarmName": "TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e"
      },
      {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2",
            "AlarmName": "TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2"
    }],
    "TargetTrackingConfiguration":{
      "PredefinedMetricSpecification":{
        "PredefinedMetricType":"ASGAverageCPUUtilization"
      },
      "TargetValue":50.0,
      "DisableScaleIn":false
    },
    "Enabled":true
  },
```

```
    ... additional policies ...

]
```

Effacer la préparation de l’instance définie précédemment pour une politique de mise à l’échelle

Après avoir activé le préchauffage de l'instance par défaut, mettez à jour toutes les politiques de dimensionnement qui ont encore leur propre temps de préchauffage pour effacer la valeur précédemment définie. Dans le cas contraire, elle remplacera la préparation de l’instance par défaut.

Vous pouvez mettre à jour les politiques de dimensionnement à l'aide de AWS CLI la console ou AWS des SDK. Cette section décrit les étapes relatives à la console. Si vous utilisez les AWS SDK AWS CLI ou, assurez-vous de conserver la configuration de politique existante, mais supprimez la `EstimatedInstanceWarmup` propriété. Lorsque vous mettez à jour une politique de dimensionnement existante, celle-ci est remplacée par celle que vous spécifiez lorsque vous appelez PutScaling Policy par programmation. Les valeurs d’origine ne sont pas conservées.

Pour effacer la préparation de l’instance précédemment définie pour une politique de mise à l’échelle (console)

1. Ouvrez la console Amazon EC2 à l’adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s’ouvre en bas de la page.
3. Dans l’onglet Mise à l’échelle automatique, dans Politiques de mise à l’échelle dynamiques, choisissez la politique qui vous intéresse, puis choisissez Actions, Modifier.
4. Pour Instance warmup, effacez la valeur de préchauffage de l'instance pour utiliser la valeur de préchauffage de l'instance par défaut à la place.
5. Choisissez Mettre à jour.

Mise à l'échelle manuelle pour Amazon EC2 Auto Scaling

Vous pouvez ajuster manuellement le nombre d'instances EC2 dans votre groupe Auto Scaling à tout moment. Ce processus de modification manuelle du nombre d'instances est appelé dimensionnement

manuel. La mise à l'échelle manuelle est une alternative à la mise à l'échelle automatique, en particulier si vous souhaitez apporter des modifications de capacité ponctuelles.

Après avoir redimensionné manuellement votre groupe, Amazon EC2 Auto Scaling reprend les activités de dimensionnement automatique normales en fonction des politiques de dimensionnement et des actions planifiées que vous avez définies. Pour les groupes pour lesquels le préchauffage d'instance par défaut est activé, toute nouvelle instance passe par une période de préchauffage avant de commencer à contribuer aux mesures utilisées pour le dimensionnement automatique. Cette période d'échauffement aide à stabiliser le groupe à sa nouvelle capacité. Pour plus d'informations, consultez Définir la préparation par défaut d'instance d'un groupe Auto Scaling.

Parfois, vous souhaiterez peut-être désactiver temporairement les politiques de dimensionnement et les actions planifiées avant de redimensionner manuellement un groupe. Cela permet d'éviter les conflits entre les actions de dimensionnement manuelles et les activités de dimensionnement automatisées. Pour plus d'informations, consultez Désactiver les activités de dimensionnement.

Table des matières

Changer la capacité souhaitée d'un groupe Auto Scaling existant

Lorsque vous modifiez la capacité souhaitée de votre groupe Auto Scaling, Amazon EC2 Auto Scaling gère le processus de lancement et de résiliation des instances pour atteindre la nouvelle taille souhaitée.

Console

Pour modifier la taille de votre groupe Auto Scaling

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet divisé s'affiche au bas de la page.
3. Sous l'onglet Details (Détails) choisissez Group details (Détails du groupe), Edit (Modifier).
4. Pour Capacité souhaitée, augmentez ou diminuez la capacité souhaitée. Par exemple, pour augmenter la taille du groupe d'une unité, si la valeur actuelle est1, entrez2.

Si votre nouvelle valeur pour la capacité souhaitée est supérieure à la capacité minimale souhaitée et à la capacité maximale souhaitée, la capacité maximale souhaitée est automatiquement augmentée à la nouvelle valeur de capacité souhaitée.

5. Une fois que vous avez terminé, choisissez Update (Mettre à jour).

Vérifiez que la taille de groupe que vous avez spécifiée a entraîné le lancement du même nombre d'instances. Par exemple, si vous avez augmenté la taille du groupe d'une unité, vérifiez que votre groupe Auto Scaling a lancé une instance supplémentaire.

Pour vérifier que la taille du groupe Auto Scaling a changé

1. Dans l'onglet Activity, dans l'historique des activités, vous pouvez voir la progression des activités associées au groupe Auto Scaling. La colonne Status (État) affiche l'état actuel de votre instance. Lorsqu'une instance est en cours de lancement, son statut est `Not yet in service`. Le statut passe à `Successful`, après le lancement de l'instance. Vous pouvez également utiliser l'icône d'actualisation pour voir l'état actuel de votre instance. Pour plus d'informations, consultez Vérifier une activité de mise à l'échelle pour un groupe Auto Scaling.
2. Dans l'onglet Gestion des instances, dans Instances, vous pouvez consulter le statut de l'instance. Il suffit de peu de temps pour lancer une instance.

   - La colonne Lifecycle (Cycle de vie) affiche l'état de votre instance. Initialement, votre instance est à l'état `Pending`. Lorsqu'une instance est prête à recevoir du trafic, son statut passe à `InService`.
   - La colonne État de santé affiche le résultat des tests de santé effectués par Amazon EC2 Auto Scaling sur votre instance.

AWS CLI

L'exemple suivant suppose que vous avez créé un groupe Auto Scaling avec une taille minimum de 1 et maximum de 5. Par conséquent, le groupe dispose actuellement d'une seule instance en cours d'exécution.

Pour modifier la taille de votre groupe Auto Scaling

Utilisez la commande set-desired-capacity pour modifier la taille du groupe Auto Scaling, comme illustré dans l'exemple suivant.

```
aws autoscaling set-desired-capacity --auto-scaling-group-name my-asg \
  --desired-capacity 2
```

Si vous choisissez de respecter le temps de stabilisation par défaut pour votre groupe Auto Scaling, vous devez spécifier l'option `--honor-cooldown`, comme illustré dans l'exemple suivant. Pour plus d'informations, consultez Temps de stabilisation de la mise à l'échelle d'Amazon EC2 Auto Scaling.

```
aws autoscaling set-desired-capacity --auto-scaling-group-name my-asg \
  --desired-capacity 2 --honor-cooldown
```

Pour vérifier la taille de votre groupe Auto Scaling

Utilisez la commande describe-auto-scaling-groups pour confirmer que la taille du groupe Auto Scaling a changé, comme illustré dans l'exemple suivant :

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Voici un exemple de sortie qui fournit des détails sur le groupe et les instances lancés.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "MinSize": 1,
            "MaxSize": 5,
            "DesiredCapacity": 2,
            "DefaultCooldown": 300,
            "AvailabilityZones": [
                "us-west-2a"
            ],
            "LoadBalancerNames": [],
            "TargetGroupARNs": [],
            "HealthCheckType": "EC2",
            "HealthCheckGracePeriod": 300,
```

```
            "Instances": [
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
                        "LaunchTemplateId": "lt-050555ad16a3f9c7f"
                    },
                    "InstanceId": "i-05b4f7d5be44822a6",
                    "InstanceType": "t3.micro",
                    "HealthStatus": "Healthy",
                    "LifecycleState": "Pending"
                },
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
                        "LaunchTemplateId": "lt-050555ad16a3f9c7f"
                    },
                    "InstanceId": "i-0c20ac468fa3049e8",
                    "InstanceType": "t3.micro",
                    "HealthStatus": "Healthy",
                    "LifecycleState": "InService"
                }
            ],
            "CreatedTime": "2019-03-18T23:30:42.611Z",
            "SuspendedProcesses": [],
            "VPCZoneIdentifier": "subnet-c87f2be0",
            "EnabledMetrics": [],
            "Tags": [],
            "TerminationPolicies": [
                "Default"
            ],
            "NewInstancesProtectedFromScaleIn": false,
            "ServiceLinkedRoleARN": "arn",
            "TrafficSources": []
        }
    ]
}
```

Notez que `DesiredCapacity` affiche la nouvelle valeur. Le groupe Auto Scaling a lancé une instance supplémentaire.

Résilier une instance de votre groupe Auto Scaling (AWS CLI)

Il peut arriver que vous souhaitiez effectuer une mise à l'échelle horizontale manuelle dans votre groupe Auto Scaling tout en mettant fin à une instance spécifique. Vous pouvez effectuer une mise à l'échelle horizontale manuelle de votre groupe Auto Scaling en utilisant la commande terminate-instance-in-auto-scaling-group et en spécifiant l'ID de l'instance que vous souhaitez résilier et l'option `--should-decrement-desired-capacity`, comme indiqué dans l'exemple suivant.

```
aws autoscaling terminate-instance-in-auto-scaling-group \
  --instance-id i-026e4c9f62c3e448c --should-decrement-desired-capacity
```

Voici un exemple de sortie qui fournit des détails sur l'activité de dimensionnement.

```
{
    "Activities": [
        {
            "ActivityId": "b8d62b03-10d8-9df4-7377-e464ab6bd0cb",
            "AutoScalingGroupName": "my-asg",
            "Description": "Terminating EC2 instance: i-026e4c9f62c3e448c",
            "Cause": "At 2023-09-23T06:39:59Z instance i-026e4c9f62c3e448c was taken
 out of service in response to a user request, shrinking the capacity from 1 to 0.",
            "StartTime": "2023-09-23T06:39:59.015000+00:00",
            "StatusCode": "InProgress",
            "Progress": 0,
            "Details": "{\"Subnet ID\":\"subnet-6194ea3b\",\"Availability Zone\":\"us-
west-2c\"}"
        }
    ]
}
```

Cette option n'est pas disponible dans la console. Cependant, vous pouvez utiliser la page Instances de la console Amazon EC2 pour mettre fin à une instance de votre groupe Auto Scaling. Lorsque vous le faites, Amazon EC2 Auto Scaling détecte que l'instance n'est plus en cours d'exécution et la remplace automatiquement dans le cadre du processus de contrôle de santé. Une minute ou deux s'écoulent entre la fin de l'instance et le lancement d'une nouvelle instance. Pour plus d'informations

sur la manière de mettre fin à une instance, consultez la section Résiliation d'une instance dans le guide de l'utilisateur Amazon EC2.

Si vous mettez fin à des instances de votre groupe et que cela entraîne une répartition inégale entre les zones de disponibilité, Amazon EC2 Auto Scaling rééquilibre le groupe afin de rétablir une distribution uniforme, sauf si vous suspendez le processus. `AZRebalance` Pour plus d'informations, consultez Suspendre et reprendre les processus Amazon EC2 Auto Scaling.

Mise à l'échelle planifiée pour Amazon EC2 Auto Scaling

Avec le dimensionnement planifié, vous pouvez configurer le dimensionnement automatique de votre application en fonction des changements de charge prévisibles. Vous créez des actions planifiées qui augmentent ou diminuent la capacité souhaitée de votre groupe à des moments précis.

Par exemple, vous êtes confronté à un schéma de trafic hebdomadaire régulier dans lequel la charge augmente en milieu de semaine et diminue vers la fin de la semaine. Vous pouvez configurer un calendrier de dimensionnement dans Amazon EC2 Auto Scaling qui s'aligne sur le modèle suivant :

- Mercredi matin, une action planifiée augmente la capacité en augmentant la capacité souhaitée précédemment définie pour le groupe Auto Scaling.
- Vendredi soir, une autre action planifiée réduit la capacité en diminuant la capacité souhaitée précédemment définie pour le groupe Auto Scaling.

Ces actions de mise à l'échelle planifiées vous permettent d'optimiser les coûts et les performances. Votre application dispose d'une capacité suffisante pour gérer le pic de trafic en milieu de semaine, mais elle ne surfournit pas de capacité inutile à d'autres moments.

Vous pouvez utiliser conjointement le dimensionnement planifié et les politiques de dimensionnement pour tirer parti des avantages des deux approches de dimensionnement. Après l'exécution d'une action de mise à l'échelle planifiée, la stratégie de mise à l'échelle peut continuer à prendre des décisions sur l'opportunité de poursuivre la mise à l'échelle de la capacité. Cela vous permet de vous assurer que vous avez une capacité suffisante pour gérer la charge de votre application. Bien que votre application soit mise à l'échelle pour répondre à la demande, la capacité actuelle doit se situer dans les limites de la capacité minimale et maximale qui a été fixée par votre action planifiée.

Table des matières

Comment fonctionne la mise à l'échelle planifiée

Pour utiliser le dimensionnement planifié, créez des actions planifiées, qui indiquent à Amazon EC2 Auto Scaling d'effectuer des activités de dimensionnement à des moments précis. Lorsque vous créez une action planifiée, vous spécifiez le groupe Auto Scaling, le moment où l'activité de dimensionnement doit avoir lieu, la nouvelle capacité souhaitée, et éventuellement une nouvelle capacité minimale et une nouvelle capacité maximale. Vous pouvez créer des actions planifiées pour une mise à l’échelle unique ou selon une planification récurrente.

À l'heure spécifiée, Amazon EC2 Auto Scaling évolue en fonction des nouvelles valeurs de capacité, en comparant la capacité actuelle à la capacité souhaitée spécifiée.

- Si la capacité actuelle est inférieure à la capacité souhaitée spécifiée, Amazon EC2 Auto Scaling augmente ou ajoute des instances à la capacité souhaitée spécifiée.
- Si la capacité actuelle est supérieure à la capacité souhaitée spécifiée, Amazon EC2 Auto Scaling intègre ou supprime des instances jusqu'à la capacité souhaitée spécifiée.

Une action planifiée définit la capacité souhaitée, minimale et maximale du groupe à la date et à l'heure spécifiées. Vous pouvez créer une action planifiée pour une seule de ces capacités à la fois, par exemple la capacité souhaitée. Cependant, dans certains cas, vous devez inclure les capacités minimale et maximale pour vous assurer que la capacité souhaitée que vous avez spécifiée dans l'action ne dépasse pas ces limites.

Planifications récurrentes

Pour créer un calendrier récurrent à l'aide du SDK AWS CLI ou d'un SDK, spécifiez une expression cron et un fuseau horaire pour décrire le moment où cette action planifiée doit se reproduire. Vous

pouvez éventuellement spécifier une date et une heure pour l'heure de début, l'heure de fin, voire les deux.

Pour créer un calendrier récurrent à l'aide du AWS Management Console, spécifiez le schéma de récurrence, le fuseau horaire, l'heure de début et l'heure de fin facultative de votre action planifiée. Toutes les options de motif de récurrence sont basées sur des expressions cron. Vous pouvez également écrire votre propre expression cron personnalisée.

L'expression cron est constituée de cinq champs séparés par des espaces : [Minute] [Heure] [Jour_du_Mois] [Mois_de_Année] [Jour_de_Semaine]. Par exemple, l'expression cron `30 6 * * 2` configure une action planifiée qui se répète tous les mardis à 6h30. L'astérisque est utilisé comme caractère générique pour correspondre à toutes les valeurs d'un champ. Pour d'autres exemples d'expressions cron, consultez https://crontab.guru/examples.html. Pour plus d'informations sur l'écriture de vos propres expressions cron dans ce format, consultez Crontab.

Choisissez avec soin vos heures de début et de fin. Gardez à l'esprit les points suivants :

- Si vous spécifiez une heure de début, Amazon EC2 Auto Scaling exécute l'action à ce moment-là, puis exécute l'action basée sur la planification récurrente.
- Si vous spécifiez une heure de fin, l'action cesse de se répéter après cette heure. Une action planifiée n'est pas conservée dans votre compte une fois qu'elle a atteint son heure de fin.
- L'heure de début et l'heure de fin doivent être définies en UTC lorsque vous utilisez le AWS CLI ou un SDK.

Fuseau horaire

Par défaut, les planifications récurrentes que vous définissez sont exprimées en heure UTC (temps universel coordonné). Vous pouvez modifier le fuseau horaire afin qu'elle corresponde à votre fuseau horaire local ou à celui d'une autre partie de votre réseau. Lorsque vous spécifiez un fuseau horaire qui observe l'heure d'été (DST), l'action s'ajuste automatiquement pour l'heure d'été.

Les valeurs valides sont les noms canoniques des fuseaux horaires issus de la base de données des fuseaux horaires de l'Internet Assigned Numbers Authority (IANA). Par exemple, l'heure de l'Est des États-Unis est identifiée canoniquement comme`America/New_York`. Pour plus d'informations, consultez https://www.iana.org/time-zones.

Fuseaux horaires basés sur la localisation, tels que l'ajustement `America/New_York` automatique pour l'heure d'été. Cependant, un fuseau horaire basé sur UTC tel que `Etc/UTC` est une heure absolue qui ne s'ajuste pas à l'heure d'été.

Par exemple, vous avez une planification récurrente dont le fuseau horaire est `America/New_York`. La première action de mise à l'échelle se produit dans le fuseau horaire `America/New_York` avant le début de l'heure d'été. La prochaine action de mise à l'échelle se produit dans le fuseau horaire `America/New_York` après le démarrage de l'heure d'été. La première action commence à 8h00 UTC-5 en heure locale, tandis que la deuxième fois commence à 8h00 UTC-4 en heure locale.

Si vous créez une action planifiée à l'aide du AWS Management Console et que vous spécifiez un fuseau horaire respectant l'heure d'été, le calendrier récurrent ainsi que les heures de début et de fin s'ajustent automatiquement à l'heure d'été.

Considérations

Lorsque vous créez une action planifiée, gardez les éléments suivants à l'esprit.

- L'ordre d'exécution des actions planifiées est garanti dans le même groupe, mais pas entre les groupes.
- Une action planifiée s'exécute généralement en quelques secondes. Cependant, l'action peut être retardée de deux minutes au plus par rapport à l'heure de début planifiée. Dans la mesure où les actions planifiées d'un groupe Auto Scaling sont exécutées dans l'ordre dans lequel elles sont spécifiées, celles dont les heures de début planifiées sont trop proches les unes des autres peuvent prendre plus de temps à s'exécuter.
- Vous pouvez désactiver temporairement la mise à l'échelle planifiée pour un groupe Auto Scaling en suspendant le processus `ScheduledActions`. Cela vous permet d'empêcher les actions planifiées d'être actives sans avoir à les supprimer. Vous pouvez ensuite reprendre la mise à l'échelle planifiée lorsque vous souhaitez l'utiliser à nouveau. Pour plus d’informations, consultez Suspendre et reprendre les processus Amazon EC2 Auto Scaling.
- Après avoir créé une action planifiée, vous pouvez mettre à jour n'importe lequel de ses paramètres, à l'exception du nom.

Création d'une action planifiée

Pour créer une action planifiée pour votre groupe Auto Scaling, appliquez l'une des méthodes suivantes :

Console

Pour créer une action planifiée

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Sous l'onglet Mise à l'échelle automatique dans Actions planifiées, choisissez Créer une action planifiée.
4. Dans le champ Name (Nom), attribuez un nom à l'action planifiée.
5. Pour Capacité souhaitée, Min, Max, choisissez la nouvelle capacité souhaitée du groupe et les nouvelles limites de taille minimale et maximale. Cette capacité souhaitée doit être supérieure ou égale à la taille minimale du groupe et inférieure ou égale à la taille maximale du groupe.
6. Pour Recurrence (Récurrence), choisissez l'une des options disponibles.
   - Si vous souhaitez effectuer des mises à l'échelle selon un calendrier récurrent, choisissez la fréquence à laquelle Amazon EC2 Auto Scaling doit exécuter l'action planifiée.
     - Si vous sélectionnez une option qui commence par Every (Toutes les), l'expression Cron est créée pour vous.
     - Si vous sélectionnez Cron, tapez une expression Cron qui spécifie l'heure à laquelle exécuter l'action.
   - Si vous ne souhaitez mettre à l'échelle qu'une seule fois, choisissez Once (Une fois).
7. Dans le champ Time zone (Fuseau horaire), choisissez un fuseau horaire. L'argument par défaut est `Etc/UTC`.

   Tous les fuseaux horaires répertoriés proviennent de la base de données des fuseaux horaires IANA. Pour plus d'informations, consultez https://en.wikipedia.org/wiki/List_of_tz_database_time_zones.
8. Définissez une date et une heure pour Specific start time (Heure de début spécifique).
   - Si vous avez choisi une planification récurrente, l'heure de début définit le moment où la première action planifiée de la série récurrente s'exécute.

- Si vous avez choisi Once (Une fois) comme récurrence, l'heure de début définit la date et l'heure d'exécution de l'action de planification.

9. (Facultatif) Pour les programmes récurrents, vous pouvez spécifier une heure de fin en sélectionnant Définir l'heure de fin puis en choisissant une date et une heure pour End by (Fin le).
10. Choisissez Créer. La console affiche les actions planifiées pour le groupe Auto Scaling.

AWS CLI

Pour créer une action planifiée, vous pouvez utiliser l'un des exemples de commandes suivants. Remplacez chaque *espace réservé à la saisie de l'utilisateur* par vos propres informations.

Exemple : Pour planifier un dimensionnement unique

Utilisez la commande put-scheduled-update-group-action suivante avec les options et. `--start-time "YYYY-MM-DDThh:mm:ssZ"` `--desired-capacity`

```
aws autoscaling put-scheduled-update-group-action --scheduled-action-name my-one-time-action \
  --auto-scaling-group-name my-asg --start-time "2021-03-31T08:00:00Z" --desired-capacity 3
```

Exemple : pour planifier le dimensionnement selon un calendrier récurrent

Utilisez la commande put-scheduled-update-group-action suivante avec les options et. `--recurrence "cron expression"` `--desired-capacity`

```
aws autoscaling put-scheduled-update-group-action --scheduled-action-name my-recurring-action \
  --auto-scaling-group-name my-asg --recurrence "0 9 * * *" --desired-capacity 3
```

Par défaut, Amazon EC2 Auto Scaling exécute le calendrier de récurrence spécifié en fonction du fuseau horaire UTC. Pour spécifier un autre fuseau horaire, incluez l'`--time-zone`option et le nom du fuseau horaire IANA, comme dans l'exemple suivant.

```
--time-zone "America/New_York"
```

Pour plus d'informations, consultez https://en.wikipedia.org/wiki/List_of_tz_database_time_zones.

Afficher les détails des actions planifiées

Pour consulter le détail des prochaines actions planifiées pour votre groupe Auto Scaling, appliquez l'une des méthodes suivantes :

Console

Pour afficher les détails des actions planifiées

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Sélectionnez votre groupe Auto Scaling.
3. Dans l'onglet Mise à l'échelle automatique, dans la section Actions planifiées, vous pouvez consulter les actions planifiées à venir.

Notez que la console affiche les valeurs de l'heure de début et de fin en heure locale avec le décalage UTC en vigueur à la date et à l'heure spécifiées. Le décalage UTC est la différence, en heures et en minutes, entre l'heure locale et l'heure UTC. La valeur de fuseau horaire affiche le fuseau horaire demandé, par exemple,`America/New_York`.

AWS CLI

Utilisez la commande describe-scheduled-actions suivante.

```
aws autoscaling describe-scheduled-actions --auto-scaling-group-name my-asg
```

Si elle aboutit, cette commande renvoie un résultat similaire à ce qui suit.

```
{
  "ScheduledUpdateGroupActions": [
    {
      "AutoScalingGroupName": "my-asg",
      "ScheduledActionName": "my-recurring-action",
      "Recurrence": "30 0 1 1,6,12 *",
      "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledUpdateGroupAction:8e86b655-b2e6-4410-8f29-
b4f094d6871c:autoScalingGroupName/my-asg:scheduledActionName/my-recurring-action",
      "StartTime": "2020-12-01T00:30:00Z",
      "Time": "2020-12-01T00:30:00Z",
      "MinSize": 1,
```

```
      "MaxSize": 6,
      "DesiredCapacity": 4
    }
  ]
}
```

Vérifier les activités de mise à l'échelle

Pour vérifier les activités de mise à l'échelle associées à la mise à l'échelle planifiée, voir Vérifier une activité de mise à l'échelle pour un groupe Auto Scaling.

Supprimer une action planifiée

Pour supprimer une action planifiée, appliquez l'une des méthodes suivantes :

Console

Pour supprimer une action planifiée

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Sélectionnez votre groupe Auto Scaling.
3. Sous l'onglet Scalabilité automatique dans Actions planifiées, sélectionnez une action planifiée.
4. Choisissez Actions, Delete (Supprimer).
5. Lorsque vous êtes invité à confirmer l'opération, choisissez Supprimer.

AWS CLI

Utilisez la commande delete-scheduled-action suivante.

```
aws autoscaling delete-scheduled-action --auto-scaling-group-name my-asg \
  --scheduled-action-name my-recurring-action
```

Limites

- Les noms des actions planifiées doivent être uniques pour chaque groupe Auto Scaling.

- Une action planifiée doit posséder une valeur de temps unique. Si vous tentez de planifier une activité à un moment où une autre activité de mise à l'échelle est déjà planifiée, l'appel est rejeté et renvoie une erreur indiquant qu'une action planifiée avec cette heure de début planifiée existe déjà.
- Vous pouvez créer 125 actions planifiées maximum par groupe Auto Scaling.

Mise à l'échelle dynamique pour Amazon EC2 Auto Scaling

La mise à l'échelle dynamique modifie la capacité de votre groupe Auto Scaling à mesure que le trafic change.

Amazon EC2 Auto Scaling prend en charge les types de politique de mise à l'échelle dynamique suivants :

- Dimensionnement du suivi des cibles : augmentez ou diminuez la capacité actuelle du groupe en fonction d'une CloudWatch métrique Amazon et d'une valeur cible. Cette option opère sous le même principe que votre thermostat qui maintient la température de votre domicile—vous sélectionnez une température et le thermostat se charge du reste.
- La mise à l'échelle par étapes – Augmente ou réduit la capacité actuelle du groupe en fonction d'un ensemble d'ajustements de mise à l'échelle appelés ajustements d'étape. Ceux-ci varient en fonction de la valeur du seuil de l'alarme.
- Mise à l'échelle simple – Augmente ou réduit la capacité actuelle du groupe en fonction d'un seul ajustement de la mise à l'échelle, avec un temps de stabilisation entre chaque opération de la mise à l'échelle.

Nous vous recommandons vivement d'utiliser des politiques de dimensionnement pour le suivi des cibles et de choisir une métrique qui change de manière inversement proportionnelle à une modification de la capacité de votre groupe Auto Scaling. Ainsi, si vous doublez la taille de votre groupe Auto Scaling, la métrique diminue de 50 %. Cela permet aux données métriques de déclencher avec précision des événements de mise à l’échelle proportionnelle. Des indicateurs tels que l'utilisation moyenne du processeur ou le nombre moyen de demandes par cible sont inclus.

Grâce au suivi des cibles, votre groupe Auto Scaling évolue en proportion directe de la charge réelle de votre application. Cela signifie qu'en plus de répondre au besoin immédiat de capacité pour résoudre le problème de changements de charge, la politique de suivi des cibles peut également s'adapter aux changements de charge qui se produisent au fil du temps, en raison de variations saisonnières, par exemple.

Les politiques de suivi des cibles éliminent également le besoin de définir manuellement les CloudWatch alarmes et les ajustements de dimensionnement. Amazon EC2 Auto Scaling gère cela automatiquement en fonction de l'objectif que vous avez défini.

Table des matières

Fonctionnement des politiques de mise à l'échelle

Une politique de dimensionnement dynamique indique à Amazon EC2 Auto Scaling de suivre une métrique CloudWatch spécifique et définit les mesures à prendre lorsque l'alarme CloudWatch associée est dans ALARM. Les métriques utilisées pour appeler l'état de l'alarme sont une agrégation de métriques provenant de toutes les instances du groupe Auto Scaling. (Par exemple, supposons que vous ayez un groupe Auto Scaling avec deux instances où une instance est à 60 % de CPU et l'autre à 40 % de CPU. En moyenne, ils sont à 50 % CPU.) Lorsque la politique est en vigueur, Amazon EC2 Auto Scaling ajuste la capacité souhaitée du groupe vers le haut ou vers le bas lorsque le seuil d'une alarme est dépassé.

Lorsqu'une politique de mise à l'échelle dynamique est appelée, si le calcul de capacité produit un nombre en dehors de la plage de tailles minimale et maximale du groupe, Amazon EC2 Auto Scaling garantit que la nouvelle capacité ne dépassera jamais les limites de taille minimale et maximale. La capacité est mesurée de deux manières : en utilisant les mêmes unités que celles que vous avez choisies lorsque vous avez défini la capacité souhaitée en termes d'instances, ou en utilisant des unités de capacité (si des poids d'instance sont appliqués).

- Exemple 1 : un groupe Auto Scaling a une capacité maximale de 3, une capacité actuelle de 2 et une politique de mise à l'échelle dynamique qui ajoute 3 instances. Lors de l'appel de cette politique, Amazon EC2 Auto Scaling ajoute 1 seule instance au groupe pour empêcher le groupe de dépasser sa taille maximale.
- Exemple 2 : un groupe Auto Scaling a une capacité minimale de 2, une capacité actuelle de 3 et une politique de mise à l'échelle dynamique qui supprime 2 instances. Lors de l'appel de cette politique, Amazon EC2 Auto Scaling ne supprime que 1 seule instance du groupe pour éviter que le groupe ne devienne inférieure à sa taille minimale.

Lorsque la capacité désirée atteint la limite de taille maximale, la mise à l'échelle s'arrête. Si la demande chute et que la capacité actuelle diminue, Amazon EC2 Auto Scaling peut réaliser à nouveau une montée en puissance.

L'exception est lorsque vous utilisez des poids d'instance. Dans ce cas, Amazon EC2 Auto Scaling peut monter en puissance au-delà de la limite de taille maximale, mais uniquement jusqu'à votre poids d'instance maximal. Son intention est de se rapprocher le plus possible de la nouvelle capacité souhaitée tout en respectant les politiques d'allocation qui sont spécifiées pour le groupe. Les politiques d'allocation déterminent les types d'instance à lancer. Les pondérations déterminent le nombre d'unités de capacité avec lequel chaque instance contribue à la capacité du groupe souhaitée selon son type d'instance.

- Exemple 3 : un groupe Auto Scaling a une capacité maximale de 12, une capacité actuelle de 10 et une politique de mise à l'échelle dynamique qui ajoute 5 unités de capacité. Les types d'instance ont l'une des trois pondérations suivantes : 1, 4 ou 6. Lors de l'appel de la politique, Amazon EC2 Auto Scaling choisit de lancer un type d'instance avec une pondération de 6 en fonction de la politique d'allocation. Le résultat de cet événement évolutif est un groupe avec une capacité désirée de 12 et une capacité actuelle de 16.

Plusieurs politiques de mise à l'échelle dynamique

Dans la plupart des cas, une politique de suivi des cibles et de mise à l’échelle est suffisante pour configurer automatiquement l’évolutivité horizontale ou la mise à l’échelle horizontale de votre groupe Auto Scaling. Une politique de suivi des cibles et de mise à l’échelle vous permet de sélectionner un résultat souhaité et que le groupe Auto Scaling ajoute ou supprime des instances en fonction des besoins pour obtenir ce résultat.

Pour une mise à l'échelle avancée de configuration, votre groupe Auto Scaling peut disposer de plus d'une politique de mise à l'échelle. Par exemple, vous pouvez définir une ou plusieurs politiques de suivi des cibles et de mise à l'échelle, une ou plusieurs politiques de mise à l'échelle par étapes, ou les deux. Cela permet une plus grande flexibilité pour couvrir plusieurs scénarios.

Pour illustrer le fonctionnement conjoint de plusieurs politiques de mise à l'échelle dynamique, prenez une application qui utilise un groupe Auto Scaling et une file d'attente Amazon SQS pour envoyer les demandes à une seule instance EC2. Pour veiller à ce que l'application fonctionne à un niveau optimal, deux politiques contrôlent le moment où le groupe Auto Scaling doit monter en puissance. La première est une politique de suivi des cibles et de mise à l'échelle qui utilise une métrique personnalisée pour ajouter et supprimer des capacités en fonction du nombre de messages SQS dans la file d'attente. L'autre est une politique de dimensionnement par étapes qui utilise la CloudWatch `CPUUtilization` métrique Amazon pour ajouter de la capacité lorsque l'instance dépasse 90 % d'utilisation pendant une durée spécifiée.

Lorsqu'il existe plusieurs politiques en vigueur en même temps, il est possible que chaque politique puisse demander au groupe Auto Scaling de se mettre à l'échelle (augmentation ou diminution) simultanément. Par exemple, il est possible que la `CPUUtilization` métrique augmente et dépasse le seuil de l' CloudWatch alarme en même temps que la métrique personnalisée SQS augmente et dépasse le seuil de l'alarme métrique personnalisée.

Lorsque ces situations se produisent, Amazon EC2 Auto Scaling choisit la politique qui fournit la plus grande capacité à la fois d'évolutivité horizontale et de mise à l'échelle horizontale. Par exemple, supposons que la politique de `CPUUtilization` lance une instance, tandis que la politique pour la file d'attente SQS en lance deux. Si le critère d'évolutivité horizontale pour les deux politiques est respecté, Amazon EC2 Auto Scaling donne la priorité à la politique de file d'attente SQS. Le groupe Auto Scaling lance donc deux instances.

L'approche qui consiste à donner la priorité à la politique qui fournit la plus grande capacité s'applique même lorsque les politiques utilisent différents critères pour la mise à l'échelle horizontale. Par exemple, si une politique suspend trois instances, qu'une autre politique diminue le nombre d'instances de 25 %, et que le groupe dispose de huit instances au moment de la mise à l'échelle horizontale, Amazon EC2 Auto Scaling donne la priorité à la politique qui fournit le plus grand nombre d'instances au groupe. Il s'ensuit que le groupe Auto Scaling résilie deux instances (25 % de 8 = 2). L'objectif est d'empêcher Amazon EC2 Auto Scaling de supprimer un trop grand nombre d'instances.

Toutefois, nous vous recommandons d'être prudent lorsque vous utilisez des politiques de suivi des objectifs et d'échelonnement avec des politiques de mise à l'échelle par étapes, car les conflits entre ces politiques peuvent entraîner un comportement indésirable. Par exemple, si la politique de

mise à l'échelle par étapes lance une activité de mise à l'échelle horizontale avant que la politique de suivi des objectifs et d'échelonnement ne soit prête pour la mise à l'échelle horizontale, l'activité de mise à l'échelle horizontale ne sera pas bloquée. Une fois l'activité de mise à l'échelle horizontale terminée, la politique de suivi des objectifs et d'échelonnement peut demander au groupe d'effectuer une évolutivité horizontale.

Politiques de suivi des objectifs et d'échelonnement pour Amazon EC2 Auto Scaling

Une politique de dimensionnement pour le suivi des cibles permet d'ajuster automatiquement la capacité de votre groupe Auto Scaling en fonction d'une valeur métrique cible. Cela permet à votre application de maintenir des performances et une rentabilité optimales sans intervention manuelle.

Avec le suivi des cibles, vous sélectionnez une métrique et une valeur cible pour représenter le niveau d'utilisation ou de débit moyen idéal pour votre application. Amazon EC2 Auto Scaling crée et gère les CloudWatch alarmes qui déclenchent des événements de dimensionnement lorsque la métrique s'écarte de la cible. Par exemple, cela est similaire à la façon dont un thermostat maintient une température cible.

Supposons par exemple que vous avez une application Web qui s'exécute actuellement sur deux instances et vous souhaitez que l'utilisation de l'UC du groupe Auto Scaling reste à environ 50 % lorsque la charge sur l'application change. Vous disposez ainsi d'une plus grande capacité pour gérer les pics de trafic sans avoir à maintenir une quantité excessive des ressources inutilisées.

Vous pouvez répondre à ce besoin en créant une stratégie de suivi des objectifs et d'échelonnement qui cible une utilisation moyenne du CPU de 50 pour cent. Ensuite, votre groupe Auto Scaling augmentera, ou augmentera sa capacité, lorsque le processeur dépasse 50 % pour faire face à une charge accrue. Il augmentera ou diminuera la capacité lorsque le processeur tombe en dessous de 50 % afin d'optimiser les coûts pendant les périodes de faible utilisation.

Rubriques

- Créer une politique de mise à l'échelle du suivi des cibles pour Amazon EC2 Auto Scaling à l'aide d'une expression mathématique appliquée à une métrique

Politiques multiple de suivi des objectifs de la mise à l'échelle

Pour vous permettre d'optimiser la performance de mise en échelle, vous pouvez disposer de plusieurs politiques de suivi des objectifs de la mise à l'échelle à la fois, dans la mesure où chacune d'entre elles utilise une métrique différente. Par exemple, l'utilisation et le débit peuvent s'influencer mutuellement. Chaque fois que l'une de ces métriques change, cela implique généralement que d'autres métriques seront également affectées. L'utilisation de plusieurs métriques fournit donc des informations supplémentaires sur la charge que subit votre groupe Auto Scaling. Cela peut aider Amazon EC2 Auto Scaling à prendre des décisions plus éclairées lorsqu'il s'agit de déterminer la capacité à ajouter à votre groupe.

L'objectif d'Amazon EC2 Auto Scaling est de toujours donner la priorité à la disponibilité. Il élargira le groupe Auto Scaling si l'une des politiques de suivi des cibles est prête à être étendue. Il ne sera étendu que si toutes les politiques de suivi des cibles (avec la partie scale-in activée) sont prêtes à être étendues.

Choisissez métriques

Vous pouvez créer des stratégies de suivi des objectifs de la mise à l'échelle avec des métriques prédéfinies ou des métriques personnalisées.

Lorsque vous créez une stratégie de suivi des objectifs de la mise à l'échelle avec une métrique prédéfinie, vous choisissez une métrique dans la liste suivante de métriques prédéfinies :

- `ASGAverageCPUUtilization` - Utilisation moyenne de l'UC du groupe Auto Scaling.
- `ASGAverageNetworkIn` - Nombre moyen d'octets reçus par une seule instance sur toutes les interfaces réseau.
- `ASGAverageNetworkOut` - Nombre moyen d'octets envoyé depuis une seule instance sur toutes les interfaces réseau.
- `ALBRequestCountPerTarget` - Nombre de demandes Application Load Balancer par cible.

> **Important**
>
> Vous trouverez d'autres informations utiles sur les mesures relatives à l'utilisation du processeur, aux E/S réseau et au nombre de demandes d'Application Load Balancer par

cible dans la rubrique Lister les métriques CloudWatch disponibles pour vos instances dans le guide de l'utilisateur Amazon EC2 et dans la rubrique relative aux métriques relatives à votre Application Load Balancer dans CloudWatch le guide de l'utilisateur pour les équilibreurs de charge d'application, respectivement.

Vous pouvez choisir d'autres CloudWatch mesures disponibles ou les vôtres en CloudWatch spécifiant une métrique personnalisée. Vous devez utiliser le AWS CLI ou un SDK pour créer une politique de suivi cible avec une spécification de métrique personnalisée. Pour un exemple qui spécifie une spécification de métrique personnalisée pour une politique de dimensionnement du suivi des cibles à l'aide du AWS CLI, voirExemple de politiques de mise à l'échelle pour AWS Command Line Interface (AWS CLI).

Gardez les points suivants à l'esprit lorsque vous choisissez une métrique :

- Nous vous recommandons d'utiliser uniquement des mesures disponibles à des intervalles d'une minute pour vous aider à évoluer plus rapidement en fonction des changements d'utilisation. Le suivi des cibles évaluera les métriques agrégées avec une granularité d'une minute pour toutes les métriques prédéfinies et les métriques personnalisées, mais la métrique sous-jacente peut publier des données moins fréquemment. Par exemple, toutes les métriques Amazon EC2 sont envoyées toutes les cinq minutes par défaut, mais elles sont configurables à une minute (ce que l'on appelle la surveillance détaillée). Ce choix appartient à chaque service. La plupart essaient d'utiliser le plus petit intervalle possible. Pour obtenir des informations sur l'activation de la surveillance détaillée, consultez Configurer la surveillance pour les instances à scalabilité automatique.
- Toutes les métriques personnalisées ne fonctionnent pas pour le suivi des cibles. La métrique doit être une métrique d'utilisation valide et décrire le degré d'occupation d'une instance. La valeur de métrique doit augmenter ou diminuer en proportion du nombre d'instances présentes dans le groupe Auto Scaling. C'est pour que les données de la métrique puissent être utilisées afin d'augmenter ou réduire proportionnellement le nombre d'instances. Par exemple, l'utilisation de l'UC d'un groupe Auto Scaling (c'est-à-dire la métrique `CPUUtilization`Amazon EC2 avec la dimension de métrique `AutoScalingGroupName`) fonctionne, si la charge du groupe Auto Scaling est répartie entre les instances.
- Les métriques suivantes ne fonctionnent pas pour le suivi de la cible :
  - Nombre de demandes reçues par l'équilibreur de charge en amont du groupe Auto Scaling (autrement dit, la métrique Elastic Load Balancing `RequestCount`). Le nombre de demandes reçues par l'équilibreur de charge ne change pas en fonction de l'utilisation du groupe Auto Scaling.

  - Latence des demandes de l'équilibreur de charge (autrement dit, la métrique Elastic Load Balancing `Latency`). La latence des demandes peut augmenter en fonction de la croissance de l'utilisation, sans toutefois évoluer proportionnellement à cette croissance.
  - La métrique de file d'attente CloudWatch Amazon SQS. `ApproximateNumberOfMessagesVisible` Le nombre de messages d'une file d'attente peut ne pas changer proportionnellement à la taille du groupe Auto Scaling qui traite les messages de la file d'attente. Toutefois, une métrique personnalisée qui mesure le nombre de messages dans la file d'attente par instance EC2 dans le groupe Auto Scaling peut fonctionner. Pour plus d'informations, consultez Mise à l'échelle basée sur Amazon SQS.
- Pour utiliser la métrique `ALBRequestCountPerTarget`, vous devez spécifier le paramètre `ResourceLabel` permettant d'identifier le groupe cible de l'équilibreur de charge associé à la métrique. Pour un exemple qui spécifie le `ResourceLabel` paramètre d'une politique de dimensionnement du suivi des cibles à l'aide du AWS CLI, voirExemple de politiques de mise à l'échelle pour AWS Command Line Interface (AWS CLI).
- Lorsqu'une métrique émet des valeurs réelles de 0 à CloudWatch (par exemple,`ALBRequestCountPerTarget`), un groupe Auto Scaling peut passer à 0 s'il n'y a aucun trafic vers votre application pendant une période prolongée. Pour que votre groupe Auto Scaling diminue à 0 lorsque aucune demande n'y est acheminée, la capacité minimale du groupe doit être définie sur 0.
- Au lieu de publier de nouvelles métriques à utiliser dans votre politique de mise à l'échelle, vous pouvez utiliser les calculs de métriques pour combiner des métriques existantes. Pour plus d'informations, consultez Créer une politique de mise à l'échelle du suivi des cibles pour Amazon EC2 Auto Scaling à l'aide d'une expression mathématique appliquée à une métrique.

Définition de la valeur cible

Lorsque vous créez une politique de suivi de la cible, vous devez spécifier une valeur cible. La valeur cible représente l'utilisation ou le débit moyen optimal pour le groupe Auto Scaling. Afin d'utiliser les ressources de manière efficiente, définissez une valeur cible aussi élevée que possible avec un tampon raisonnable en cas d'augmentation inattendue du trafic. Lorsque votre application est mise à l'échelle de manière optimale pour un flux de trafic normal, la valeur de métrique réelle doit être égale ou sensiblement inférieure à la valeur cible.

Lorsqu'une stratégie de mise à l'échelle est basée sur le débit, tel que le nombre de demandes par cible pour un Application Load Balancer, les I/O réseau ou d'autres métriques de nombre, la

valeur cible représente le débit moyen optimal depuis une seule instance, pendant une période d'une minute.

Définir le temps de préchauffage de l'instance

Vous pouvez facultativement spécifier le nombre de secondes nécessaires pour la préparation d'une instance nouvellement lancée. Tant que le temps de préchauffage spécifié n'est pas expiré, une instance n'est pas prise en compte dans les métriques d'instance EC2 agrégées du groupe Auto Scaling.

Lorsque les instances sont en période de préchauffage, vos politiques de dimensionnement ne sont redimensionnées que si la valeur métrique des instances qui ne sont pas en phase de préchauffage est supérieure à l'utilisation cible de la politique.

Si le groupe est à nouveau monté en puissance, les instances qui sont toujours en cours de préparation sont comptées dans le cadre de la capacité souhaitée pour la prochaine activité de montée en puissance. L'objectif est d'effectuer une montée en puissance continue (mais pas excessive).

Pendant que l'activité de mise à l'échelle est en cours, toutes les activités de mise à l'échelle initiées par les politiques de mise à l'échelle sont bloquées jusqu'à ce que les instances aient fini leur préparation. Lorsque les instances ont terminé la préparation, si un événement de mise à l'échelle horizontale se produit, toutes les instances en cours de résiliation seront prises en compte dans la capacité actuelle du groupe lors du calcul de la nouvelle capacité souhaitée. Par conséquent, nous n'enlevons pas plus d'instances du groupe Auto Scaling que nécessaire.

Valeur par défaut

Si aucune valeur n'est définie, la politique de dimensionnement utilisera la valeur par défaut, qui est la valeur du préchauffage d'instance par défaut défini pour le groupe. Si le préchauffage de l'instance par défaut est nul, il revient à la valeur du temps de recharge par défaut. Nous vous recommandons d'utiliser le préchauffage de l'instance par défaut pour faciliter la mise à jour de toutes les politiques de dimensionnement lorsque le temps de préchauffage change.

Considérations

Les points suivants s'appliquent lors de l'utilisation des politiques de suivi des objectifs et d'échelonnement

- Ne créez pas, ne modifiez ni ne supprimez les CloudWatch alarmes utilisées avec une politique de dimensionnement du suivi des cibles. Amazon EC2 Auto Scaling crée et gère les CloudWatch

alarmes associées à vos politiques de dimensionnement du suivi des cibles et les supprime lorsqu'elles ne sont plus nécessaires.

- Une politique de mise à l'échelle du suivi cible priorise la disponibilité pendant les périodes de fluctuation des niveaux de trafic en augmentant plus progressivement lorsque le trafic diminue. Si vous souhaitez que votre groupe Auto Scaling soit mis à l'échelle immédiatement à la fin d'une charge de travail, vous pouvez désactiver la fonction diminuer de la politique. Cela vous donne la latitude d'utiliser la méthode de mise à l'échelle qui répond le mieux à vos besoins lorsque l'utilisation est faible. Pour garantir que la mise à l'échelle se fasse le plus rapidement possible, nous vous recommandons de ne pas utiliser de politique de mise à l'échelle simple pour empêcher l'ajout d'une période de recharge.
- S'il manque des points de données à la métrique, l'état de l' CloudWatch alarme passe à`INSUFFICIENT_DATA`. Dans ce cas, Amazon EC2 Auto Scaling ne peut pas mettre à l'échelle votre groupe tant que de nouveaux points de données ne sont pas trouvés.
- Si la métrique est rarement rapportée, les calculs de métriques peuvent s'avérer utiles. Par exemple, pour utiliser les valeurs les plus récentes, utilisez la fonction `FILL(m1,REPEAT)` là où `m1` est la métrique.
- Vous pouvez constater des écarts entre la valeur cible et les points de données de métrique réels. Ceci est dû au fait que nous agissons toujours avec prudence en déterminant un arrondi vers le haut ou vers le bas quand il détermine le nombre d'instances à ajouter ou enlever. Cela empêche nous empêche d'ajouter un nombre insuffisant d'instances ou de supprimer trop d'instances. Cependant, pour les groupes Auto Scaling de petite taille avec moins d'instances, l'utilisation du groupe peut sembler éloignée de la valeur cible. Définissez, par exemple, une valeur cible de 50 % pour l'utilisation de l'UC et le groupe Auto Scaling dépasse alors la cible. Nous pouvons déterminer que l'ajout de 1,5 instance diminuera l'utilisation de l'UC d'environ 50 %. Comme il n'est pas possible d'ajouter 1,5 instance, nous arrondissons à la valeur supérieure et ajoutons deux instances. Cela peut diminuer l'utilisation de la CPU à une valeur inférieure à 50 % mais cela garantit que votre application dispose de suffisamment de ressources pour le prendre en charge. De même, si nous déterminons que la suppression de 1,5 instance diminue l'utilisation de la CPU à une valeur supérieure à 50 %, nous ne retirons qu'une seule instance.

  Pour les groupes Auto Scaling ayant plus d'instances, l'utilisation est répartie sur un plus grand nombre d'instances, auquel cas l'ajout ou la suppression d'instances entraîne moins d'écarts entre la valeur cible et les points de données de métrique réels.
- Une politique de suivi des objectifs et d'échelonnement suppose qu'elle doit effectuer une montée en puissance de votre groupe Auto Scaling lorsque la métrique spécifiée est au-dessus de la valeur cible. Vous ne pouvez pas utiliser une politique de suivi des objectifs et d'échelonnement pour

effectuer une montée en puissance de votre groupe Auto Scaling lorsque la métrique spécifiée est en dessous de la valeur cible.

Création d'une politique de suivi des cibles et d'échelonnement

Pour créer une politique de dimensionnement du suivi des cibles pour votre groupe Auto Scaling, appliquez l'une des méthodes suivantes.

Avant de commencer, vérifiez que votre métrique préférée est disponible à intervalles d’une minute (par rapport à l’intervalle de 5 minutes par défaut pour les métriques Amazon EC2).

Console

Pour créer une politique de suivi des cibles et de mise à l’échelle pour un groupe Auto Scaling existant

1. Ouvrez la console Amazon EC2 à l’adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Choisissez Créer un groupe Auto Scaling.
3. Dans les étapes 1, 2 et 3, choisissez les options souhaitées et passez à l'Étape 4 : configurer la taille du groupe et des politiques de mise à l'échelle.
4. Sous Taille du groupe, spécifiez la plage entre laquelle vous souhaitez mettre à l’échelle en mettant à jour la capacité minimale et la capacité maximale. Ces deux paramètres permettent à votre groupe Auto Scaling d'effectuer une mise à l'échelle dynamique. Pour plus d’informations, consultez Définissez des limites de mise à l’échelle pour votre groupe Auto Scaling.
5. Sous Mise à l’échelle automatique, choisissez Politique de suivi des cibles et de mise à l’échelle.
6. Pour définir une politique, procédez comme suit :

   a. Attribuez un nom à la politique.

   b. Pour Metric type (Type de métrique), choisissez une métrique.

      Si vous avez choisi Application Load Balancer request count per target (Nombre de demandes d'Application Load Balancer par cible), choisissez un groupe cible dans Target group (Groupe cible).

   c. Spécifiez une Valeur cible pour la métrique.

d. (Facultatif) Pour le préchauffage de l'instance, mettez à jour la valeur de préchauffage de l'instance selon les besoins.

e. (Facultatif) Sélectionnez Désactiver la mise à l'échelle horizontale pour créer uniquement une politique de montée en puissance. Cela vous permet de créer une politique de mise à l'échelle horizontale distincte avec un type différent si vous le souhaitez.

7. Procédez à la création du groupe Auto Scaling. Votre politique de mise à l'échelle sera créée après la création du groupe Auto Scaling.

Pour créer une politique de suivi des objectifs et d'échelonnement pour un groupe Auto Scaling existant

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Vérifiez que les limites de mise à l'échelle sont correctement définies. Par exemple, si le groupe est déjà au maximum de sa taille, vous devez spécifier un nouveau maximum pour monter en puissance. Pour plus d'informations, consultez Définissez des limites de mise à l'échelle pour votre groupe Auto Scaling.
4. Sous l'onglet Scalabilité automatique, dans Politiques de mise à l'échelle dynamique, choisissez Créer une politique de mise à l'échelle dynamique.
5. Pour définir une politique, procédez comme suit :

   a. Pour le Type de politique, conservez la valeur par défaut de Suivi des cibles et de mise à l'échelle.

   b. Attribuez un nom à la politique.

   c. Pour Metric type (Type de métrique), choisissez une métrique. Vous ne pouvez choisir qu'un seul type de métrique. Pour utiliser plusieurs métriques, créez différentes politiques.

      Si vous avez choisi Application Load Balancer request count per target (Nombre de demandes d'Application Load Balancer par cible), choisissez un groupe cible dans Target group (Groupe cible).

   d. Spécifiez une Valeur cible pour la métrique.

e. (Facultatif) Pour le préchauffage de l'instance, mettez à jour la valeur de préchauffage de l'instance selon les besoins.

f. (Facultatif) Sélectionnez Désactiver la mise à l'échelle horizontale pour créer uniquement une politique de montée en puissance. Cela vous permet de créer une politique de mise à l'échelle horizontale distincte avec un type différent si vous le souhaitez.

6. Choisissez Créer.

AWS CLI

Pour créer une politique de dimensionnement du suivi des cibles, vous pouvez utiliser l'exemple suivant pour vous aider à démarrer. Remplacez chaque *espace réservé à la saisie de l'utilisateur* par vos propres informations.

> **Note**
>
> Pour obtenir plus d'exemples, consultez Exemple de politiques de mise à l'échelle pour AWS Command Line Interface (AWS CLI).

Pour créer une politique de suivi des cibles et de mise à l'échelle (AWS CLI)

1. Utilisez la `cat` commande suivante pour stocker une valeur cible pour votre politique de dimensionnement et une spécification de métrique prédéfinie dans un fichier JSON nommé `config.json` dans votre répertoire de base. Voici un exemple de configuration de suivi des cibles qui maintient l'utilisation moyenne du processeur à 50 %.

```
$ cat ~/config.json
{
  "TargetValue": 50.0,
  "PredefinedMetricSpecification":
    {
      "PredefinedMetricType": "ASGAverageCPUUtilization"
    }
}
```

Pour plus d'informations, consultez la section PredefinedMetricSpécification dans le manuel Amazon EC2 Auto Scaling API Reference.

2. Utilisez la commande put-scaling-policy ainsi que le fichier `config.json` créé à l'étape précédente pour élaborer la politique de mise à l'échelle.

```
aws autoscaling put-scaling-policy --policy-name cpu50-target-tracking-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
```

En cas de succès, cette commande renvoie les ARN et les noms des deux CloudWatch alarmes créées en votre nom.

```
{
    "PolicyARN": "arn:aws:autoscaling:us-
west-2:123456789012:scalingPolicy:228f02c2-c665-4bfd-
aaac-8b04080bea3c:autoScalingGroupName/my-asg:policyName/cpu50-target-tracking-
scaling-policy",
    "Alarms": [
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e",
            "AlarmName": "TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e"
        },
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2",
            "AlarmName": "TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2"
        }
    ]
}
```

Créer une politique de mise à l'échelle du suivi des cibles pour Amazon EC2 Auto Scaling à l'aide d'une expression mathématique appliquée à une métrique

À l'aide des mathématiques métriques, vous pouvez interroger plusieurs CloudWatch métriques et utiliser des expressions mathématiques pour créer de nouvelles séries chronologiques basées sur ces métriques. Vous pouvez visualiser les séries chronologiques obtenues dans la CloudWatch

console et les ajouter aux tableaux de bord. Pour plus d'informations sur les mathématiques métriques, consultez la section Utilisation des mathématiques métriques dans le guide de CloudWatch l'utilisateur Amazon.

Les considérations suivantes s'appliquent aux expressions mathématiques appliquées aux métriques :

- Vous pouvez interroger n'importe quelle CloudWatch métrique disponible. Chaque métrique est une combinaison unique du nom de la métrique, de l'espace de noms et de zéro dimension ou plus.
- Vous pouvez utiliser n'importe quel opérateur arithmétique (+ - */^), fonction statistique (telle que AVG ou SUM) ou toute autre fonction compatible. CloudWatch
- Vous pouvez utiliser à la fois des métriques et les résultats d'autres expressions mathématiques dans les formules de l'expression mathématique.
- Toutes les expressions utilisées dans une spécification de métrique doivent finalement retourner une seule séries temporelles.
- Vous pouvez vérifier la validité d'une expression mathématique métrique à l'aide de la CloudWatch console ou de l'API de CloudWatch GetMetricdonnées.

> **Note**
>
> Vous pouvez créer une politique de dimensionnement du suivi des cibles à l'aide de mathématiques métriques uniquement si vous utilisez le AWS CLI AWS CloudFormation, ou un SDK. Cette fonctionnalité n’est pas disponible depuis la console.

Exemple : file Amazon SQS des éléments en attente par instance

Pour calculer la file Amazon SQS des éléments en attente par instance, prenez le nombre approximatif de messages disponibles à la récupération dans la file d'attente et divisez ce nombre par la capacité d'exécution du groupe Auto Scaling, ce qui correspond au nombre d'instances dans l'état `InService`. Pour plus d’informations, consultez Mise à l'échelle basée sur Amazon SQS.

La logique de l'expression est la suivante :

```
sum of (number of messages in the queue)/(number of InService instances)
```

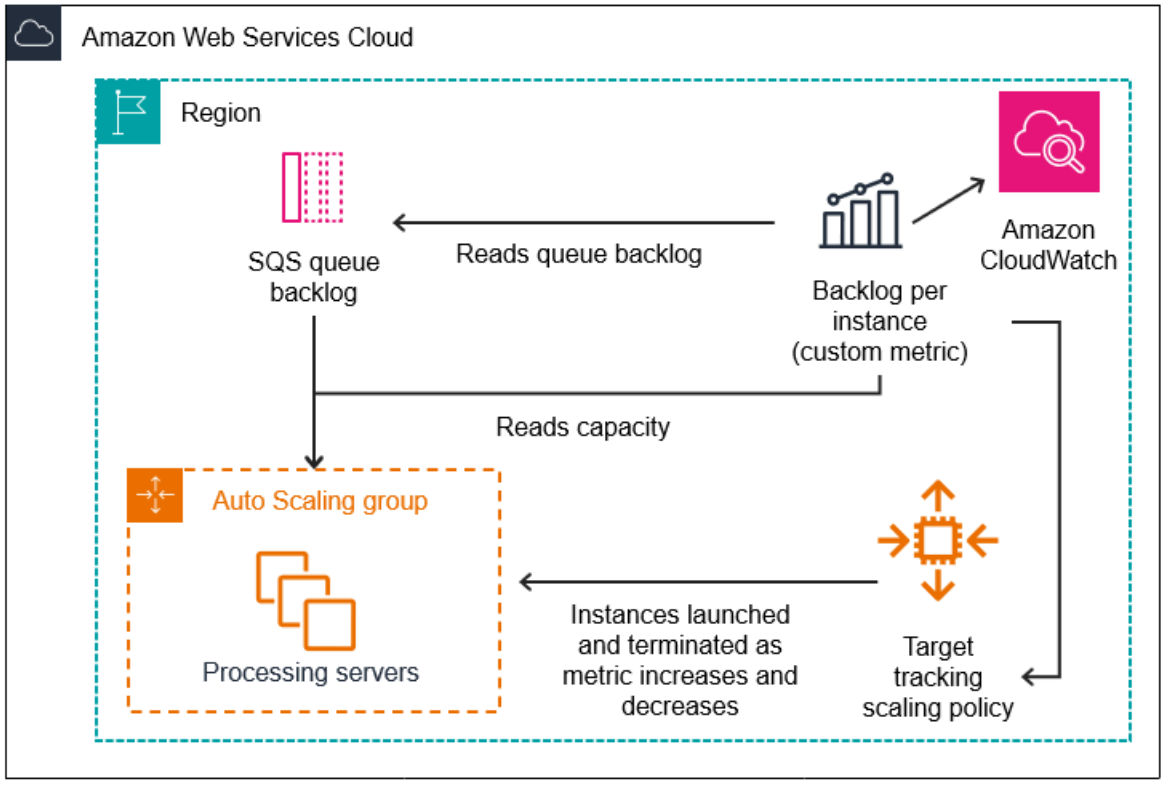
Vos informations CloudWatch métriques sont alors les suivantes.

ID	CloudWatch métrique	Statistique	Période
m1	ApproximateNumberOfMessagesVisible	Somme	1 minute
m2	GroupInServiceInstances	Moyenne	1 minute

Votre ID de mathématiques appliquées aux métriques et votre expression sont les suivantes :

ID	Expression
e1	`(m1)/(m2)`

Le schéma suivant illustre l'architecture de cette métrique :



Pour utiliser cette expression mathématique appliquée à une métrique pour créer une politique de suivi des cibles (AWS CLI)

1. Stockez l'expression mathématique appliquée aux métriques dans le cadre d'une spécification métrique personnalisée dans un fichier JSON nommé `config.json`.

   Utilisez l'exemple suivant pour vous aider à démarrer. Remplacez chaque *espace réservé à la saisie de l'utilisateur* par vos propres informations.

```
{
    "CustomizedMetricSpecification": {
        "Metrics": [
            {
                "Label": "Get the queue size (the number of messages waiting to be
 processed)",
                "Id": "m1",
                "MetricStat": {
                    "Metric": {
                        "MetricName": "ApproximateNumberOfMessagesVisible",
                        "Namespace": "AWS/SQS",
                        "Dimensions": [
                            {
                                "Name": "QueueName",
                                "Value": "my-queue"
                            }
                        ]
                    },
                    "Stat": "Sum"
                },
                "ReturnData": false
            },
            {
                "Label": "Get the group size (the number of InService instances)",
                "Id": "m2",
                "MetricStat": {
                    "Metric": {
                        "MetricName": "GroupInServiceInstances",
                        "Namespace": "AWS/AutoScaling",
                        "Dimensions": [
                            {
                                "Name": "AutoScalingGroupName",
                                "Value": "my-asg"
                            }
```

```
                    ]
                },
                "Stat": "Average"
            },
            "ReturnData": false
        },
        {
            "Label": "Calculate the backlog per instance",
            "Id": "e1",
            "Expression": "m1 / m2",
            "ReturnData": true
        }
    ]
  },
  "TargetValue": 100
}
```

Pour plus d'informations, consultez la section TargetTrackingConfiguration dans le manuel Amazon EC2 Auto Scaling API Reference.

> **Note**
>
> Voici quelques ressources supplémentaires qui peuvent vous aider à trouver des noms de métriques, des espaces de noms, des dimensions et des statistiques pour les CloudWatch métriques :
>
> - Pour plus d'informations sur les métriques disponibles pour les AWS services, consultez les AWS services qui publient CloudWatch des métriques dans le guide de CloudWatch l'utilisateur Amazon.
> - Pour obtenir le nom, l'espace de noms et les dimensions exacts (le cas échéant) d'une CloudWatch métrique comportant le AWS CLI, consultez list-metrics.

2. Pour créer cette politique, exécutez la commande put-scaling-policy avec le fichier JSON comme entrée, tel qu'illustré dans l'exemple suivant.

```
aws autoscaling put-scaling-policy --policy-name sqs-backlog-target-tracking-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
```

En cas de succès, cette commande renvoie le nom de ressource Amazon (ARN) de la politique et les ARN des deux CloudWatch alarmes créées en votre nom.

```
{
    "PolicyARN": "arn:aws:autoscaling:us-
west-2:123456789012:scalingPolicy:228f02c2-c665-4bfd-
aaac-8b04080bea3c:autoScalingGroupName/my-asg:policyName/sqs-backlog-target-
tracking-scaling-policy",
    "Alarms": [
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e",
            "AlarmName": "TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e"
        },
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2",
            "AlarmName": "TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2"
        }
    ]
}
```

> **Note**
>
> Si cette commande génère une erreur, assurez-vous d'avoir mis à jour AWS CLI localement la dernière version.

Politiques de mise à l'échelle par étapes et simples pour Amazon EC2 Auto Scaling

Le dimensionnement par étapes et les politiques de dimensionnement simples permettent d'ajuster la capacité de votre groupe Auto Scaling par incréments prédéfinis en fonction des CloudWatch alarmes. Vous pouvez définir des stratégies de mise à l'échelle distinctes pour gérer la montée en

puissance (augmentation de la capacité) et la mise à l'échelle horizontale (diminution de la capacité) en cas de dépassement d'un seuil d'alarme.

Grâce à la mise à l'échelle par étapes et à la mise à l'échelle simple, vous pouvez créer et gérer les CloudWatch alarmes qui déclenchent le processus de dimensionnement. Lorsqu'une alarme est violée, Amazon EC2 Auto Scaling initie la politique de dimensionnement associée à cette alarme.

Nous vous recommandons vivement d'utiliser des politiques de dimensionnement du suivi des cibles pour effectuer une mise à l'échelle en fonction de mesures telles que l'utilisation moyenne du processeur ou le nombre moyen de demandes par cible. Les métriques qui diminuent lorsque la capacité augmente et augmentent lorsque la capacité diminue peuvent être utilisées pour monter ou diminuer en charge proportionnellement le nombre d'instances utilisant le suivi de cible. Cela permet de s'assurer qu'Amazon EC2 Auto Scaling suit de près la courbe des demandes pour vos applications. Pour plus d'informations, consultez Politiques de suivi des objectifs de la mise à l'échelle.

Table des matières

Comment fonctionnent les politiques de mise à l'échelle par étapes

Pour utiliser le step scaling, vous devez d'abord créer une CloudWatch alarme qui surveille une métrique pour votre groupe Auto Scaling. Définissez la métrique, la valeur de seuil et le nombre de périodes d'évaluation qui déterminent le déclenchement d'une alarme. Créez ensuite une politique d'échelonnement qui définit comment redimensionner votre groupe lorsque le seuil d'alarme est dépassé.

Ajoutez les ajustements par étapes dans la stratégie. Vous pouvez définir différents ajustements par étapes en fonction de la taille du déclenchement de l'alarme. Par exemple :

- Diminution de 10 instances si la métrique d'alarme atteint 60 %
- Diminution de 30 instances si la métrique d'alarme atteint 75 %
- Diminution de 40 instances si la métrique d'alarme atteint 85 %

Lorsque le seuil d'alarme est dépassé pendant le nombre de périodes d'évaluation spécifié, Amazon EC2 Auto Scaling applique les ajustements d'étape définis dans la politique. Les ajustements peuvent se poursuivre pour d'autres déclenchements d'alarme jusqu'à ce que l'état de l'alarme revienne à OK.

Chaque instance dispose d'une période de préchauffage afin d'éviter que les activités de dimensionnement ne soient trop réactives aux modifications survenant sur de courtes périodes. Vous pouvez éventuellement configurer la période de préchauffage pour votre politique de dimensionnement. Toutefois, nous vous recommandons d'utiliser le préchauffage de l'instance par défaut pour faciliter la mise à jour de toutes les politiques de dimensionnement lorsque le temps de préchauffage change. Pour plus d'informations, consultez Définir la préparation par défaut d'instance d'un groupe Auto Scaling.

Les politiques de dimensionnement simples sont similaires aux politiques de dimensionnement par étapes, sauf qu'elles sont basées sur un seul ajustement de mise à l'échelle, avec un délai de recharge entre chaque activité de dimensionnement. Pour plus d'informations, consultez Politiques de mise à l'échelle simples.

Ajustements d'étape pour la mise à l'échelle par étapes

Lorsque vous créez une politique de mise à l'échelle par étapes, vous spécifiez un ou plusieurs ajustements par étapes qui redimensionnent automatiquement le nombre d'instances de manière dynamique en fonction de la taille du seuil de l'alarme. Chaque ajustement par étapes précise ce qui suit :

- Une limite inférieure pour la valeur de la métrique
- Une limite supérieure pour la valeur de la métrique
- L'ampleur de la mise à l'échelle, en fonction du type d'ajustement de la mise à l'échelle

CloudWatch agrège les points de données métriques en fonction des statistiques de la métrique associée à votre CloudWatch alarme. En cas de violation de l'alarme, la stratégie de mise à l'échelle appropriée est appelée. Amazon EC2 Auto Scaling applique le type d'agrégation aux points de données métriques les plus récents provenant CloudWatch (par opposition aux données métriques

brutes). Il compare cette valeur de métrique regroupée aux limites supérieures et inférieures définies par les ajustements d'étape afin de déterminer l'ajustement d'étape à réaliser.

Vous spécifiez les limites supérieure et inférieure par rapport au seuil d'une utilisation hors limites. Supposons, par exemple, que vous ayez défini une CloudWatch alarme et une politique de scale-out lorsque la métrique est supérieure à 50 %. Vous avez ensuite défini une deuxième alarme et une stratégie de mise à l'échelle horizontale pour les cas où la métrique est inférieure à 50 %. Vous avez effectué une série d'ajustements par étapes avec un type d'ajustement `PercentChangeInCapacity` (ou pourcentage du groupe dans la console) pour chaque politique :

Exemple : ajustements par étapes pour la politique d'évolutivité horizontale

Limite inférieure	Limite supérieure	Ajustement
0 USD	10	0 USD
10	20	10
20	null	30

Exemple : ajustements par étapes pour la politique de mise à l'échelle horizontale

Limite inférieure	Limite supérieure	Ajustement
-10	0	0
-20	-10	-10
null	-20	-30

La configuration de mise à l'échelle suivante est ainsi créée.

```
Metric value

-infinity          30%    40%          60%     70%             infinity
-------------------------------------------------------------------------
          -30%      | -10% | Unchanged   | +10%  |       +30%
-------------------------------------------------------------------------
```

Supposons maintenant que vous utilisiez cette configuration de dimensionnement sur un groupe Auto Scaling qui possède à la fois une capacité actuelle et une capacité souhaitée de 10. Les points suivants résument le comportement de la configuration de mise à l'échelle par rapport à la capacité souhaitée et actuelle du groupe :

- La capacité actuelle et souhaitée est conservée, tandis que la valeur de la métrique regroupée est supérieure à 40 et inférieure à 60.
- Si la valeur métrique atteint 60, la capacité souhaitée du groupe augmente d'1 instance pour atteindre 11 instances, en fonction du deuxième ajustement par étapes de la politique d'évolutivité horizontale (ajoutez 10 pour cent de 10 instances). Une fois que la nouvelle instance est en cours d'exécution et que le temps de préchauffage spécifié a expiré, la capacité actuelle du groupe passe à 11 instances. Si la valeur de la métrique s'élève à 70 même après cette augmentation de capacité, la capacité désirée du groupe augmente de 3 nouvelles instances, soit 14 instances. Le comportement est basé sur l'ajustement de la troisième étape de la politique d'évolutivité horizontale (ajouter 30 % des 11 instances, soit 3,3 instances, arrondies à 3 instances).
- Si la valeur de métrique atteint 40, la capacité souhaitée du groupe est réduite d'1 instance pour atteindre 13 instances, en fonction du deuxième ajustement d'étape de la politique de mise à l'échelle horizontale (enlevez 10 pour cent de 14 instances, 1.4 instances arrondies à 1). Si la valeur de la métrique tombe à 30 même après cette diminution de capacité, la capacité désirée du groupe diminue de 3 instances, soit 10 instances. Le comportement est basé sur l'ajustement de la troisième étape de la politique de mise à l'échelle horizontale (supprimer 30 % des 13 instances, soit 3,9 instances, arrondies à 3 instances).

Lorsque vous spécifiez les ajustements d'étape pour votre politique de mise à l'échelle, notez les points suivants :

- Si vous utilisez le AWS Management Console, vous spécifiez les limites supérieure et inférieure sous forme de valeurs absolues. Si vous utilisez le AWS CLI ou un SDK, vous spécifiez les limites supérieure et inférieure par rapport au seuil de violation.
- Les plages d'ajustements d'étape peuvent se chevaucher ou avoir un écart.
- Un seul ajustement d'étape peut avoir une limite inférieure null (infini négatif). Si un seul ajustement d'étape possède une limite inférieure négative, un ajustement d'étape avec une limite inférieure null doit donc exister.
- Un seul ajustement d'étape peut avoir une limite supérieure null (infini positif). Si un seul ajustement d'étape possède une limite supérieure positive, un ajustement d'étape avec une limite supérieure null doit donc exister.

- Les limites supérieure et inférieure ne peuvent pas être null dans le même ajustement d'étape.
- Si la valeur métrique dépasse le seuil, la limite inférieure est inclusive et la limite supérieure est exclusive. Si la valeur métrique n'atteint pas le seuil, la limite inférieure est exclusive et la limite supérieure est inclusive.

Types d'ajustement de la mise à l'échelle

Vous pouvez définir une politique de mise à l'échelle qui exécute l'action de mise à l'échelle optimale, en fonction du type d'ajustement de mise à l'échelle que vous choisissez. Vous pouvez spécifier le type d'ajustement en pourcentage de la capacité actuelle de votre groupe Auto Scaling ou en unités de capacité. Normalement, une unité de capacité signifie une instance, sauf si vous utilisez la fonction de pondération des instances.

Amazon EC2 Auto Scaling prend en charge les types d'ajustement suivants pour la mise à l'échelle par étapes et la mise à l'échelle simple :

- `ChangeInCapacity` — Augmenter ou réduire la capacité actuelle du groupe en fonction de la valeur spécifiée. Une valeur positive augmente la capacité et une valeur d'ajustement négative la réduit. Par exemple : si la capacité actuelle du groupe est de 3 et que l'ajustement est de 5, lorsque cette politique est exécutée, nous ajoutons 5 unités de capacité à la capacité pour un total de 8 unités de capacité.
- `ExactCapacity` — Modifier la capacité actuelle du groupe à la valeur spécifiée. Spécifiez une valeur non négative avec ce type d'ajustement. Exemple : si la capacité actuelle du groupe est de 3 instances et l'ajustement de 5, lorsque la politique est appliquée, la nouvelle capacité est de 5 unités.
- `PercentChangeInCapacity` — Augmenter ou réduire la capacité actuelle du groupe en fonction du pourcentage spécifié. Une valeur positive augmente la capacité et une valeur négative la réduit. Par exemple : si la capacité actuelle est de 10 et que l'ajustement est de 10 %, lorsque cette politique est exécutée, nous ajoutons 1 unité de capacité à la capacité pour un total de 11 unités de capacité.

> **Note**
>
> Si la valeur générée n'est pas un nombre entier, elle est arrondie comme suit :
>
> - Les valeurs supérieures à 1 sont arrondies à l'unité inférieure. Par exemple, `12.7` est arrondi à `12`.
> - Les valeurs comprises entre 0 et 1 sont arrondies à 1. Par exemple, `.67` est arrondi à `1`.

- Les valeurs comprises entre 0 et -1 sont arrondies à -1. Par exemple, `-.58` est arrondi à `-1`.
- Les valeurs inférieures à -1 sont arrondies à l'unité supérieure. Par exemple, `-6.67` est arrondi à `-6`.

Avec `PercentChangeInCapacity`, vous pouvez également spécifier le nombre minimal d'instances à mettre à l'échelle à l'aide du paramètre `MinAdjustmentMagnitude`. Par exemple, imaginons que vous ayez créé une politique qui ajoute 25 % et que vous spécifiiez un incrément minimal de 2 instances. Si vous disposez d'un groupe Auto Scaling avec 4 instances et que la politique de mise à l'échelle est exécutée, 25 % de 4 est égal à 1 instance. Cependant, comme vous avez spécifié un incrément minimal de 2, 2 instances sont ajoutées.

Lorsque vous utilisez des pondérations d'instance, l'effet du réglage du `MinAdjustmentMagnitude` paramètre sur une valeur différente de zéro change. La valeur est exprimée en unités de capacité. Pour définir le nombre minimal d'instances à mettre à l'échelle, définissez ce paramètre sur une valeur au moins égale à la pondération maximale de votre instance.

Si vous utilisez des poids d'instance, gardez à l'esprit que la capacité actuelle de votre groupe Auto Scaling peut dépasser la capacité souhaitée selon les besoins. Si votre nombre absolu à décrémenter, ou le montant que le pourcentage indique à décrémenter, est inférieur à la différence entre la capacité actuelle et la capacité souhaitée, aucune action de mise à l'échelle n'est effectuée. Vous devez prendre ces comportements en compte lorsque vous examinez les résultats d'une stratégie de mise à l'échelle lorsqu'un seuil d'alarme est dépassé. Par exemple, supposons que la capacité désirée soit de 30 et la capacité actuelle de 32. En cas de violation de l'alarme, si la stratégie de mise à l'échelle réduit de 1 la capacité souhaitée, aucune action de mise à l'échelle n'est effectuée.

Préparation d'instance

Pour les stratégies de mise à l'échelle par étapes, vous pouvez facultativement spécifier le nombre de secondes nécessaires pour la préparation d'une instance nouvellement lancée. Tant que le temps de préchauffage spécifié n'est pas expiré, une instance n'est pas prise en compte dans les métriques d'instance EC2 agrégées du groupe Auto Scaling.

Lorsque les instances sont en période de préchauffage, vos politiques de dimensionnement ne sont redimensionnées que si la valeur métrique des instances qui ne sont pas en phase de préchauffage est supérieure au seuil d'alarme maximal de la politique.

Si le groupe est à nouveau monté en puissance, les instances qui sont toujours en cours de préparation sont comptées dans le cadre de la capacité souhaitée pour la prochaine activité de montée en puissance. Par conséquent, plusieurs seuils de l'alarme inclus dans la plage du même ajustement par étapes génèrent une activité de mise à l'échelle unique. L'objectif est d'effectuer une montée en puissance continue (mais pas excessive).

Par exemple, supposons que vous créez une politique en deux étapes. La première étape ajoute 10 % lorsque la métrique atteint 60, et la deuxième étape ajoute 30 % lorsque la métrique atteint 70 %. Votre groupe Auto Scaling a une capacité souhaitée et actuelle de 10. La capacité actuelle et souhaitée est conservée, tandis que la valeur de la métrique agrégée est inférieure à 60. Supposons que la métrique atteigne 60, donc qu'une instance soit ajoutée (10 % des 10 instances). Ensuite, la métrique atteint 62 alors que la nouvelle instance est toujours en cours de préparation. La politique de mise à l'échelle calcule la nouvelle capacité souhaitée en fonction de la capacité actuelle, qui est toujours de 10. Cependant, la capacité souhaitée du groupe est déjà de 11 instances. Donc, la politique de mise à l'échelle n'augmente pas davantage la capacité souhaitée. Si la valeur métrique atteint 70 alors que l'instance est toujours en cours de préparation, nous devons ajouter 3 instances (30 pour cent de 10 instances). Cependant, la capacité souhaitée du groupe est déjà de 11, donc nous ajoutons uniquement 2 instances, pour atteindre une nouvelle capacité souhaitée de 13 instances.

Pendant que l'activité de mise à l'échelle est en cours, toutes les activités de mise à l'échelle initiées par les politiques de mise à l'échelle sont bloquées jusqu'à ce que les instances aient fini leur préparation. Lorsque les instances ont terminé la préparation, si un événement de mise à l'échelle horizontale se produit, toutes les instances en cours de résiliation seront prises en compte dans la capacité actuelle du groupe lors du calcul de la nouvelle capacité souhaitée. Par conséquent, nous n'enlevons pas plus d'instances du groupe Auto Scaling que nécessaire. Par exemple, alors qu'une instance est déjà en cours de résiliation, si un seuil d'alarme est dépassé dans la plage de réglage de la même étape qui a réduit de 1 la capacité souhaitée, aucune action de mise à l'échelle n'est effectuée.

Valeur par défaut

Si aucune valeur n'est définie, la politique de dimensionnement utilisera la valeur par défaut, qui est la valeur du préchauffage d'instance par défaut défini pour le groupe. Si le préchauffage de l'instance par défaut est nul, il revient à la valeur du temps de recharge par défaut.

Considérations

Les considérations suivantes s’appliquent lors de l’utilisation de politiques de mise à l’échelle par étapes et simples :

- Déterminez si vous pouvez prédire les ajustements d'étape sur l'application avec suffisamment de précision pour utiliser la mise à l'échelle par étapes. Si votre métrique de mise à l'échelle augmente ou diminue proportionnellement à la capacité de la cible évolutive, nous vous recommandons d'utiliser plutôt une politique de suivi des cibles et de mise à l'échelle. Vous avez toujours la possibilité d'utiliser la mise à l'échelle par étapes comme politique supplémentaire pour une configuration plus avancée. Par exemple, si vous le souhaitez, vous pouvez configurer une réponse plus agressive lorsque l'utilisation atteint un certain niveau.
- Assurez-vous de choisir une marge adéquate entre les seuils de mise à l'échelle horizontale et de mise à l'échelle avec montée en puissance parallèle, afin d'éviter tout battement. Le battement est une boucle infinie de mise à l'échelle horizontale et de montage en puissance. En d'autres termes, si une action de mise à l'échelle est effectuée, la valeur de la métrique changera et déclenchera une autre action de mise à l'échelle dans le sens inverse.

Créez une politique de dimensionnement par étapes pour le scalage externe

Pour créer une politique de dimensionnement par étapes à appliquer à votre groupe Auto Scaling pour votre groupe Auto Scaling, appliquez l'une des méthodes suivantes :

Console

Étape 1 : créer une CloudWatch alarme pour le seuil supérieur de la métrique

1. Ouvrez la CloudWatch console à l'adresse https://console.aws.amazon.com/cloudwatch/.
2. Si nécessaire, changez la région. Dans la barre de navigation, sélectionnez la région où réside votre groupe Auto Scaling.
3. Dans le panneau de navigation, choisissez Alarmes, Toutes les Alarmes, puis choisissez Créer une alarme.
4. Choisissez Select metric (Sélectionner une métrique).
5. Sous l'onglet Toutes les métriques, choisissez EC2, Par groupe Auto Scaling, puis saisissez le nom du groupe Auto Scaling dans le champ de recherche. Ensuite, sélectionnez `CPUUtilization` et choisissez Select metric (Sélectionner une métrique). La page Specify

metric and conditions (Spécifier les métriques et les conditions) apparaît, présentant un graphique et d'autres informations sur la métrique.

6. Sous Period (Période), choisissez la période d'évaluation de l'alarme, par exemple, 1 minute. Lors de l'évaluation de l'alarme, chaque période est regroupée en un point de données.

> **Note**
>
> Une période plus courte crée une alarme plus sensible.

7. Sous Conditions, procédez comme suit :

   - Pour Threshold type (Type de seuil), choisissez Static (Statique).
   - Pour **`CPUUtilization`**Whenever is, spécifiez si vous souhaitez que la valeur de la métrique soit supérieure, supérieure ou égale au seuil de violation de l'alarme. Ensuite, sous than (à), entrez la valeur de seuil qui doit appeler l'alarme.

> **Important**
>
> Pour qu'une alarme puisse être utilisée avec une politique de montée en puissance (métrique élevée), assurez-vous de ne pas choisir une valeur inférieure ou égale au seuil.

8. Sous Additional configuration (Configuration supplémentaire), procédez comme suit :

   - Pour Datapoints to alarm (Points de données pour le déclenchement d'alarme), saisissez le nombre de points de données (périodes d'évaluation) au cours desquels la valeur de métrique doit répondre aux conditions de seuil pour appeler l'alarme. Par exemple, l'alarme se déclenchera au bout de 10 minutes si vous sélectionnez deux périodes consécutives de 5 minutes pour invoquer l'état de l'alarme.
   - Pour Missing data treatment (Traitement des données manquantes), choisissez Treat missing data as bad (breaching threshold) (Traiter les données manquantes comme incorrectes [seuil dépassé]). Pour plus d'informations, consultez la section Configuration de la manière dont les CloudWatch alarmes traitent les données manquantes dans le guide de CloudWatch l'utilisateur Amazon.

9. Choisissez Suivant.

   La page Configure actions (Configurer des actions) apparait.

10. Sous Notification, sélectionnez la rubrique Amazon SNS qui doit recevoir une notification lorsque l'alerte passe à l'état `ALARM`, `OK` ou `INSUFFICIENT_DATA`.

    Pour que l'alerte envoie plusieurs notifications pour le même état d'alerte ou pour les différents états d'alerte, choisissez Add notification (Ajouter une notification).

    Pour que l'alerte n'envoie pas de notifications, choisissez Remove (Supprimer).
11. Vous pouvez quitter les autres sections de la page Configurer les actions qui est vide. Laisser les autres sections vides crée une alarme sans l'associer à une politique de stabilisation. Vous pouvez ensuite associer l'alarme à une politique de mise à l'échelle à partir de la console Amazon EC2 Auto Scaling.
12. Choisissez Suivant.
13. Saisissez un nom (par exemple, `Step-Scaling-AlarmHigh-AddCapacity`) et, éventuellement, une description de l'alarme, puis choisissez Next (Suivant).
14. Sélectionnez Créer une alerte.

Suivez la procédure ci-dessous pour continuer là où vous vous êtes arrêté après avoir créé votre CloudWatch alarme.

Étape 2 : Création d'une politique de dimensionnement par étapes pour le scaling out

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Vérifiez que les limites de mise à l'échelle sont correctement définies. Par exemple, si le groupe est déjà au maximum de sa taille, vous devez spécifier un nouveau maximum pour monter en puissance. Pour plus d'informations, consultez Définissez des limites de mise à l'échelle pour votre groupe Auto Scaling.
4. Sous l'onglet Scalabilité automatique, dans Politiques de mise à l'échelle dynamique, choisissez Créer une politique de mise à l'échelle dynamique.
5. Pour Type de stratégie, choisissez Step scaling, puis spécifiez le nom de la stratégie.
6. Pour l'CloudWatch alarme, choisissez votre alarme. Si vous n'avez pas encore créé d'alarme, choisissez Créer une CloudWatch alarme et effectuez les étapes 4 à 14 de la procédure précédente pour créer une alarme.

7. Spécifiez la modification de la taille de groupe actuelle que cette politique effectuera lors de l'exécution à l'aide de Take the action (Exécuter l'action). Vous pouvez ajouter un nombre spécifique d'instances ou un pourcentage de la taille de groupe existante, ou définir le groupe avec une taille exacte.

   Par exemple, pour créer une politique de scale-out qui augmente la capacité du groupe de 30 %, choisissezAdd, entrez `30` dans le champ suivant, puis choisissez. `percent of group` Par défaut, la limite inférieure pour cet ajustement d'étape est le seuil de l'alarme et la limite supérieure est l'infini positif (+).
8. Pour ajouter une autre étape, choisissez Add step (Ajouter étape) puis définissez la quantité de mise à l'échelle et les limites inférieure et supérieure de l'étape par rapport au seuil d'alarme.
9. Pour définir un nombre minimal d'instances à mettre à l'échelle, mettez à jour le champ numérique dans Add capacity units in increments of at least (Ajouter des unités de capacité par incréments d'au moins) 1 unité de capacité.
10. (Facultatif) Pour le préchauffage de l'instance, mettez à jour la valeur de préchauffage de l'instance selon les besoins.
11. Choisissez Créer.

AWS CLI

Pour créer une politique de dimensionnement par étapes pour le scalage (augmentation de la capacité), vous pouvez utiliser les exemples de commandes suivants. Remplacez chaque *`espace réservé à la saisie de l'utilisateur`* par vos propres informations.

Lorsque vous utilisez le AWS CLI, vous créez d'abord une politique de dimensionnement par étapes qui fournit des instructions à Amazon EC2 Auto Scaling sur la manière de procéder à une mise à l'échelle lorsque la valeur d'une métrique augmente. Vous créez ensuite l'alarme en identifiant la métrique à surveiller, en définissant le seuil supérieur de la métrique et d'autres détails relatifs aux alarmes, et en associant l'alarme à la politique de dimensionnement.

Étape 1 : créer une politique de mise à l'échelle

Utilisez la commande put-scaling-policy suivante pour créer une politique de dimensionnement par étapes nomméemy-`step-scale-out-policy`, avec un type d'ajustement `PercentChangeInCapacity` qui augmente la capacité du groupe en fonction des ajustements d'étape suivants (en supposant un seuil CloudWatch d'alarme de 60 %) :

- Augmentez le nombre d'instances de 10 % lorsque la valeur de la métrique est supérieure ou égale à 60 %, mais inférieure à 75 %
- Augmentez le nombre d'instances de 20 % lorsque la valeur de la métrique est supérieure ou égale à 75 %, mais inférieure à 85 %
- Augmenter le nombre d'instances de 30 % quand la valeur de la métrique est supérieure ou égale à 85 %

```
aws autoscaling put-scaling-policy \
  --auto-scaling-group-name my-asg  \
  --policy-name my-step-scale-out-policy \
  --policy-type StepScaling \
  --adjustment-type PercentChangeInCapacity \
  --metric-aggregation-type Average \
  --step-adjustments
 MetricIntervalLowerBound=0.0,MetricIntervalUpperBound=15.0,ScalingAdjustment=10 \

 MetricIntervalLowerBound=15.0,MetricIntervalUpperBound=25.0,ScalingAdjustment=20 \
                     MetricIntervalLowerBound=25.0,ScalingAdjustment=30 \
  --min-adjustment-magnitude 1
```

Notez le nom Amazon Resource Name (ARN) de la politique. Vous en avez besoin pour créer une CloudWatch alarme pour la politique.

```
{
    "PolicyARN":
 "arn:aws:autoscaling:region:123456789012:scalingPolicy:4ee9e543-86b5-4121-b53b-
aa4c23b5bbcc:autoScalingGroupName/my-asg:policyName/my-step-scale-in-policy
}
```

Étape 2 : créer une CloudWatch alarme pour le seuil supérieur de la métrique

Utilisez la commande CloudWatch put-metric-alarm suivante pour créer une alarme qui augmente la taille du groupe Auto Scaling sur la base d'une valeur seuil moyenne du processeur de 60 % pendant au moins deux périodes d'évaluation consécutives de deux minutes. Pour utiliser votre propre métrique personnalisée, spécifiez son nom dans `--metric-name` et son espace de noms dans `--namespace`.

```
aws cloudwatch put-metric-alarm --alarm-name Step-Scaling-AlarmHigh-AddCapacity \
```

```
--metric-name CPUUtilization --namespace AWS/EC2 --statistic Average \
--period 120 --evaluation-periods 2 --threshold 60 \
--comparison-operator GreaterThanOrEqualToThreshold \
--dimensions "Name=AutoScalingGroupName,Value=my-asg" \
--alarm-actions PolicyARN
```

Créez une politique de dimensionnement par étapes pour la mise à l'échelle dans

Pour créer une politique de dimensionnement par étapes à appliquer à votre groupe Auto Scaling pour votre groupe Auto Scaling, appliquez l'une des méthodes suivantes :

Console

Étape 1 : créer une CloudWatch alarme pour le seuil bas de la métrique

1. Ouvrez la CloudWatch console à l'adresse https://console.aws.amazon.com/cloudwatch/.
2. Si nécessaire, changez la région. Dans la barre de navigation, sélectionnez la région où réside votre groupe Auto Scaling.
3. Dans le panneau de navigation, choisissez Alarmes, Toutes les Alarmes, puis choisissez Créer une alarme.
4. Choisissez Select metric (Sélectionner une métrique).
5. Sous l'onglet Toutes les métriques, choisissez EC2, Par groupe Auto Scaling, puis saisissez le nom du groupe Auto Scaling dans le champ de recherche. Ensuite, sélectionnez `CPUUtilization` et choisissez Select metric (Sélectionner une métrique). La page Specify metric and conditions (Spécifier les métriques et les conditions) apparaît, présentant un graphique et d'autres informations sur la métrique.
6. Sous Period (Période), choisissez la période d'évaluation de l'alarme, par exemple, 1 minute. Lors de l'évaluation de l'alarme, chaque période est regroupée en un point de données.

   > **Note**
   >
   > Une période plus courte crée une alarme plus sensible.

7. Sous Conditions, procédez comme suit :

   - Pour Threshold type (Type de seuil), choisissez Static (Statique).

- Pour **`CPUUtilization`**Whenever is, spécifiez si vous souhaitez que la valeur de la métrique soit inférieure, inférieure ou égale au seuil de violation de l'alarme. Ensuite, sous than (à), entrez la valeur de seuil qui doit appeler l'alarme.

> **Important**
>
> Pour qu'une alarme puisse être utilisée avec une politique de descente en puissance (métrique faible), assurez-vous de ne pas choisir une valeur supérieure ou supérieure ou égale au seuil.

8. Sous Additional configuration (Configuration supplémentaire), procédez comme suit :

   - Pour Datapoints to alarm (Points de données pour le déclenchement d'alarme), saisissez le nombre de points de données (périodes d'évaluation) au cours desquels la valeur de métrique doit répondre aux conditions de seuil pour appeler l'alarme. Par exemple, l'alarme se déclenchera au bout de 10 minutes si vous sélectionnez deux périodes consécutives de 5 minutes pour invoquer l'état de l'alarme.
   - Pour Missing data treatment (Traitement des données manquantes), choisissez Treat missing data as bad (breaching threshold) (Traiter les données manquantes comme incorrectes [seuil dépassé]). Pour plus d'informations, consultez la section Configuration de la manière dont les CloudWatch alarmes traitent les données manquantes dans le guide de CloudWatch l'utilisateur Amazon.

9. Choisissez Suivant.

   La page Configure actions (Configurer des actions) apparait.

10. Sous Notification, sélectionnez la rubrique Amazon SNS qui doit recevoir une notification lorsque l'alerte passe à l'état `ALARM`, `OK` ou `INSUFFICIENT_DATA`.

    Pour que l'alerte envoie plusieurs notifications pour le même état d'alerte ou pour les différents états d'alerte, choisissez Add notification (Ajouter une notification).

    Pour que l'alerte n'envoie pas de notifications, choisissez Remove (Supprimer).

11. Vous pouvez quitter les autres sections de la page Configurer les actions qui est vide. Laisser les autres sections vides crée une alarme sans l'associer à une politique de stabilisation. Vous pouvez ensuite associer l'alarme à une politique de mise à l'échelle à partir de la console Amazon EC2 Auto Scaling.

12. Choisissez Suivant.

13. Saisissez un nom (par exemple, `Step-Scaling-AlarmLow-RemoveCapacity`) et, éventuellement, une description de l'alarme, puis choisissez Next (Suivant).
14. Sélectionnez Créer une alerte.

Suivez la procédure ci-dessous pour continuer là où vous vous êtes arrêté après avoir créé votre CloudWatch alarme.

Étape 2 : Création d'une politique de dimensionnement par étapes pour la mise à l'échelle

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Vérifiez que les limites de mise à l'échelle sont correctement définies. Par exemple, si la capacité souhaitée par votre groupe est déjà au minimum, vous devez spécifier un nouveau minimum pour pouvoir l'augmenter. Pour plus d'informations, consultez Définissez des limites de mise à l'échelle pour votre groupe Auto Scaling.
4. Sous l'onglet Scalabilité automatique, dans Politiques de mise à l'échelle dynamique, choisissez Créer une politique de mise à l'échelle dynamique.
5. Pour Type de stratégie, choisissez Step scaling, puis spécifiez le nom de la stratégie.
6. Pour l'CloudWatch alarme, choisissez votre alarme. Si vous n'avez pas encore créé d'alarme, choisissez Créer une CloudWatch alarme et effectuez les étapes 4 à 14 de la procédure précédente pour créer une alarme.
7. Spécifiez la modification de la taille de groupe actuelle que cette politique effectuera lors de l'exécution à l'aide de Take the action (Exécuter l'action). Vous pouvez supprimer un nombre spécifique d'instances ou un pourcentage de la taille de groupe existante, ou définir le groupe sur une taille exacte.

   Par exemple, pour créer une politique d'évolutivité qui réduit la capacité du groupe de deux instances, choisissez`Remove`, entrez 2 dans le champ suivant, puis choisissez. `capacity units` Par défaut, la limite supérieure de cet ajustement d'étape est le seuil de l'alarme et la limite inférieure est l'infini négatif (-).
8. Pour ajouter une autre étape, choisissez Ajouter étape puis définissez la quantité de mise à l'échelle et les limites inférieure et supérieure de l'étape par rapport au seuil d'alarme.
9. Choisissez Créer.

AWS CLI

Pour créer une politique de dimensionnement par étapes pour la mise à l'échelle (diminution de la capacité), vous pouvez utiliser les exemples de commandes suivants. Remplacez chaque *espace réservé à la saisie de l'utilisateur* par vos propres informations.

Lorsque vous utilisez le AWS CLI, vous créez d'abord une politique de dimensionnement par étapes qui fournit des instructions à Amazon EC2 Auto Scaling sur la manière de procéder lorsque la valeur d'une métrique diminue. Vous créez ensuite l'alarme en identifiant la métrique à surveiller, en définissant le seuil bas de la métrique et d'autres détails relatifs aux alarmes, et en associant l'alarme à la politique de dimensionnement.

Étape 1 : créer une politique de mise à l'échelle

Utilisez la commande put-scaling-policy suivante pour créer une politique de dimensionnement par étapes nomméemy-step-scale-in-policy, avec un type d'ajustement de type ChangeInCapacity qui réduit la capacité du groupe de 2 instances lorsque l' CloudWatch alarme associée dépasse le seuil bas de la métrique.

```
aws autoscaling put-scaling-policy \
  --auto-scaling-group-name my-asg  \
  --policy-name my-step-scale-in-policy \
  --policy-type StepScaling \
  --adjustment-type ChangeInCapacity \
  --step-adjustments MetricIntervalUpperBound=0.0,ScalingAdjustment=-2
```

Notez le nom Amazon Resource Name (ARN) de la politique. Vous en avez besoin pour créer l' CloudWatch alarme pour la politique.

```
{
    "PolicyARN": "arn:aws:autoscaling:region:123456789012:scalingPolicy:ac542982-
cbeb-4294-891c-a5a941dfa787:autoScalingGroupName/my-asg:policyName/my-step-scale-
out-policy
}
```

Étape 2 : créer une CloudWatch alarme pour le seuil bas de la métrique

Utilisez la commande CloudWatch put-metric-alarm suivante pour créer une alarme qui réduit la taille du groupe Auto Scaling sur la base d'une valeur seuil moyenne du processeur de 40 % pendant au moins deux périodes d'évaluation consécutives de deux minutes. Pour utiliser votre

propre métrique personnalisée, spécifiez son nom dans `--metric-name` et son espace de noms dans `--namespace`.

```
aws cloudwatch put-metric-alarm --alarm-name Step-Scaling-AlarmLow-RemoveCapacity \
  --metric-name CPUUtilization --namespace AWS/EC2 --statistic Average \
  --period 120 --evaluation-periods 2 --threshold 40 \
  --comparison-operator LessThanOrEqualToThreshold \
  --dimensions "Name=AutoScalingGroupName,Value=my-asg" \
  --alarm-actions PolicyARN
```

Politiques de mise à l'échelle simples

Les exemples suivants montrent comment utiliser les commandes CLI pour créer des politiques de dimensionnement simples. Elles figurent toujours dans ce document à titre de référence pour tous les clients qui souhaitent les utiliser, mais nous vous recommandons d'utiliser plutôt des politiques de suivi des cibles ou de dimensionnement par étapes.

À l'instar des politiques de dimensionnement par étapes, les politiques de dimensionnement simples vous obligent à créer des CloudWatch alarmes pour vos politiques de dimensionnement. Dans les politiques que vous créez, vous devez également définir s'il faut ajouter ou supprimer des instances, et combien, ou définir la taille exacte du groupe.

L'une des principales différences entre les politiques de dimensionnement par étapes et les politiques de dimensionnement simples réside dans les ajustements d'étapes que vous pouvez obtenir avec les politiques de dimensionnement par étapes. Avec la mise à l'échelle des étapes, vous pouvez apporter des modifications plus ou moins importantes à la taille du groupe en fonction des ajustements d'étapes que vous spécifiez.

Une politique de dimensionnement simple doit également attendre la fin d'une activité de dimensionnement en cours ou le remplacement d'un bilan de santé et la fin d'une période de recharge avant de répondre à des alarmes supplémentaires. En revanche, avec la mise à l'échelle progressive, la politique continue de répondre à des alarmes supplémentaires, même lorsqu'une activité de dimensionnement ou le remplacement du bilan de santé est en cours. Cela signifie qu'Amazon EC2 Auto Scaling évalue toutes les violations d'alarme à mesure qu'il reçoit les messages d'alarme. Pour cette raison, nous vous recommandons d'utiliser plutôt des politiques de dimensionnement par étapes, même si vous ne disposez que d'un seul ajustement de mise à l'échelle.

Amazon EC2 Auto Scaling prenait initialement en charge uniquement les politiques de mise à l'échelle simple. Si vous avez créé votre politique de dimensionnement avant l'introduction des politiques de suivi des cibles et de dimensionnement par étapes, votre stratégie est traitée comme une simple politique de dimensionnement.

Créez une politique de dimensionnement simple pour le scaling out

Utilisez la commande put-scaling-policy suivante pour créer une politique de dimensionnement simple nomméemy-simple-scale-out-policy, avec un type d'ajustement PercentChangeInCapacity qui augmente la capacité du groupe de 30 % lorsque l' CloudWatch alarme associée dépasse le seuil élevé de la métrique.

```
aws autoscaling put-scaling-policy --policy-name my-simple-scale-out-policy \
  --auto-scaling-group-name my-asg --scaling-adjustment 30 \
  --adjustment-type PercentChangeInCapacity
```

Notez le nom Amazon Resource Name (ARN) de la politique. Vous en avez besoin pour créer l' CloudWatch alarme pour la politique.

Créez une politique de dimensionnement simple pour la mise à l'échelle

Utilisez la commande put-scaling-policy suivante pour créer une politique de dimensionnement simple nomméemy-simple-scale-in-policy, avec un type d'ajustement ChangeInCapacity qui réduit la capacité du groupe d'une instance lorsque l' CloudWatch alarme associée dépasse le seuil bas de la métrique.

```
aws autoscaling put-scaling-policy --policy-name my-simple-scale-in-policy \
  --auto-scaling-group-name my-asg --scaling-adjustment -1 \
  --adjustment-type ChangeInCapacity --cooldown 180
```

Notez le nom Amazon Resource Name (ARN) de la politique. Vous en avez besoin pour créer l' CloudWatch alarme pour la politique.

Temps de stabilisation de la mise à l'échelle d'Amazon EC2 Auto Scaling

> **Important**
>
> Comme bonne pratique, nous vous recommandons de ne pas utiliser de politiques de mise à l'échelle simple et de stabilisation de mise à l'échelle. Une politique de suivi des cibles et de mise à l’échelle ou une politique de mise à l’échelle par étapes est meilleure

pour mettre à l’échelle les performances. Pour une politique de mise à l'échelle qui modifie proportionnellement la taille de votre groupe Auto Scaling au fur et à mesure que la valeur de la métrique de mise à l'échelle diminue ou augmente, nous vous recommandons de suivre les objectifs par une mise à l'échelle simple ou par étapes.

Lorsque vous créez des politiques de mise à l’échelle simples pour votre groupe Auto Scaling, nous vous recommandons de configurer le temps de stabilisation de mise à l’échelle en même temps.

Une fois que votre groupe Auto Scaling lance ou résilie des instances, il attend la fin du temps de stabilisation avant de commencer toute autre activité de mise à l'échelle initiale. L’objectif du temps de stabilisation est de laisser votre groupe Auto Scaling se stabiliser et de l’empêcher de lancer ou résilier des instances supplémentaires avant que les effets des activités de mise à l’échelle précédentes ne soient visibles.

Par exemple, supposons qu'une politique de mise à l'échelle simple pour l'utilisation du processeur recommande de lancer deux instances. Amazon EC2 Auto Scaling lance deux instances, puis met en pause les activités de mise à l'échelle jusqu'à la fin du temps de stabilisation. À la fin du temps de stabilisation, toutes les activités de mise à l'échelle initiées par des politiques de mise à l'échelle simple peuvent reprendre. Si l'utilisation de la CPU dépasse de nouveau le seuil élevé de l'alarme, le groupe Auto Scaling est de nouveau mis à l'échelle et le temps de stabilisation redevient effectif. Cependant, si les deux instances ont suffi à réduire la valeur de la métrique, le groupe conserve sa taille actuelle.

Table des matières

Considérations

Les considérations suivantes s'appliquent lorsque vous travaillez avec des politiques de mise à l'échelle simples et des temps de stabilisation de mise à l'échelle :

- Les politiques de suivi de cible et de mise à l'échelle par étapes peuvent lancer une activité évolutive immédiatement sans attendre la fin du temps de stabilisation. Au lieu de cela, chaque

fois que votre groupe Auto Scaling lance des instances, les instances individuelles bénéficient d'une période de préchauffage. Pour plus d'informations, consultez Définir la préparation par défaut d'instance d'un groupe Auto Scaling.

- Lorsqu'une action planifiée démarre à l'heure prévue, elle peut aussi initier une activité de mise à l'échelle immédiatement sans attendre la fin du temps de stabilisation.
- Si une instance devient défectueuse, Amazon EC2 Auto Scaling n'attend pas la fin du temps de stabilisation avant de la remplacer.
- Lorsque plusieurs instances lancent ou résilient, le temps de stabilisation (par défaut ou spécifique à la mise à l'échelle) prend effet après la fin du lancement ou de la résiliation de la dernière instance.
- Lorsque vous mettez manuellement votre groupe Auto Scaling à l'échelle, la valeur par défaut ne consiste pas à attendre la fin du temps de stabilisation. Toutefois, vous pouvez annuler ce comportement et respecter le temps de recharge par défaut lorsque vous utilisez le AWS CLI ou un SDK pour effectuer une mise à l'échelle manuelle.
- Par défaut, Elastic Load Balancing attend 300 secondes avant de terminer le processus de désinscription (drainage de la connexion). Si le groupe se trouve derrière un équilibreur de charge Elastic Load Balancing, il attend que les instances terminales se désenregistrent avant de commencer le temps de stabilisation.

Les hooks de cycle de vie peuvent entraîner des retards supplémentaires.

Si un hook de cycle de vie est appelé, le temps de stabilisation commence une fois que vous avez terminé l'action du cycle de vie ou une fois le temps d'expiration terminé. Prenons l'exemple d'un groupe Auto Scaling avec un hook de cycle de vie pour le lancement de l'instance. Lorsque l'application connaît une hausse de la demande, le groupe lance une instance pour ajouter de la capacité. En raison de la présence d'un hook de cycle de vie, l'instance est dans un état d'attente et les activités de mise à l'échelle dues à des politiques de mise à l'échelle simple sont suspendues. Lorsque l'instance a comme statut `InService`, le temps de stabilisation démarre. À la fin du temps de stabilisation, les activités de politiques de mise à l'échelle simple reprennent.

Lorsque Elastic Load Balancing est activé, à des fins de dimensionnement, le délai de recharge commence lorsque l'instance sélectionnée pour être résiliée commence à drainer la connexion (délai de désenregistrement). La période de recharge n'attend pas la fin de la vidange de la connexion ou la fin du cycle de vie du hook pour terminer son action. Cela signifie que toutes les activités de mise à l'échelle dues à de simples politiques de mise à l'échelle peuvent reprendre dès que le résultat de l'événement évolutif se reflète dans la capacité du groupe. Sinon, attendre la fin des trois

activités (drainage de la connexion, hook de cycle de vie et temps de stabilisation) augmenterait considérablement le temps nécessaire au groupe Auto Scaling pour interrompre la mise à l'échelle.

Modifier le temps de stabilisation par défaut

Vous ne pouvez pas définir le temps de stabilisation par défaut lorsque vous créez un groupe Auto Scaling dans la console Amazon EC2 Auto Scaling. Par défaut, ce temps de stabilisation est fixé à 300 secondes (5 minutes). Au besoin, vous pouvez le mettre à jour une fois le groupe créé.

Pour modifier le temps de stabilisation par défaut (console)

Après avoir créé le groupe Auto Scaling, sur l'onglet Details (Détails), choisissez Advanced configurations (Configurations avancées), Edit (Modifier). Pour Default cooldown (Temps de stabilisation par défaut), choisissez le temps que vous souhaitez en fonction des besoins de la période de démarrage de votre instance ou d'autres applications.

Pour modifier le temps de stabilisation par défaut (AWS CLI)

Utilisez les commandes suivantes pour modifier le temps de stabilisation par défaut des groupes Auto Scaling nouveaux ou existants. Si le temps de stabilisation par défaut n'est pas défini, la valeur par défaut de 300 secondes est utilisée.

- create-auto-scaling-group
- update-auto-scaling-group

Pour confirmer la valeur du temps de stabilisation par défaut, utilisez la commande describe-auto-scaling-groups.

Définir un temps de stabilisation pour des politiques de mise à l'échelle simples particulières

Par défaut, toutes les politiques de mise à l'échelle simple utilisent le temps de stabilisation par défaut défini pour le groupe Auto Scaling. Pour définir un temps de stabilisation spécifique à des politiques de mise à l'échelle simples, utilisez le paramètre de stabilisation facultatif lorsque vous créez ou mettez à jour une politique. Lorsqu'un temps de stabilisation est spécifié pour une politique, il remplace le temps de stabilisation par défaut.

Un temps de stabilisation spécifique à politique de mise à l'échelle est souvent utilisé avec une politique de mise à l'échelle horizontale. Cette politique résilie les instances, Amazon EC2

Auto Scaling a donc besoin de moins de temps pour déterminer s'il doit résilier des instances supplémentaires. La résiliation d'instances doit être une opération beaucoup plus rapide que le lancement d'instances. Le temps de stabilisation par défaut de 300 secondes est donc trop long. Dans ce cas, un temps de stabilisation spécifique à une politique de mise à l'échelle avec une valeur inférieure pour votre politique de mise à l'échelle horizontale peut vous aider à réduire les coûts en permettant au groupe de diminuer plus rapidement.

Pour créer ou mettre à jour des politiques de mise à l'échelle simples dans la console, choisissez l'onglet Automatic scaling (Scalabilité automatique) après avoir créé le groupe. Pour créer ou mettre à jour des politiques de dimensionnement simples à l'aide de AWS CLI, utilisez la commande put-scaling-policy. Pour plus d'informations, consultez Politiques de mise à l'échelle simple et par étapes.

Mise à l'échelle basée sur Amazon SQS

> **Important**
>
> Les informations et étapes suivantes vous montrent comment calculer le backlog de file d'attente Amazon SQS par instance à l'aide de l'attribut `ApproximateNumberOfMessages` queue avant de le publier sous forme de métrique personnalisée sur. CloudWatch Cependant, vous pouvez désormais réduire les coûts et les efforts consacrés à la publication de votre propre métrique en utilisant une expression mathématique appliquée à une métrique. Pour plus d'informations, consultez Créer une politique de mise à l'échelle du suivi des cibles pour Amazon EC2 Auto Scaling à l'aide d'une expression mathématique appliquée à une métrique.

Cette section vous montre comment mettre à l'échelle votre groupe Auto Scaling en réponse aux modifications de la charge du système dans une file d'attente Amazon Simple Queue Service (Amazon SQS). Pour en savoir plus sur la façon dont vous pouvez utiliser Amazon SQS, veuillez consulter le Guide du développeur Amazon Simple Queue Service.

Il existe des scénarios dans lesquels vous pourriez envisager la mise à l'échelle en réponse à une activité dans une file d'attente Amazon SQS. Supposons par exemple que vous disposez d'une application web qui permet aux utilisateurs de charger des images et de les utiliser en ligne. Dans ce scénario, chaque image doit être codée et redimensionnée avant de pouvoir être publiée. L'application s'exécute sur des instances EC2 dans un groupe Auto Scaling et elle est configurée pour gérer les taux de chargement classiques. Les instances non saines sont résiliées et remplacées pour maintenir des niveaux d'instance actuels à tout moment. L'application place les données bitmap

brutes des images dans une file d'attente SQS afin qu'elles soient traitées. Elle traite les images, puis publie les images traitées à un emplacement où elles peuvent être affichées par les utilisateurs. L'architecture de ce scénario fonctionne correctement si le nombre de chargements d'images ne varie pas au fil du temps. En revanche, si le nombre de chargements varie au fil du temps, vous pouvez envisager d'utiliser la mise à l'échelle dynamique pour mettre à l'échelle la capacité de votre groupe Auto Scaling.

Table des matières

Utiliser un suivi de la cible avec la métrique appropriée

Si vous utilisez une politique de suivi des objectifs et d'échelonnement basée sur une métrique de file d'attente Amazon SQS personnalisée, la mise à l'échelle dynamique peut s'adapter plus efficacement à la courbe de la demande de votre application. Pour de plus amples informations sur le choix des métriques pour le suivi de la cible, veuillez consulter Choisissez métriques.

Le problème lié à l'utilisation d'une métrique CloudWatch Amazon SQS, comme `ApproximateNumberOfMessagesVisible` pour le suivi des cibles, est que le nombre de messages dans la file d'attente peut ne pas changer proportionnellement à la taille du groupe Auto Scaling qui traite les messages de la file d'attente. En effet, le nombre de messages dans votre file d'attente SQS ne définit pas uniquement le nombre d'instances nécessaires. Le nombre d'instances du groupe Auto Scaling peut être dicté par plusieurs facteurs, y compris le temps nécessaire pour traiter un message et la durée de latence acceptable (délai de file d'attente).

La solution consiste à utiliser une métrique d'éléments en attente par instance avec la valeur cible constituant les éléments en attente acceptables par instance à conserver. Vous pouvez calculer ces valeurs comme suit :

- Éléments en attente par instance : pour calculer les éléments en attente par instance, commencez avec l'attribut de file d'attente `ApproximateNumberOfMessages` pour déterminer la longueur de la file d'attente SQS (nombre de messages disponibles dans cette file d'attente). Divisez ce nombre par la capacité d'exécution du parc, ce qui correspond dans le cas d'un groupe Auto Scaling au nombre d'instances dans l'état `InService`, pour obtenir les éléments en attente par instance.

- Éléments en attente acceptables par instance : pour calculer votre valeur cible, commencez par déterminer ce que votre application peut accepter en termes de latence. Prenez ensuite la valeur de latence acceptable et divisez-la par le temps moyen nécessaire à une instance EC2 pour traiter un message.

Par exemple, disons que vous avez actuellement un groupe Auto Scaling avec 10 instances et le nombre de messages visibles dans la file d'attente (`ApproximateNumberOfMessages`) s'élève à 1500. Si le temps de traitement moyen est de 0,1 seconde pour chaque message et si la plus grande latence acceptable est de 10 secondes, alors les éléments en attente acceptables par instance sont de 10/0,1, ce qui équivaut à 100 messages. Cela signifie que 100 est la valeur cible pour votre politique de suivi des objectifs et d'échelonnement. Lorsque les éléments en attente par instance atteignent la valeur cible, une diminution se produit. Si les éléments en attente par instance sont actuellement à 150 messages (1500 messages / 10 instances), votre groupe augmente de cinq instances pour maintenir la proportion de la valeur cible.

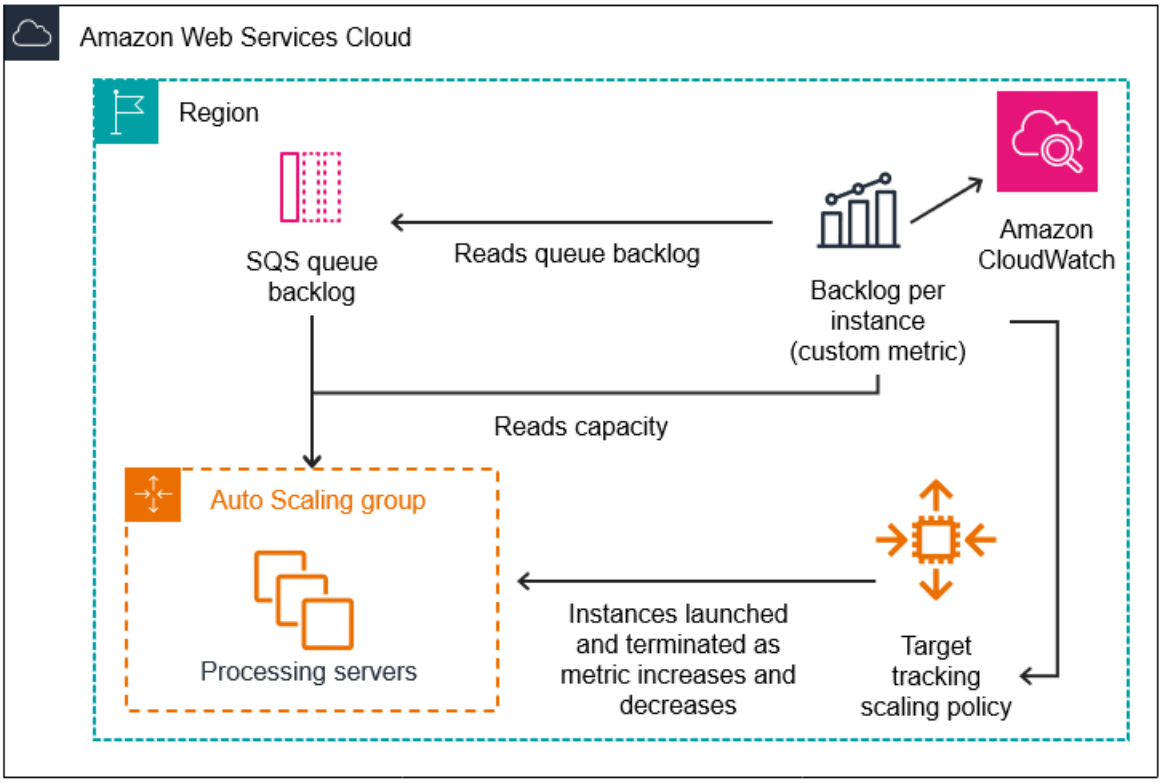
Les procédures suivantes montrent comment publier la métrique personnalisée et créer la politique de suivi des objectifs et d'échelonnement qui configure la mise à l'échelle du groupe Auto Scaling en fonction de ces calculs.

> **Important**
>
> N'oubliez pas que pour réduire les coûts, vous pouvez utiliser une expression mathématique appliquée à une métrique. Pour plus d'informations, consultez Créer une politique de mise à l'échelle du suivi des cibles pour Amazon EC2 Auto Scaling à l'aide d'une expression mathématique appliquée à une métrique.

Il existe trois parties principales pour cette configuration :

- Un groupe Auto Scaling pour gérer les instances EC2 afin de traiter les messages d'une file d'attente SQS.
- Une métrique personnalisée à envoyer à Amazon CloudWatch qui mesure le nombre de messages dans la file d'attente par instance EC2 du groupe Auto Scaling.
- Une politique de suivi des cibles qui configure votre groupe Auto Scaling pour qu'il évolue en fonction de la métrique personnalisée et d'une valeur cible définie. CloudWatch les alarmes invoquent la politique de dimensionnement.

Le graphique suivant illustre l'architecture de cette configuration.



Limitations et prérequis

Pour utiliser cette configuration, vous devez être conscient des limitations suivantes :

- Vous devez utiliser le AWS CLI ou un SDK pour publier votre métrique personnalisée sur. CloudWatch Vous pouvez ensuite surveiller votre métrique à l'aide du AWS Management Console.
- La console Amazon EC2 Auto Scaling ne prend pas en charge les politiques de suivi des objectifs et d'échelonnement utilisant des métriques personnalisés. Vous devez utiliser le AWS CLI ou un SDK pour spécifier une métrique personnalisée pour votre politique de dimensionnement.

Les sections suivantes vous indiquent comment utiliser le AWS CLI pour les tâches que vous devez effectuer. Par exemple, pour obtenir des données de métrique qui reflètent l'utilisation actuelle de la file d'attente, utilisez la commande SQS get-queue-attributes. Assurez-vous d'avoir installé et configuré l'interface de ligne de commande.

Avant de commencer, vous devez disposer d'une file d'attente Amazon SQS afin de l'utiliser. Dans les sections suivantes, il est supposé que vous disposez déjà d'une file d'attente (standard ou FIFO), d'un groupe Auto Scaling et d'instances EC2 exécutant l'application qui utilise la file d'attente. Pour

plus d'informations sur Amazon SQS, consultez le Guide du développeur Amazon Simple Queue Service.

Configuration de la mise à l'échelle en fonction d'Amazon SQS

Tâches

- Étape 1 : créer une métrique CloudWatch personnalisée
- Étape 2 : créer une politique de suivi des objectifs et d'échelonnement
- Étape 3 : tester votre politique de mise à l'échelle

Étape 1 : créer une métrique CloudWatch personnalisée

Une métrique personnalisée est définie au moyen d'un nom de métrique et d'un espace de noms de votre choix. Les espaces de noms pour les métriques personnalisées ne peuvent pas commencer par `AWS/`. Pour plus d'informations sur la publication de métriques personnalisées, consultez la rubrique Publier des métriques personnalisées dans le guide de CloudWatch l'utilisateur Amazon.

Suivez cette procédure pour créer la métrique personnalisée en lisant d'abord les informations de votre AWS compte. Calculez ensuite les éléments en attente par métrique d'instance, comme recommandé précédemment. Enfin, publiez ce numéro avec une granularité d'une minute. CloudWatch Nous vous recommandons vivement de mettre à l'échelle les métriques avec une granularité d'une minute afin de garantir une réponse plus rapide aux modifications de la charge du système, dans la mesure du possible.

Pour créer une métrique CloudWatch personnalisée (AWS CLI)

1. Utilisez la commande SQS get-queue-attributes pour obtenir le nombre de messages en attente dans la file d'attente (`ApproximateNumberOfMessages`).

```
aws sqs get-queue-attributes --queue-url https://
sqs.region.amazonaws.com/123456789/MyQueue \
  --attribute-names ApproximateNumberOfMessages
```

2. Utilisez la commande describe-auto-scaling-groups pour obtenir la capacité d'exécution du groupe, qui correspond au nombre d'instances dans l'état du cycle de vie `InService`. Cette commande renvoie les instances d'un groupe Auto Scaling, avec leur état de cycle de vie.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-names my-asg
```

3. Calculez les éléments en attente par instance en divisant le nombre approximatif de messages disponibles pour la récupération dans la file d'attente par la capacité d'exécution du groupe.
4. Créez un script qui s'exécute toutes les minutes pour récupérer le backlog par valeur d'instance et le publier dans une métrique CloudWatch personnalisée. Lorsque vous publiez une métrique personnalisée, vous spécifiez son nom, son espace de noms et aucune ou plusieurs dimensions. Une dimension se compose d'un nom de dimension et d'une valeur de dimension.

   Pour publier votre métrique personnalisée, remplacez les valeurs d'espace réservé en *italique* par le nom de métrique de votre choix, la valeur de la métrique, un espace de noms (à condition qu'il ne commence pas par « AWS ») et des dimensions (facultatif), puis exécutez la commande put-metric-data suivante.

   ```
   aws cloudwatch put-metric-data --metric-name MyBacklogPerInstance --
   namespace MyNamespace \
     --unit None --value 20 --
   dimensions MyOptionalMetricDimensionName=MyOptionalMetricDimensionValue
   ```

Une fois que votre application a émis la métrique souhaitée, les données sont envoyées à CloudWatch. La métrique est visible dans la CloudWatch console. Vous pouvez y accéder en vous connectant AWS Management Console et en accédant à la CloudWatch page. Consultez ensuite la métrique en accédant à la page des métriques ou en la recherchant à l'aide de la zone de recherche. Pour plus d'informations sur l'affichage des métriques, consultez la section Afficher les métriques disponibles dans le guide de CloudWatch l'utilisateur Amazon.

Étape 2 : créer une politique de suivi des objectifs et d'échelonnement

La métrique que vous avez créée peut désormais être ajoutée à une politique de suivi des cibles et de mise à l'échelle.

Pour créer une politique de suivi des cibles et de mise à l'échelle (AWS CLI)

1. Utilisez la commande `cat` suivante pour spécifier une valeur cible pour votre politique de mise à l'échelle et une spécification métrique personnalisée dans un fichier JSON appelé `config.json` dans votre répertoire de base. Remplacez chaque *espace réservé à la saisie de l'utilisateur* par vos propres informations. Pour `TargetValue`, calculez la métrique des éléments en attente acceptables par instance et saisissez la valeur ici. Pour calculer cette valeur, décidez d'une valeur de latence normale et divisez-la par la durée moyenne nécessaire au traitement d'un message, comme décrit dans une précédente section.

Si vous n'avez spécifié aucune dimension pour la métrique que vous avez créée à l'étape 1, n'incluez aucune dimension dans la spécification de métrique personnalisée.

```
$ cat ~/config.json
{
   "TargetValue":100,
   "CustomizedMetricSpecification":{
      "MetricName":"MyBacklogPerInstance",
      "Namespace":"MyNamespace",
      "Dimensions":[
         {
            "Name":"MyOptionalMetricDimensionName",
            "Value":"MyOptionalMetricDimensionValue"
         }
      ],
      "Statistic":"Average",
      "Unit":"None"
   }
}
```

2. Utilisez la commande put-scaling-policy ainsi que le fichier `config.json` créé à l'étape précédente pour élaborer la politique de mise à l'échelle.

```
aws autoscaling put-scaling-policy --policy-name sqs100-target-tracking-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://~/config.json
```

Cela crée deux alarmes, une pour augmenter et une pour réduire la taille des instances, Il renvoie également l'Amazon Resource Name (ARN) de la politique enregistrée CloudWatch, qui est CloudWatch utilisée pour invoquer le dimensionnement chaque fois que le seuil métrique est dépassé.

Étape 3 : tester votre politique de mise à l'échelle

Une fois votre configuration terminée, vérifiez que votre politique de mise à l'échelle fonctionne. Vous pouvez tester la politique de montée en puissance en augmentant le nombre de messages dans la file d'attente SQS, puis vérifier que le groupe Auto Scaling a lancé une instance EC2 supplémentaire. De la même façon, vous pouvez tester la politique de mise à l'échelle horizontale en réduisant le

nombre de messages dans la file d'attente SQS, puis vérifier que le groupe Auto Scaling a résilié l'instance EC2.

Pour tester la fonction d'évolutivité horizontale

1. Suivez les étapes décrites dans Création d'une file d'attente standard Amazon SQS et envoi d'un message ou Création d'une file d'attente FIFO Amazon SQS et envoi d'un message pour ajouter des messages à votre file d'attente. Assurez-vous que vous avez augmenté le nombre de messages dans la file d'attente, afin que la métrique d'éléments en attente par instance dépasse la valeur cible.

   Il peut s'écouler quelques minutes avant que les modifications n'appellent l'alarme.
2. Utilisez la commande describe-auto-scaling-groups pour vérifier si le groupe a lancé une instance.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Pour tester la fonction de mise à l'échelle horizontale

1. Suivez les étapes décrites dans la section Recevoir et supprimer un message (console) pour supprimer des messages de la file d'attente. Assurez-vous que vous avez réduit le nombre de messages dans la file d'attente, afin que la métrique d'éléments en attente par instance soit inférieure à la valeur cible.

   Il peut s'écouler quelques minutes avant que les modifications n'appellent l'alarme.
2. Utilisez la commande describe-auto-scaling-groups suivante pour vérifier si le groupe a résilié une instance.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Protection contre la mise à l'échelle horizontale d'instance et Amazon SQS

Les messages qui n'ont pas été traités au moment de la résiliation d'une instance sont renvoyés à la file d'attente SQS où ils peuvent être traités par une autre instance qui est encore en cours d'exécution. Pour les applications dans lesquelles des tâches longues sont exécutées, vous pouvez éventuellement utiliser la protection contre la mise à l'échelle horizontale d'instance pour contrôler

les processus Worker de file d'attente qui sont résiliés lorsque votre groupe Auto Scaling est mis à l'échelle.

Le pseudocode suivant illustre une façon de protéger les processus Worker de longue durée pilotés par la file d'attente contre la résiliation de mise à l'échelle horizontale.

```
while (true)
{
  SetInstanceProtection(False);
  Work = GetNextWorkUnit();
  SetInstanceProtection(True);
  ProcessWorkUnit(Work);
  SetInstanceProtection(False);
}
```

Pour plus d'informations, consultez Concevez vos applications sur Amazon EC2 Auto Scaling pour gérer de manière optimale la résiliation des instances.

Vérifier une activité de mise à l'échelle pour un groupe Auto Scaling

Dans la section Amazon EC2 Auto Scaling de la console Amazon EC2, l'Activity history (Historique de l'activité) pour un groupe Auto Scaling vous permet de visualiser l'état actuel d'une activité de mise à l'échelle qui est en cours. Lorsque l'activité de mise à l'échelle est terminée, vous pouvez voir si elle a réussi ou non. Ceci est particulièrement utile lorsque vous créez des groupes Auto Scaling ou que vous ajoutez des conditions de mise à l'échelle à des groupes existants.

Lorsque vous ajoutez une politique de suivi de cible, d'étape ou de mise à l'échelle simple à votre groupe Auto Scaling, Amazon EC2 Auto Scaling commence immédiatement à évaluer la politique par rapport à la métrique. L'alarme de la métrique passe à l'état ALARM quand la métrique dépasse le seuil pendant un certain nombre de périodes d'évaluation. Cela signifie qu'une politique de mise à l'échelle peut entraîner une activité de mise à l'échelle peu de temps après sa création. Après qu'Amazon EC2 Auto Scaling ait ajusté la capacité souhaitée en réponse à une politique de mise à l'échelle, vous pouvez vérifier l'activité de mise à l'échelle dans votre compte. Si vous souhaitez recevoir une notification par e-mail d'Amazon EC2 Auto Scaling vous informant d'une activité de mise à l'échelle, suivez les instructions de la section Options de notification Amazon SNS pour Amazon EC2 Auto Scaling.

> **Tip**
>
> Dans la procédure suivante, vous consultez les sections Activity history (Historique des activités), et Instances pour le groupe Auto Scaling. Dans les deux sections, les colonnes nommées doivent déjà être affichées. Pour afficher les colonnes masquées ou modifier le nombre de lignes affichées, cliquez sur l'icône en forme de roue dentée dans le coin supérieur droit de chaque section pour ouvrir les préférences modeales, mettez à jour les paramètres au besoin et cliquez sur Confirm (Confirmer).

Pour afficher les activités de mise à l'échelle de votre groupe Auto Scaling (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Dans la barre de navigation située en haut de l'écran, choisissez la région dans laquelle vous avez créé votre groupe Auto Scaling.
3. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
4. Sous l'onglet Activité sous Historique de l'activité, la colonne État indique si votre groupe Auto Scaling a réussi à lancer ou à résilier des instances, ou si l'activité de mise à l'échelle est toujours en cours.
5. (Facultatif) Si vous avez beaucoup d'activités de mise à l'échelle, vous pouvez choisir l'icône **>** sur le bord supérieur de l'historique de l'activité pour voir la page suivante des activités de mise à l'échelle.
6. Sous l'onglet Instance management (Gestion des instances) dans Instances, la colonne Lifecycle (Cycle de vie) contient l'état de vos instances. Une fois l'instance démarrée et tous les hooks de cycle de vie terminés, son cycle de vie passe à l'état `InService`. La colonne Health Status (État d'intégrité) affiche le résultat des surveillances de l'état de l'instance EC2 sur votre instance.

Pour afficher les activités de mise à l'échelle d'un groupe Auto Scaling (AWS CLI)

Utilisez la commande describe-scaling-activités suivante.

```
aws autoscaling describe-scaling-activities --auto-scaling-group-name my-asg
```

Voici un exemple de sortie.

Les activités de mise à l'échelle sont classées par heure de début. Les activités toujours en cours sont décrites en premier lieu.

```
{
  "Activities": [
    {
      "ActivityId": "5e3a1f47-2309-415c-bfd8-35aa06300799",
      "AutoScalingGroupName": "my-asg",
      "Description": "Terminating EC2 instance: i-06c4794c2499af1df",
      "Cause": "At 2020-02-11T18:34:10Z a monitor alarm TargetTracking-my-asg-AlarmLow-
b9376cab-18a7-4385-920c-dfa3f7783f82 in state ALARM triggered policy my-target-
tracking-policy changing the desired capacity from 3 to 2.  At 2020-02-11T18:34:31Z
 an instance was taken out of service in response to a difference between desired and
 actual capacity, shrinking the capacity from 3 to 2.  At 2020-02-11T18:34:31Z instance
 i-06c4794c2499af1df was selected for termination.",
      "StartTime": "2020-02-11T18:34:31.268Z",
      "EndTime": "2020-02-11T18:34:53Z",
      "StatusCode": "Successful",
      "Progress": 100,
      "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":\"us-west-2a
\"...}",
      "AutoScalingGroupARN": "arn"
    },
...
  ]
}
```

Pour obtenir une description des champs de la sortie, consultez Activité dans la Référence de l'API Amazon EC2 Auto Scaling.

Pour obtenir de l’aide sur la récupération des activités de mise à l’échelle d’un groupe supprimé et pour obtenir des informations sur les types d’erreurs que vous pouvez rencontrer et sur la façon de les traiter, consultez Résoudre les problèmes d'Amazon EC2 Auto Scaling.

Désactiver une politique de mise à l'échelle pour un groupe Auto Scaling

Cette rubrique décrit comment désactiver temporairement une politique de mise à l'échelle afin qu'elle ne modifie pas le nombre d'instances que contient le groupe Auto Scaling. Lorsque vous désactivez une politique de mise à l'échelle, les détails de configuration sont conservés, de sorte que vous pouvez réactiver rapidement la politique. Cela est plus facile que de supprimer temporairement une politique lorsque vous n'en avez pas besoin et de la recréer ultérieurement.

Lorsqu'une politique de mise à l'échelle est désactivée, le groupe Auto Scaling ne monte pas ou ne diminue pas en charge en fonction des alarmes de métrique qui sont enfreintes pendant que la politique de mise à l'échelle est désactivée. Toutefois, les activités de mise à l'échelle qui sont encore en cours ne sont pas arrêtées.

Notez que les politiques de mise à l'échelle désactivées comptent toujours dans vos quotas sur le nombre de politiques de mise à l'échelle que vous pouvez ajouter à un groupe Auto Scaling.

Pour désactiver une politique de mise à l'échelle (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Dans l'onglet Mise à l'échelle automatique, sous Politiques de mise à l'échelle dynamique, cochez la case dans le coin supérieur droit de la politique de mise à l'échelle souhaitée.
4. Faites défiler jusqu'en haut de la section Dynamic scaling policies (Politiques de mise à l'échelle dynamique), puis choisissez Actions, Disable (Désactiver).

Lorsque vous êtes prêt à réactiver la politique de mise à l'échelle, répétez ces étapes, puis choisissez Actions, Enable (Activer). Après avoir réactivé une politique de mise à l'échelle, votre groupe Auto Scaling peut immédiatement lancer une action de mise à l'échelle si des alarmes sont actuellement dans l'état ALARM.

Pour désactiver une politique de mise à l'échelle (AWS CLI)

Utilisez la commande put-scaling-policy avec l'option `--no-enabled` comme suit. Spécifiez toutes les options de la commande comme vous le feriez lors de la création de la politique.

```
aws autoscaling put-scaling-policy --auto-scaling-group-name my-asg \
   --policy-name my-scaling-policy --policy-type TargetTrackingScaling \
   --estimated-instance-warmup 360 \
   --target-tracking-configuration '{ "TargetValue": 70,
 "PredefinedMetricSpecification": { "PredefinedMetricType":
 "ASGAverageCPUUtilization" } }' \
   --no-enabled
```

Pour réactiver une politique de mise à l'échelle (AWS CLI)

Utilisez la commande put-scaling-policy avec l'option `--enabled` comme suit. Spécifiez toutes les options de la commande comme vous le feriez lors de la création de la politique.

```
aws autoscaling put-scaling-policy --auto-scaling-group-name my-asg \
   --policy-name my-scaling-policy --policy-type TargetTrackingScaling \
   --estimated-instance-warmup 360 \
   --target-tracking-configuration '{ "TargetValue": 70,
 "PredefinedMetricSpecification": { "PredefinedMetricType":
 "ASGAverageCPUUtilization" } }' \
   --enabled
```

Pour décrire une politique de mise à l'échelle (AWS CLI)

Utilisez la commande describe-policies pour vérifier l'état activé d'une politique de mise à l'échelle.

```
aws autoscaling describe-policies --auto-scaling-group-name my-asg \
   --policy-names my-scaling-policy
```

Voici un exemple de sortie.

```
{
    "ScalingPolicies": [
        {
            "AutoScalingGroupName": "my-asg",
            "PolicyName": "my-scaling-policy",
            "PolicyARN": "arn:aws:autoscaling:us-
west-2:123456789012:scalingPolicy:1d52783a-b03b-4710-
bb0e-549fd64378cc:autoScalingGroupName/my-asg:policyName/my-scaling-policy",
            "PolicyType": "TargetTrackingScaling",
            "StepAdjustments": [],
            "Alarms": [
                {
                    "AlarmName": "TargetTracking-my-asg-
AlarmHigh-9ca53fdd-7cf5-4223-938a-ae1199204502",
                    "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmHigh-9ca53fdd-7cf5-4223-938a-
ae1199204502"
                },
                {
                    "AlarmName": "TargetTracking-my-asg-AlarmLow-7010c83d-d55a-4a7a-
abe0-1cf8b9de6d6c",
```

```
                    "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-7010c83d-d55a-4a7a-
abe0-1cf8b9de6d6c"
                }
            ],
            "TargetTrackingConfiguration": {
                "PredefinedMetricSpecification": {
                    "PredefinedMetricType": "ASGAverageCPUUtilization"
                },
                "TargetValue": 70.0,
                "DisableScaleIn": false
            },
            "Enabled": true
        }
    ]
}
```

Suppression d'une stratégie de mise à l'échelle

Lorsque vous n'avez plus besoin d'une politique de mise à l'échelle, vous pouvez la supprimer. Selon le type de politique de dimensionnement, vous devrez peut-être également supprimer les CloudWatch alarmes. La suppression d'une politique de dimensionnement du suivi des cibles supprime également toutes les CloudWatch alarmes associées. La suppression d'une politique de dimensionnement par étapes ou d'une simple politique de dimensionnement supprime l'action d'alarme sous-jacente, mais elle ne supprime pas l' CloudWatch alarme, même si aucune action n'y est plus associée.

Pour supprimer une politique de mise à l'échelle (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s’ouvre en bas de la page.
3. Dans l’onglet Mise à l’échelle automatique, sous Politiques de mise à l’échelle dynamique, cochez la case dans le coin supérieur droit de la politique de mise à l’échelle souhaitée.
4. Faites défiler jusqu'en haut de la section Dynamic scaling policies (Politiques de mise à l'échelle dynamique), puis choisissez Actions, Delete (Supprimer).
5. Lorsque vous êtes invité à confirmer l’opération, choisissez Supprimer.

6. (Facultatif) Si vous avez supprimé une politique de dimensionnement par étapes ou une simple politique de dimensionnement, procédez comme suit pour supprimer l' CloudWatch alarme associée à la politique. Vous pouvez ignorer ces sous-étapes pour conserver l'alarme en vue d'une utilisation ultérieure.

   a. Ouvrez la CloudWatch console à l'adresse https://console.aws.amazon.com/cloudwatch/.

   b. Dans le panneau de navigation, choisissez Alarmes.

   c. Choisissez l'alarme (par exemple, `Step-Scaling-AlarmHigh-AddCapacity`), puis choisissez Action, Delete (Supprimer).

   d. Lorsque vous êtes invité à confirmer l'opération, choisissez Supprimer.

Pour obtenir les politiques de mise à l'échelle d'un groupe Auto Scaling (AWS CLI)

Avant de supprimer une politique de mise à l'échelle, utilisez la commande describe-policies suivante pour voir quelles politiques de mise à l'échelle ont été créées pour le groupe Auto Scaling. Vous pouvez utiliser la sortie lors de la suppression de la politique et des CloudWatch alarmes.

```
aws autoscaling describe-policies --auto-scaling-group-name my-asg
```

Vous pouvez filtrer les résultats par le type de politique de mise à l'échelle à l'aide du paramètre `--query`. Cette syntaxe pour `query` fonctionne sur Linux ou macOS. Sous Windows, remplacez les guillemets simples par des guillemets doubles.

```
aws autoscaling describe-policies --auto-scaling-group-name my-asg
  --query 'ScalingPolicies[?PolicyType==`TargetTrackingScaling`]'
```

Voici un exemple de sortie.

```
[
    {
        "AutoScalingGroupName": "my-asg",
        "PolicyName": "cpu50-target-tracking-scaling-policy",
        "PolicyARN": "PolicyARN",
        "PolicyType": "TargetTrackingScaling",
        "StepAdjustments": [],
        "Alarms": [
            {
```

```
                "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e",
                "AlarmName": "TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e"
            },
            {
                "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2",
                "AlarmName": "TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2"
            }
        ],
        "TargetTrackingConfiguration": {
            "PredefinedMetricSpecification": {
                "PredefinedMetricType": "ASGAverageCPUUtilization"
            },
            "TargetValue": 50.0,
            "DisableScaleIn": false
        },
        "Enabled": true
    }
]
```

Pour supprimer votre politique de mise à l'échelle (AWS CLI)

Utilisez la commande delete-policy suivante.

```
aws autoscaling delete-policy --auto-scaling-group-name my-asg \
  --policy-name cpu50-target-tracking-scaling-policy
```

Pour supprimer votre CloudWatch alarme (AWS CLI)

Pour les politiques de dimensionnement par étapes et simples, utilisez la commande delete-alarm pour supprimer l' CloudWatch alarme associée à la politique. Vous pouvez ignorer cette étape pour conserver l'alarme pour une utilisation ultérieure. Vous pouvez supprimer une ou plusieurs alarmes en même temps. Par exemple, utilisez la commande suivante pour supprimer les alarmes `Step-Scaling-AlarmHigh-AddCapacity` et `Step-Scaling-AlarmLow-RemoveCapacity`.

```
aws cloudwatch delete-alarms --alarm-name Step-Scaling-AlarmHigh-AddCapacity Step-
Scaling-AlarmLow-RemoveCapacity
```

Exemple de politiques de mise à l'échelle pour AWS Command Line Interface (AWS CLI)

Vous pouvez créer des politiques de dimensionnement pour Amazon EC2 Auto Scaling via le ou AWS Management Console AWS CLI les SDK.

Les exemples suivants montrent comment créer des politiques de dimensionnement pour Amazon EC2 Auto Scaling à l'aide de la commande AWS CLI put-scaling-policy. Remplacez chaque *espace réservé à la saisie de l'utilisateur* par vos propres informations.

Pour commencer à rédiger des politiques de dimensionnement à l'aide du AWS CLI, consultez les exercices d'introduction dans Politiques de suivi des objectifs de la mise à l'échelle etPolitiques de mise à l'échelle simple et par étapes.

Exemple 1 : pour appliquer une politique de suivi des objectifs et d'échelonnement avec une spécification de métrique prédéfinie

```
aws autoscaling put-scaling-policy --policy-name cpu50-target-tracking-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
{
  "TargetValue": 50.0,
  "PredefinedMetricSpecification": {
    "PredefinedMetricType": "ASGAverageCPUUtilization"
  }
}
```

Pour plus d'informations, consultez la section PredefinedMetricSpécification dans le manuel Amazon EC2 Auto Scaling API Reference.

> **Note**
>
> Si le fichier ne se trouve pas dans le répertoire actuel, saisissez le chemin complet du fichier. Pour plus d'informations sur la lecture de valeurs de AWS CLI paramètres depuis un fichier, consultez la section Chargement de AWS CLI paramètres depuis un fichier dans le Guide de AWS Command Line Interface l'utilisateur.

Exemple 2 : pour appliquer une politique de suivi des objectifs et d'échelonnement avec une spécification de métrique personnalisée

```
aws autoscaling put-scaling-policy --policy-name sqs100-target-tracking-scaling-policy
 \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
{
  "TargetValue": 100.0,
  "CustomizedMetricSpecification": {
    "MetricName": "MyBacklogPerInstance",
    "Namespace": "MyNamespace",
    "Dimensions": [{
      "Name": "MyOptionalMetricDimensionName",
      "Value": "MyOptionalMetricDimensionValue"
    }],
    "Statistic": "Average",
    "Unit": "None"
  }
}
```

Pour plus d'informations, consultez la section CustomizedMetricSpécification dans le manuel Amazon EC2 Auto Scaling API Reference.

Exemple 3 : pour appliquer une politique de suivi des objectifs et d'échelonnement uniquement en vue d'une évolutivité horizontale

```
aws autoscaling put-scaling-policy --policy-name alb1000-target-tracking-scaling-policy
 \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
{
  "TargetValue": 1000.0,
  "PredefinedMetricSpecification": {
    "PredefinedMetricType": "ALBRequestCountPerTarget",
    "ResourceLabel": "app/my-alb/778d41231b141a0f/targetgroup/my-alb-target-
group/943f017f100becff"
  },
  "DisableScaleIn": true
}
```

Exemple 4 : pour appliquer une politique de mise à l'échelle par étapes pour en vue d'une évolutivité horizontale

```
aws autoscaling put-scaling-policy \
```

```
  --auto-scaling-group-name my-asg  \
  --policy-name my-step-scale-out-policy \
  --policy-type StepScaling \
  --adjustment-type PercentChangeInCapacity \
  --metric-aggregation-type Average \
  --step-adjustments
 MetricIntervalLowerBound=10.0,MetricIntervalUpperBound=20.0,ScalingAdjustment=10 \

 MetricIntervalLowerBound=20.0,MetricIntervalUpperBound=30.0,ScalingAdjustment=20 \
                     MetricIntervalLowerBound=30.0,ScalingAdjustment=30 \
  --min-adjustment-magnitude 1
```

Notez le nom Amazon Resource Name (ARN) de la politique. Vous avez besoin de l'ARN lorsque vous créez l' CloudWatch alarme.

Exemple 5 : pour appliquer une politique de mise à l'échelle par étapes en vue d'une mise à l'échelle horizontale

```
aws autoscaling put-scaling-policy \
  --auto-scaling-group-name my-asg  \
  --policy-name my-step-scale-in-policy \
  --policy-type StepScaling \
  --adjustment-type ChangeInCapacity \
  --step-adjustments MetricIntervalUpperBound=0.0,ScalingAdjustment=-2
```

Notez le nom Amazon Resource Name (ARN) de la politique. Vous avez besoin de l'ARN lorsque vous créez l' CloudWatch alarme.

Exemple 6 : pour appliquer une politique de mise à l'échelle simple en vue d'une évolutivité horizontale

```
aws autoscaling put-scaling-policy --policy-name my-simple-scale-out-policy \
  --auto-scaling-group-name my-asg --scaling-adjustment 30 \
  --adjustment-type PercentChangeInCapacity --min-adjustment-magnitude 2
```

Notez le nom Amazon Resource Name (ARN) de la politique. Vous avez besoin de l'ARN lorsque vous créez l' CloudWatch alarme.

Exemple 7 : pour appliquer une politique de mise à l'échelle simple en vue de la mise à l'échelle horizontale

```
aws autoscaling put-scaling-policy --policy-name my-simple-scale-in-policy \
```

```
  --auto-scaling-group-name my-asg --scaling-adjustment -1 \
  --adjustment-type ChangeInCapacity --cooldown 180
```

Notez le nom Amazon Resource Name (ARN) de la politique. Vous avez besoin de l'ARN lorsque vous créez l' CloudWatch alarme.

Mise à l'échelle prédictive pour Amazon EC2 Auto Scaling

La mise à l'échelle prédictive fonctionne en analysant les données de charge historiques afin de détecter les tendances quotidiennes ou hebdomadaires des flux de trafic. Il utilise ces informations pour prévoir les besoins de capacité futurs afin qu'Amazon EC2 Auto Scaling puisse augmenter de manière proactive la capacité de votre groupe Auto Scaling pour qu'il corresponde à la charge prévue.

La mise à l'échelle prédictive se prête particulièrement bien aux situations suivantes :

- Trafic cyclique, tel qu'une utilisation intensive des ressources pendant les heures de bureau et une faible utilisation le soir et le week-end
- Modèles on-and-off de charge de travail récurrents, tels que le traitement par lots, les tests ou l'analyse périodique des données
- Applications dont l'initialisation prend beaucoup de temps, ce qui en termes de performances se traduit par une latence notable lors des événements de montée en puissance

En règle générale, si vous avez des pics de trafic réguliers et des applications dont l'initialisation prend beaucoup de temps, n'hésitez pas à utiliser la mise à l'échelle prédictive. La mise à l'échelle prédictive peut vous permettre de vous adapter plus rapidement en fournissant de la capacité avant l'arrivée de la charge prévue, par opposition à la mise à l'échelle dynamique, qui est réactive par nature. La mise à l'échelle prédictive peut également vous faire économiser de l'argent sur votre facture EC2 en vous évitant de devoir surapprovisionner la capacité.

Prenons l'exemple d'une application très utilisée pendant les heures de bureau et peu utilisée la nuit. Au début de chaque jour ouvrable, la mise à l'échelle prédictive peut augmenter la capacité avant le premier afflux de trafic. Cela permet à votre application de maintenir une disponibilité et des performances élevées lorsqu'elle passe d'une période de faible utilisation à une période d'utilisation plus intensive. Vous n'avez pas besoin d'attendre que la mise à l'échelle dynamique réagisse à l'évolution du trafic. Vous n'avez pas non plus à examiner les tendances de charge de votre application et à essayer de planifier la capacité adéquate à l'aide d'une mise à l'échelle planifiée.

Rubriques

Fonctionnement de la mise à l'échelle prédictive

Cette rubrique explique le fonctionnement du dimensionnement prédictif et décrit les éléments à prendre en compte lors de la création d'une politique de dimensionnement prédictif.

Rubriques

Comment ça marche

Pour utiliser la mise à l'échelle prédictive, créez une politique de mise à l'échelle prédictive qui spécifie la CloudWatch métrique à surveiller et à analyser. Pour que la mise à l'échelle prédictive commence à prévoir les valeurs futures, cette métrique doit disposer d'au moins 24 heures de données.

Une fois que vous avez créé la politique, le dimensionnement prédictif commence à analyser les données métriques des 14 derniers jours afin d'identifier des modèles. Il utilise cette analyse pour générer une prévision horaire des besoins de capacité pour les 48 prochaines heures. Les prévisions sont mises à jour toutes les 6 heures en utilisant les dernières CloudWatch données. À mesure que de nouvelles données arrivent, la mise à l'échelle prédictive est en mesure d'améliorer en permanence la précision des prévisions futures.

Lorsque vous activez la mise à l'échelle prédictive pour la première fois, elle s'exécute en mode prévision uniquement. Dans ce mode, il génère des prévisions de capacité mais ne redimensionne pas réellement votre groupe Auto Scaling en fonction de ces prévisions. Cela vous permet d'évaluer

la précision et la pertinence des prévisions. Vous pouvez consulter les données de prévision à l'aide de l'opération `GetPredictiveScalingForecast` API ou du AWS Management Console.

Après avoir examiné les données de prévision et décidé de commencer le dimensionnement en fonction de ces données, passez la politique de dimensionnement en mode prévision et échelle. Dans ce mode :

- Si les prévisions prévoient une augmentation de la charge, Amazon EC2 Auto Scaling augmentera la capacité en la redimensionnant.
- Si les prévisions prévoient une diminution de la charge, elles ne seront pas ajustées pour réduire la capacité. Si vous souhaitez supprimer une capacité qui n'est plus nécessaire, vous devez créer des politiques de dimensionnement dynamiques.

Par défaut, Amazon EC2 Auto Scaling redimensionne votre groupe Auto Scaling au début de chaque heure en fonction des prévisions pour cette heure. Vous pouvez éventuellement spécifier une heure de début antérieure en utilisant la `SchedulingBufferTime` propriété dans l'opération d'`PutScalingPolicy`API ou le paramètre Instances de pré-lancement dans le AWS Management Console. Amazon EC2 Auto Scaling lance donc de nouvelles instances avant la demande prévue, ce qui leur laisse le temps de démarrer et de se préparer à gérer le trafic.

Pour permettre le lancement de nouvelles instances avant la demande prévue, nous vous recommandons vivement d'activer le préchauffage des instances par défaut pour votre groupe Auto Scaling. Cela indique une période après une activité de scale-out pendant laquelle Amazon EC2 Auto Scaling n'interviendra pas, même si les politiques de dimensionnement dynamique indiquent que la capacité doit être réduite. Cela vous permet de vous assurer que les instances nouvellement lancées disposent de suffisamment de temps pour commencer à traiter le trafic accru avant d'être prises en compte pour des opérations d'extension. Pour plus d’informations, consultez Définir la préparation par défaut d'instance d'un groupe Auto Scaling.

Limite de capacité maximale

Les groupes Auto Scaling ont un paramètre de capacité maximale qui limite le nombre maximum d'instances EC2 pouvant être lancées pour le groupe. Par défaut, lorsque des politiques de dimensionnement sont définies, elles ne peuvent pas augmenter la capacité au-delà de sa capacité maximale.

Vous pouvez également autoriser l'augmentation automatique de la capacité maximale du groupe si la capacité prévue approche ou dépasse la capacité maximale du groupe Auto Scaling. Pour activer

ce comportement, utilisez les `MaxCapacityBuffer` propriétés `MaxCapacityBreachBehavior` et dans le fonctionnement de l'`PutScalingPolicy`API ou le paramètre de comportement de capacité maximale dans le AWS Management Console.

> **Warning**
>
> Soyez prudent lorsque vous autorisez l'augmentation automatique de la capacité maximale. Cela peut entraîner le lancement d'un plus grand nombre d'instances que prévu si l'augmentation de la capacité maximale n'est pas surveillée et gérée. La capacité maximale accrue devient alors la nouvelle capacité maximale normale pour le groupe Auto Scaling jusqu'à ce que vous la mettiez à jour manuellement. La capacité maximale ne diminue pas automatiquement pour revenir à la capacité maximale initiale.

Considérations

- Vérifiez que la mise à l'échelle prédictive se prête à votre charge de travail. Une charge de travail se prête bien à la mise à l'échelle prédictive si elle présente des tendances de charge récurrentes spécifiques au jour de la semaine ou à l'heure de la journée. Pour le vérifier, configurez des politiques de mise à l'échelle prédictive en mode prévision uniquement, puis consultez les recommandations dans la console. Amazon EC2 Auto Scaling fournit des recommandations sur la base d'observations relatives aux performances potentielles des politiques. Évaluez la prévision et les recommandations avant de laisser la mise à l'échelle prédictive adapter activement votre application.
- La mise à l'échelle prédictive requiert au moins 24 heures de données historiques pour commencer à élaborer des prévisions. Toutefois, les prévisions sont plus efficaces si les données historiques couvrent deux semaines complètes. Si vous mettez à jour votre application en créant un nouveau groupe Auto Scaling et en supprimant l'ancien, le nouveau groupe doit disposer de 24 heures de données de charge historiques avant que la mise à l'échelle prédictive puisse recommencer à générer des prévisions. Vous pouvez utiliser des métriques personnalisées pour agréger des métriques entre les anciens et les nouveaux groupes Auto Scaling. Autrement, vous devrez peut-être attendre quelques jours pour obtenir une prévision plus précise.
- Choisissez une métrique de charge qui représente avec précision la charge complète de votre application et qui constitue l'aspect le plus important de votre application à prendre en compte.
- L'utilisation de la mise à l'échelle dynamique associée à la mise à l'échelle prédictive vous permet de suivre de près la courbe de demande de votre application, en augmentant l'échelle pendant les périodes de faible trafic et en la redimensionnant lorsque le trafic est plus élevé que prévu. Lorsque

plusieurs stratégies de mise à l'échelle sont actives, chaque stratégie détermine indépendamment la capacité souhaitée, et la capacité souhaitée est définie sur le maximum de celles-ci. Par exemple, si 10 instances sont requises pour respecter l'objectif d'utilisation dans une stratégie de suivi des objectifs et de mise à l'échelle, et que 8 instances sont requises pour respecter l'objectif d'utilisation dans une stratégie de mise à l'échelle prédictive, la capacité souhaitée du groupe est définie sur 10. Si vous débutez dans le domaine de la mise à l'échelle dynamique, nous vous recommandons d'utiliser des politiques de dimensionnement pour le suivi des cibles. Pour plus d'informations, consultez Mise à l'échelle dynamique pour Amazon EC2 Auto Scaling.

- Une hypothèse fondamentale de la mise à l'échelle prédictive est que le groupe Auto Scaling est homogène et que toutes les instances ont une capacité égale. Si ce n'est pas le cas pour votre groupe, la capacité prévue peut être inexacte. Par conséquent, soyez prudent lorsque vous créez des politiques de dimensionnement prédictif pour des groupes d'instances mixtes, car des instances de différents types peuvent être provisionnées avec une capacité inégale. Voici quelques exemples de cas où la capacité prévue sera inexacte :
  - Votre politique de mise à l'échelle prédictive est basée sur l'utilisation du processeur, mais le nombre de vCPUs sur chaque instance Auto Scaling varie d'un type d'instance à l'autre.
  - Votre politique de mise à l'échelle prédictive est basée sur l'entrée réseau ou la sortie réseau, mais le débit de bande passante réseau pour chaque instance Auto Scaling varie d'un type d'instance à l'autre. Par exemple, les types d'instances M5 et M5n sont similaires, mais le type d'instance M5n offre un débit réseau nettement supérieur.

Régions prises en charge

Amazon EC2 Auto Scaling prend en charge les politiques de dimensionnement prédictif dans les pays suivants Régions AWS : USA Est (Virginie du Nord), USA Est (Ohio), USA Ouest (Oregon), USA Ouest (Californie du Nord), Afrique (Le Cap), Canada (Centre), UE (Francfort), UE (Hong Kong), Asie-Pacifique (Jakarta), Asie-Pacifique (Mumbai), Asie-Pacifique (Osaka), Asie-Pacifique (Tokyo), Asie-Pacifique (Singapour), Asie-Pacifique (Séoul), Asie-Pacifique (Sydney), Moyen-Orient (Bahreïn), Moyen-Orient (Émirats arabes unis), Amérique du Sud (Sao Paulo), Chine (Pékin), Chine (Ningxia), AWS GovCloud (États-Unis de l'Est) et AWS GovCloud (États-Unis de l'Ouest).

Création d'une politique de dimensionnement prédictive

Les procédures suivantes vous aident à créer une politique de dimensionnement prédictive à l'aide du AWS Management Console ou AWS CLI.

Si le groupe Auto Scaling est nouveau, il doit fournir au moins 24 heures de données avant qu'Amazon EC2 Auto Scaling puisse générer une prévision.

Table des matières

Créer une politique de mise à l'échelle prédictive (console)

Si c'est la première fois que vous créez une politique de dimensionnement prédictif, nous vous recommandons d'utiliser la console pour créer plusieurs politiques de dimensionnement prédictif en mode prévision uniquement. Cela vous permet de tester les effets potentiels de différentes mesures et valeurs cibles. Vous pouvez créer plusieurs politiques de mise à l'échelle prédictive pour chaque groupe Auto Scaling, mais une seule de ces politiques peut être utilisée pour la mise à l'échelle active.

Créer une politique de mise à l'échelle prédictive dans la console (métriques prédéfinies)

Utilisez la procédure suivante pour créer une politique de mise à l'échelle prédictive à l'aide de métriques prédéfinies (CPU, I/O réseau ou nombre de requêtes Application Load Balancer par cible). La manière la plus simple de créer une politique de mise à l'échelle prédictive consiste à utiliser des métriques prédéfinies. Si vous préférez utiliser des métriques personnalisées, consultez Créer une politique de mise à l'échelle prédictive dans la console (métriques personnalisées).

Pour créer une politique de mise à l'échelle prédictive

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Sous l'onglet Mise à l'échelle automatique, accédez au champ Stratégies de mise à l'échelle et choisissez Créer une stratégie de mise à l'échelle prédictive.
4. Entrez le nom de la politique.
5. Activez Mise à l'échelle basée sur la prévision pour autoriser Amazon EC2 Auto Scaling à lancer immédiatement la mise à l'échelle.

Pour que la stratégie reste en mode prévision uniquement, n'activez pas Mise à l'échelle basée sur la prévision.

6. Sous Métriques, choisissez vos métriques dans la liste des options. Les options incluent Processeur, Entrée réseau, Sortie réseau, Nombre de demandes Application Load Balancer et Paire de métriques personnalisées.

   Si vous avez choisi Nombre de demandes Application Load Balancer par cible, choisissez un groupe cible dans le champ Groupe cible. L'option Nombre de demandes Application Load Balancer par cible n'est prise en charge que si vous avez attaché un groupe cible Application Load Balancer à votre groupe Auto Scaling.

   Si vous avez choisi Paire de métriques personnalisées, sélectionnez les métriques individuelles dans les listes déroulantes Métrique de charge et Métrique de mise à l'échelle.

7. Dans le champ Objectif d'utilisation, saisissez la valeur cible à maintenir par Amazon EC2 Auto Scaling. Amazon EC2 Auto Scaling adapte votre capacité jusqu'à ce que l'utilisation moyenne corresponde à l'objectif, ou jusqu'à ce qu'elle atteigne le nombre maximal d'instances que vous avez spécifié.

Si votre métrique de mise à l'échelle est...	L'objectif d'utilisation représente...
CPU	Pourcentage du processeur que chaque instance devrait idéalement utiliser.
Entrée réseau	Nombre moyen d'octets par minute que chaque instance devrait idéalement recevoir.
Sortie réseau	Nombre moyen d'octets par minute que chaque instance devrait idéalement envoyer.
Nombre de demandes Application Load Balancer par cible	Nombre moyen de demandes par minute que chaque instance devrait idéalement recevoir.

8. (Facultatif) Dans le champ Pré-lancement des instances, choisissez combien de temps à l'avance vous souhaitez que vos instances soient lancées avant que la prévision n'appelle une augmentation de la charge.

9. (Facultatif) Dans le champ Max capacity behavior (Comportement de capacité maximale), choisissez d'autoriser ou non Amazon EC2 Auto Scaling à monter en puissance au-delà de la capacité maximale du groupe lorsque la capacité prévue dépasse le maximum défini. L'activation de ce paramètre permet de monter en puissance pendant les périodes où l'on prédit que le trafic sera le plus élevé.
10. (Facultatif) Dans le champ Capacité maximale du tampon supérieure à la capacité prédite, choisissez la capacité supplémentaire à utiliser lorsque la capacité prévue est proche de la capacité maximale ou la dépasse. La valeur est exprimée en pourcentage par rapport à la capacité prévue. Par exemple, si la valeur du tampon est 10, cela implique un tampon de 10 %. Par conséquent, si la capacité prévue est 50 et que la capacité maximale est 40, la capacité maximale effective est 55.

    Si la valeur est définie sur 0, Amazon EC2 Auto Scaling peut augmenter la capacité au-delà de la capacité maximale pour égaler la capacité prévue, mais pas la dépasser.
11. Choisissez Créer une stratégie de mise à l'échelle prédictive.

Créer une politique de mise à l'échelle prédictive dans la console (métriques personnalisées)

Utilisez la procédure suivante pour créer une politique de mise à l'échelle prédictive à l'aide de métriques personnalisées. Les métriques personnalisées peuvent inclure d'autres métriques fournies par CloudWatch ou sur lesquelles vous publiez CloudWatch. Pour utiliser les métriques de CPU, d'I/O réseau ou de nombre de demandes Application Load Balancer par cible, consultez Créer une politique de mise à l'échelle prédictive dans la console (métriques prédéfinies).

Pour créer une politique de mise à l'échelle prédictive à l'aide de métriques personnalisées, procédez comme suit :

- Vous devez fournir les requêtes brutes qui permettent à Amazon EC2 Auto Scaling d'interagir avec les métriques contenues. CloudWatch Pour plus d'informations, consultez Configurations avancées de la politique de mise à l'échelle prédictive à l'aide de métriques personnalisées. Pour être sûr qu'Amazon EC2 Auto Scaling peut extraire les données métriques CloudWatch, vérifiez que chaque requête renvoie des points de données. Confirmez cela à l'aide de la CloudWatch console ou de l'opération CloudWatch GetMetricDataAPI.

  > **Note**
  >
  > Nous fournissons des exemples de charges utiles JSON dans l'éditeur JSON de la console Amazon EC2 Auto Scaling. Ces exemples vous fournissent une référence pour les paires

clé-valeur requises pour ajouter d'autres CloudWatch mesures fournies par AWS ou sur lesquelles vous avez précédemment publié. CloudWatch Vous pouvez les utiliser comme point de départ, puis les personnaliser en fonction de vos besoins.

- Si vous utilisez des mathématiques de métriques, vous devez construire manuellement la charge utile JSON pour l'adapter à votre scénario unique. Pour plus d'informations, consultez Utiliser des expressions mathématiques de métrique. Avant d'utiliser les mathématiques de métriques dans votre politique, vérifiez que les requêtes de métriques basées sur des expressions mathématiques de métriques sont valides et renvoient une seule série temporelle. Confirmez cela à l'aide de la CloudWatch console ou de l'opération CloudWatch GetMetricDataAPI.

Si vous faites une erreur dans une requête en fournissant des données incorrectes, par exemple un nom de groupe Auto Scaling erroné, la prévision ne contiendra aucune donnée. Pour résoudre les problèmes liés aux métriques personnalisées, consultez Considérations et dépannage.

Pour créer une politique de mise à l'échelle prédictive

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Sous l'onglet Mise à l'échelle automatique, accédez au champ Stratégies de mise à l'échelle et choisissez Créer une stratégie de mise à l'échelle prédictive.
4. Entrez le nom de la politique.
5. Activez Mise à l'échelle basée sur la prévision pour autoriser Amazon EC2 Auto Scaling à lancer immédiatement la mise à l'échelle.

   Pour que la stratégie reste en mode prévision uniquement, n'activez pas Mise à l'échelle basée sur la prévision.
6. Pour Metrics (Métriques), choisissez Custom metric pair (Paire de métriques personnalisées).

   a. Pour Métrique de charge, choisissez CloudWatch Métrique personnalisée pour utiliser une métrique personnalisée. Construisez la charge utile JSON qui contient la définition de la métrique de charge de la politique et collez-la dans la zone de l'éditeur JSON pour remplacer le contenu qui s'y trouve déjà.

b. Pour Scaling metric, choisissez Custom CloudWatch metric pour utiliser une métrique personnalisée. Construisez la charge utile JSON qui contient la définition de la métrique de mise à l'échelle de la politique et collez-la dans la zone de l'éditeur JSON pour remplacer le contenu qui s'y trouve déjà.

c. (Facultatif) Pour ajouter une métrique de capacité personnalisée, cochez la case Add custom capacity metric (Ajouter une métrique de capacité personnalisée). Construisez la charge utile JSON qui contient la définition de la métrique de capacité personnalisée de la politique et collez-la dans la zone de l'éditeur JSON pour remplacer le contenu qui s'y trouve déjà.

   Vous devez uniquement activer cette option pour créer une nouvelle série temporelle de capacité si vos données de métriques de capacité couvrent plusieurs groupes Auto Scaling. Dans ce cas, vous devez utiliser les mathématiques de métriques pour agréger les données en une seule série temporelle.

7. Dans le champ Objectif d'utilisation, saisissez la valeur cible à maintenir par Amazon EC2 Auto Scaling. Amazon EC2 Auto Scaling adapte votre capacité jusqu'à ce que l'utilisation moyenne corresponde à l'objectif, ou jusqu'à ce qu'elle atteigne le nombre maximal d'instances que vous avez spécifié.
8. (Facultatif) Dans le champ Pré-lancement des instances, choisissez combien de temps à l'avance vous souhaitez que vos instances soient lancées avant que la prévision n'appelle une augmentation de la charge.
9. (Facultatif) Dans le champ Max capacity behavior (Comportement de capacité maximale), choisissez d'autoriser ou non Amazon EC2 Auto Scaling à monter en puissance au-delà de la capacité maximale du groupe lorsque la capacité prévue dépasse le maximum défini. L'activation de ce paramètre permet de monter en puissance pendant les périodes où l'on prédit que le trafic sera le plus élevé.
10. (Facultatif) Dans le champ Capacité maximale du tampon supérieure à la capacité prédite, choisissez la capacité supplémentaire à utiliser lorsque la capacité prévue est proche de la capacité maximale ou la dépasse. La valeur est exprimée en pourcentage par rapport à la capacité prévue. Par exemple, si la valeur du tampon est 10, cela implique un tampon de 10 %. Par conséquent, si la capacité prévue est 50 et que la capacité maximale est 40, la capacité maximale effective est 55.

    Si la valeur est définie sur 0, Amazon EC2 Auto Scaling peut augmenter la capacité au-delà de la capacité maximale pour égaler la capacité prévue, mais pas la dépasser.
11. Choisissez Créer une stratégie de mise à l'échelle prédictive.

Pour créer une stratégie de mise à l'échelle prédictive (AWS CLI)

Utilisez ce qui AWS CLI suit pour configurer les politiques de dimensionnement prédictif pour votre groupe Auto Scaling. Remplacez chaque *espace réservé à la saisie de l'utilisateur* par vos propres informations.

Pour plus d'informations sur les CloudWatch métriques que vous pouvez spécifier, consultez le PredictiveScalingMetricSpecificationmanuel Amazon EC2 Auto Scaling API Reference.

Exemple 1 : stratégie de mise à l'échelle prédictive qui crée des prévisions mais ne met pas à l'échelle

L'exemple de stratégie suivant illustre une configuration de stratégie complète qui utilise les métriques d'utilisation du processeur pour la mise à l'échelle prédictive avec un objectif d'utilisation de `40`. Le mode `ForecastOnly` est utilisé par défaut, sauf si vous spécifiez explicitement le mode à utiliser. Enregistrez cette configuration dans un fichier nommé `config.json`.

```
{
    "MetricSpecifications": [
        {
            "TargetValue": 40,
            "PredefinedMetricPairSpecification": {
                "PredefinedMetricType": "ASGCPUUtilization"
            }
        }
    ]
}
```

Pour créer la politique à partir de la ligne de commande, exécutez la put-scaling-policycommande avec le fichier de configuration spécifié, comme illustré dans l'exemple suivant.

```
aws autoscaling put-scaling-policy --policy-name cpu40-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
```

Si elle aboutit, cette commande renvoie l'Amazon Resource Name (ARN) de la stratégie.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:2f4f5048-d8a8-4d14-
b13a-d1905620f345:autoScalingGroupName/my-asg:policyName/cpu40-predictive-scaling-
policy",
```

```
  "Alarms": []
}
```

Exemple 2 : stratégie de mise à l'échelle prédictive qui prédit et met à l'échelle

Pour créer une stratégie permettant à Amazon EC2 Auto Scaling de prédire et de mettre à l'échelle, ajoutez la propriété `Mode` associée à la valeur `ForecastAndScale`. L'exemple suivant illustre une configuration de stratégie qui utilise les métriques Nombre de demandes Application Load Balancer de charge. L'objectif d'utilisation est de `1000`, et la mise à l'échelle prédictive est définie sur le mode `ForecastAndScale`.

```
{
    "MetricSpecifications": [
        {
            "TargetValue": 1000,
            "PredefinedMetricPairSpecification": {
                "PredefinedMetricType": "ALBRequestCount",
                "ResourceLabel": "app/my-alb/778d41231b141a0f/targetgroup/my-alb-
target-group/943f017f100becff"
            }
        }
    ],
    "Mode": "ForecastAndScale"
}
```

Pour créer cette politique, exécutez la put-scaling-policycommande avec le fichier de configuration spécifié, comme illustré dans l'exemple suivant.

```
aws autoscaling put-scaling-policy --policy-name alb1000-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
```

Si elle aboutit, cette commande renvoie l'Amazon Resource Name (ARN) de la stratégie.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-
id:scalingPolicy:19556d63-7914-4997-8c81-d27ca5241386:autoScalingGroupName/my-
asg:policyName/alb1000-predictive-scaling-policy",
  "Alarms": []
}
```

Exemple 3 : stratégie de mise à l'échelle prédictive qui peut augmenter la capacité au-delà de la capacité maximale

L'exemple suivant montre comment créer une stratégie permettant d'aller au-delà de la limite de taille maximale du groupe lorsque vous en avez besoin pour gérer une charge supérieure à la normale. Par défaut, Amazon EC2 Auto Scaling ne va pas au-delà de la capacité maximale définie pour votre instance EC2. Toutefois, un léger dépassement de la capacité maximale peut parfois être utile afin d'éviter les problèmes de performances ou de disponibilité.

Pour permettre à Amazon EC2 Auto Scaling d'approvisionner de la capacité supplémentaire lorsque les prévisions sont égales ou très proches de la taille maximale de votre groupe, définissez les propriétés `MaxCapacityBreachBehavior` et `MaxCapacityBuffer`, comme illustré dans l'exemple suivant. Vous devez définir la propriété `MaxCapacityBreachBehavior` sur `IncreaseMaxCapacity`. Le nombre maximal d'instances que votre groupe peut accueillir dépend de la valeur de la propriété `MaxCapacityBuffer`.

```
{
    "MetricSpecifications": [
        {
            "TargetValue": 70,
            "PredefinedMetricPairSpecification": {
                "PredefinedMetricType": "ASGCPUUtilization"
            }
        }
    ],
    "MaxCapacityBreachBehavior": "IncreaseMaxCapacity",
    "MaxCapacityBuffer": 10
}
```

Dans cet exemple, la stratégie est configurée pour utiliser un tampon de 10 % (`"MaxCapacityBuffer": 10`). Ainsi, si la capacité prévue est de 50 et la capacité maximale de 40, la capacité maximale effective est de 55. Une stratégie permettant d'augmenter la capacité au-delà de la capacité maximale pour égaler mais pas dépasser la capacité prévue aurait un tampon de 0 (`"MaxCapacityBuffer": 0`).

Pour créer cette politique, exécutez la put-scaling-policycommande avec le fichier de configuration spécifié, comme illustré dans l'exemple suivant.

```
aws autoscaling put-scaling-policy --policy-name cpu70-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
```

```
  --predictive-scaling-configuration file://config.json
```

Si elle aboutit, cette commande renvoie l'Amazon Resource Name (ARN) de la stratégie.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:d02ef525-8651-4314-
bf14-888331ebd04f:autoScalingGroupName/my-asg:policyName/cpu70-predictive-scaling-
policy",
  "Alarms": []
}
```

Évaluer vos politiques de mise à l'échelle prédictive

Avant d'utiliser une politique de mise à l'échelle prédictive pour mettre à l'échelle votre groupe Auto Scaling, consultez les recommandations et les autres données relatives à votre politique dans la console Amazon EC2 Auto Scaling. C'est une étape importante pour éviter qu'une politique de mise à l'échelle prédictive mette votre capacité réelle à l'échelle tant que vous ne savez pas si ses prévisions sont exactes.

Si le groupe Auto Scaling est nouveau, laissez 24 heures à Amazon EC2 Auto Scaling pour créer la première prévision.

Lorsqu'Amazon EC2 Auto Scaling crée une prévision, il utilise des données historiques. Si votre groupe Auto Scaling ne dispose pas encore de nombreuses données historiques récentes, Amazon EC2 Auto Scaling peut temporairement remplir la prévision avec des agrégats créés à partir des agrégats historiques actuellement disponibles. Les prévisions sont remplies jusqu'à deux semaines avant la date de création d'une politique.

Table des matières

Afficher vos recommandations de mise à l'échelle prédictive

Pour une analyse efficace, Amazon EC2 Auto Scaling doit disposer d'au moins deux politiques de mise à l'échelle prédictive à comparer. (Toutefois, vous pouvez toujours consulter les résultats d'une seule politique.) Lorsque vous créez plusieurs politiques, vous pouvez évaluer une politique qui utilise

une seule métrique par rapport à une autre qui utilise une autre métrique. Vous pouvez également évaluer l'impact de différentes combinaisons de valeurs cibles et de métriques. Une fois les politiques de mise à l'échelle prédictive créées, Amazon EC2 Auto Scaling commence immédiatement à évaluer la politique la plus appropriée pour mettre votre groupe à l'échelle.

Pour afficher vos recommandations dans la console Amazon EC2 Auto Scaling

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s’ouvre en bas de la page.
3. Dans l'onglet Auto Scaling, sous Politiques de mise à l'échelle prédictive, vous pouvez afficher les détails d'une politique ainsi que notre recommandation. La recommandation vous indique si la politique de mise à l'échelle prédictive est plus efficace que si vous ne l'utilisez pas.

   Si vous ne savez pas si une politique de mise à l'échelle prédictive convient à votre groupe, consultez les colonnes Impact sur la disponibilité et Impact sur les coûts pour choisir la politique appropriée. Les informations de chaque colonne vous indiquent l'impact de la politique.

   - Impact sur la disponibilité : indique si la politique permettrait d'éviter un impact négatif sur la disponibilité en provisionnant suffisamment d'instances pour gérer la charge de travail, en comparaison avec sa non-utilisation.
   - Impact sur les coûts : indique si la politique permettrait d'éviter un impact négatif sur vos coûts en ne surprovisionnant pas les instances, en comparaison avec sa non-utilisation. En cas de surprovisionnement excessif, vos instances sont sous-utilisées ou inactives, ce qui ne fait qu'augmenter l'impact sur les coûts.

   Si vous avez plusieurs politiques, la balise Meilleure prévision s'affiche à côté du nom de la politique qui offre le plus d'avantages en matière de disponibilité à moindre coût. Une plus grande importance est accordée à l'impact sur la disponibilité.
4. (Facultatif) Pour sélectionner la période souhaitée pour les résultats des recommandations, choisissez la valeur de votre choix dans la liste déroulante Période d'évaluation : 2 jours, 1 semaine, 2 semaines, 4 semaines, 6 semaines ou 8 semaines. Par défaut, la période d'évaluation est réglée sur les deux dernières semaines. Une période d'évaluation plus longue fournit davantage de points de données pour les résultats des recommandations. Toutefois, l'ajout de points de données supplémentaires risque de ne pas améliorer les résultats si vos

modèles de charge ont changé, par exemple après une période de demande exceptionnelle. Dans ce cas, vous pouvez obtenir une recommandation plus ciblée en consultant des données plus récentes.

> **Note**
>
> Les recommandations sont générées uniquement pour les politiques qui sont en mode Prévision uniquement. La fonctionnalité des recommandations offre de meilleurs résultats lorsqu'une politique est en mode Prévision uniquement pendant toute la période d'évaluation. Si vous lancez une politique en mode Prévision et mise à l'échelle et que vous la passez ultérieurement en mode Prévision uniquement, les résultats de cette politique risquent d'être biaisés. Cela s'explique par le fait que la politique a déjà contribué à la capacité réelle.

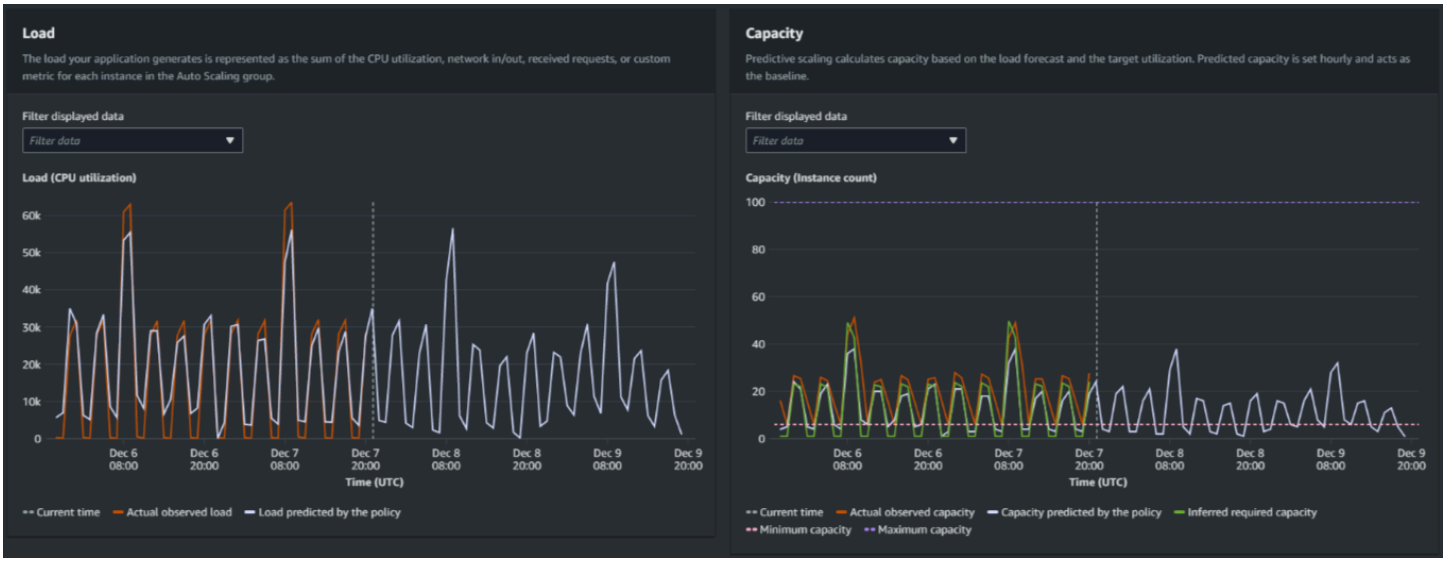
Consulter les graphiques de surveillance de la mise à l'échelle prédictive

Dans la console Amazon EC2 Auto Scaling, vous pouvez consulter les prévisions des jours, semaines ou mois précédents afin de visualiser les performances de la politique au fil du temps. Vous pouvez également utiliser ces informations pour évaluer la précision des prévisions lorsque vous décidez de laisser une politique mettre votre capacité réelle à l'échelle.

Pour consulter les graphiques de surveillance de la mise à l'échelle prédictive dans la console Amazon EC2 Auto Scaling

1. Choisissez une politique dans la liste Politiques de mise à l'échelle prédictive.
2. Dans la section Surveillance, vous pouvez afficher les prévisions passées et futures de votre politique concernant la charge et la capacité par rapport aux valeurs réelles. Le graphique Charge présente la prévision de charge et les valeurs réelles pour la métrique de charge que vous avez choisie. Le graphique Capacité indique le nombre d'instances prédit par la politique. Il inclut également le nombre réel d'instances lancées. La ligne verticale sépare les valeurs historiques des prévisions futures. Ces graphiques sont disponibles peu de temps après la création de la politique.
3. (Facultatif) Pour modifier la quantité de données historiques affichées dans le graphique, choisissez la valeur de votre choix dans la liste déroulante Période d'évaluation en haut de la page. La période d'évaluation ne transforme en rien les données de cette page. Elle ne fait que modifier la quantité de données historiques affichées.

L'image suivante montre les graphiques Charge et Capacité lorsque les prévisions ont été appliquées plusieurs fois. Les prévisions de mise à l'échelle prédictive se chargent en fonction de vos données de charge historiques. La charge générée par votre application est représentée comme la somme de l'utilisation du processeur, des entrées/sorties réseau, des demandes reçues ou des métriques personnalisées pour chaque instance du groupe Auto Scaling. La mise à l'échelle prédictive calcule les besoins de capacité future en fonction de la prévision de charge et de l'utilisation cible que vous souhaitez atteindre pour la métrique de mise à l'échelle.



Comparer les données du graphique Charge

Chaque ligne horizontale représente un ensemble différent de points de données rapportés à des intervalles d'une heure :

1. La charge observée réelle utilise la statistique SUM correspondant à la métrique de charge de votre choix afin d'afficher la charge horaire totale dans le passé.
2. La charge prévue par la politique indique la prévision de charge horaire. Cette prévision se base sur les observations de charge réelles des deux semaines précédentes.

Comparez les données du graphique Capacité

Chaque ligne horizontale représente un ensemble différent de points de données rapportés à des intervalles d'une heure :

1. La capacité observée réelle indique la capacité réelle de votre groupe Auto Scaling dans le passé, qui dépend de vos autres politiques de mise à l'échelle et de la taille minimale du groupe en vigueur pour la période sélectionnée.

2. La capacité prévue par la politique indique la capacité de base à laquelle vous pouvez vous attendre au début de chaque heure lorsque la politique est en mode Prévision et mise à l'échelle.
3. La capacité requise déduite indique la capacité idéale pour maintenir la métrique de mise à l'échelle à la valeur cible que vous avez définie.
4. La capacité minimale indique la capacité minimale du groupe Auto Scaling.
5. La capacité maximale indique la capacité maximale du groupe Auto Scaling.

Afin de calculer la capacité requise déduite, nous partons du principe que chaque instance est utilisée de manière égale à une valeur cible spécifiée. Dans la pratique, les instances ne sont pas utilisées de manière égale. En supposant que l'utilisation est uniformément répartie entre les instances, nous pouvons toutefois établir une estimation probable de la quantité de capacité requise. La capacité requise est ensuite calculée de manière à être inversement proportionnelle à la métrique de mise à l'échelle que vous avez utilisée pour votre politique de mise à l'échelle prédictive. En d'autres termes, la métrique de mise à l'échelle diminue à mesure que la capacité augmente. Par exemple, si la capacité double, la métrique de mise à l'échelle doit être réduite de moitié.

La formule de la capacité requise déduite est la suivante :

```
sum of (actualCapacityUnits*scalingMetricValue)/(targetUtilization)
```

Par exemple, nous prenons les `actualCapacityUnits` (`10`) et la `scalingMetricValue` (`30`) pour une heure donnée. Nous prenons ensuite la `targetUtilization` que vous avez spécifiée dans votre politique de mise à l'échelle prédictive (`60`) et calculons la capacité requise déduite pour la même heure. Elle renvoie la valeur `5`. Cela signifie que cinq est la quantité de capacité requise déduite pour maintenir la capacité en proportion inverse directe à la valeur cible de la métrique de mise à l'échelle.

> **Note**
>
> Différents leviers sont à votre disposition pour ajuster et améliorer les économies de coûts et la disponibilité de votre application.
>
> - Vous utilisez la mise à l'échelle prédictive pour la capacité de base et la mise à l'échelle dynamique afin de gérer la capacité supplémentaire. La mise à l'échelle dynamique fonctionne indépendamment de la mise à l'échelle prédictive, avec une mise à l'échelle horizontale et une montée en puissance en fonction de l'utilisation actuelle. Tout d'abord, Amazon EC2 Auto Scaling calcule le nombre d'instances recommandé pour chaque

politique de mise à l'échelle dynamique. Il est ensuite mis à l'échelle selon la politique qui fournit le plus grand nombre d'instances.

- Pour permettre la mise à l'échelle horizontale lorsque la charge diminue, votre groupe Auto Scaling doit toujours disposer d'au moins une politique de mise à l'échelle dynamique avec la partie de mise à l'échelle horizontale activée.
- Vous pouvez améliorer les performances de mise à l'échelle en vous assurant que vos capacités minimale et maximale ne sont pas trop restrictives. Une politique comportant un nombre recommandé d'instances qui ne se situe pas dans la plage de capacité minimale et maximale ne pourra pas être mise à l'échelle horizontalement ni montée en puissance.

Surveillez les mesures de dimensionnement prédictives avec CloudWatch

Selon vos besoins, vous préférerez peut-être accéder aux données de surveillance à des fins de dimensionnement prédictif depuis Amazon CloudWatch plutôt que depuis la console Amazon EC2 Auto Scaling. Une fois que vous avez créé une politique de mise à l'échelle prédictif, celle-ci collecte des données qui sont utilisées pour prévoir votre charge et votre capacité future. Une fois ces données collectées, elles sont automatiquement stockées CloudWatch à intervalles réguliers. Vous pouvez ensuite l'utiliser CloudWatch pour visualiser les performances de la politique au fil du temps. Vous pouvez également créer des CloudWatch alarmes pour vous avertir lorsque les indicateurs de performance changent au-delà des limites que vous avez définies CloudWatch.

Rubriques

- Visualisez les données de prévision historiques
- Créer des mesures de précision à l'aide de mathématiques

Visualisez les données de prévision historiques

Vous pouvez consulter les données de prévision de charge et de capacité pour une politique de dimensionnement prédictive dans CloudWatch. Cela peut être utile lorsque vous visualisez les prévisions par rapport à d'autres CloudWatch indicateurs dans un seul graphique. Cela peut également être utile lors de l'affichage d'une plage de temps plus large, afin de voir les tendances au fil du temps. Vous pouvez accéder aux métriques historiques jusqu'à 15 mois pour acquérir un meilleur point de vue de la façon dont votre politique s'exécute.

Pour plus d’informations, consultez Métriques et dimensions de mise à l'échelle.

Pour consulter les données de prévision historiques à l'aide de la CloudWatch console

1. Ouvrez la CloudWatch console à l'adresse https://console.aws.amazon.com/cloudwatch/.
2. Dans le panneau de navigation, choisissez Metrics (Métriques), All metrics (Toutes les métriques).
3. Cliquez sur l'ongletAuto Scalingmetric namespace.
4. Choisissez l'une des options suivantes pour afficher les mesures de prévision de charge ou de prévision de capacité :

   - Prévisions de charge prédictive Scaling
   - Prévisions de capacité d'évolutivité prédictive
5. Dans le champ de recherche, entrez le nom de la stratégie de mise à l'échelle prédictive ou le nom du groupe Auto Scaling, puis appuyez sur Entrée pour filtrer les résultats.
6. Pour représenter graphiquement une métrique, cochez la case en regard de la métrique. Pour modifier le nom du graphique, choisissez l'icône représentant un crayon. Pour modifier la plage de temps, sélectionnez l'une des valeurs prédéfinies ou choisissez custom (personnalisé). Pour plus d'informations, consultez la section Représentation graphique d'une métrique dans le guide de CloudWatch l'utilisateur Amazon.
7. Pour modifier les statistiques, choisissez l'onglet Graphed metrics (Graphique des métriques). Sélectionnez l'en-tête de colonne ou une valeur individuelle et choisissez une autre valeur. Bien que vous puissiez choisir n'importe quelle statistique pour chaque métrique, toutes les statistiques ne sont pas utiles pour PredictiveScalingLoadForecastles PredictiveScalingCapacityForecastmétriques. Par exemple, les calculs statistiques de moyenne, minimum et maximum de l'utilisation de l'UC sont utiles, mais le calcul statistique de somme ne l'est pas.
8. (En option) Pour ajouter une autre métrique à utiliser dans l'expression mathématique, sous Toutes les métriques, choisissez Tous, recherchez la métrique spécifique, puis activez la case à cocher en regard de celle-ci. Vous pouvez ajouter jusqu'à 10 métriques.

   Par exemple, pour ajouter les valeurs réelles de l'utilisation de l'UC au graphique, choisissez l'optionEC2espace de noms, puis choisissezPar Auto Scaling Group. Activez ensuite la case à cocherCPUUtilizationmetric et le groupe Auto Scaling.
9. (Facultatif) Pour ajouter le graphique à un CloudWatch tableau de bord, choisissez Actions, puis Ajouter au tableau de bord.

Créer des mesures de précision à l'aide de mathématiques

Avec les mathématiques métriques, vous pouvez interroger plusieurs CloudWatch métriques et utiliser des expressions mathématiques pour créer de nouvelles séries chronologiques basées sur ces métriques. Vous pouvez visualiser les séries chronologiques obtenues sur la CloudWatch console et les ajouter aux tableaux de bord. Pour plus d'informations sur les mathématiques métriques, consultez la section Utilisation des mathématiques métriques dans le guide de CloudWatch l'utilisateur Amazon.

À l'aide de métriques mathématiques, vous pouvez représenter graphiquement les données générées par Amazon EC2 Auto Scaling pour une mise à l'échelle prédictive de différentes manières. Cela vous permet de surveiller les performances des politiques au fil du temps et de comprendre si votre combinaison de mesures peut être améliorée.

Par exemple, vous pouvez utiliser une expression mathématique de métrique pour surveillererreur absolue moyenne en pourcentage(MÂLE). La métrique MAPE permet de surveiller la différence entre les valeurs prévues et les valeurs réelles observées pendant une fenêtre de prévision donnée. Les modifications de la valeur de MAPE peuvent indiquer si les performances de la stratégie se dégradent au fil du temps à mesure que la nature de votre application change. Une augmentation de l'EMAP indique un écart plus important entre les valeurs prévues et les valeurs réelles.

Exemple : expressions mathématiques appliquées aux métriques

Pour commencer avec ce type de graphique, vous pouvez créer une expression mathématique de métrique comme celle présentée dans l'exemple suivant.

```
{
  "MetricDataQueries": [
    {
      "Expression": "TIME_SERIES(AVG(ABS(m1-m2)/m1))",
      "Id": "e1",
      "Period": 3600,
      "Label": "MeanAbsolutePercentageError",
      "ReturnData": true
    },
    {
      "Id": "m1",
      "Label": "ActualLoadValues",
      "MetricStat": {
        "Metric": {
          "Namespace": "AWS/EC2",
          "MetricName": "CPUUtilization",
```

```
          "Dimensions": [
            {
              "Name": "AutoScalingGroupName",
              "Value": "my-asg"
            }
          ]
        },
        "Period": 3600,
        "Stat": "Sum"
      },
      "ReturnData": false
    },
    {
      "Id": "m2",
      "Label": "ForecastedLoadValues",
      "MetricStat": {
        "Metric": {
          "Namespace": "AWS/AutoScaling",
          "MetricName": "PredictiveScalingLoadForecast",
          "Dimensions": [
            {
              "Name": "AutoScalingGroupName",
              "Value": "my-asg"
            },
            {
              "Name": "PolicyName",
              "Value": "my-predictive-scaling-policy"
            },
            {
              "Name": "PairIndex",
              "Value": "0"
            }
          ]
        },
        "Period": 3600,
        "Stat": "Average"
      },
      "ReturnData": false
    }
  ]
}
```

Au lieu d'une seule métrique, il existe un tableau de structures de requêtes de données de métriques pour`MetricDataQueries`. Chaque élément de`MetricDataQueries`obtient une métrique ou exécute une expression mathématique. Le premier point,`e1`, est l'expression mathématique. L'expression désignée définit le`ReturnData`Paramètre to`true`, qui génère une seule série chronologique. Pour toutes les autres métriques, le kit`ReturnData`valeur est`false`.

Dans l'exemple, l'expression désignée utilise les valeurs réelles et prévues comme entrée et renvoie la nouvelle métrique (MAPE). `m1`est la CloudWatch métrique qui contient les valeurs de charge réelles (en supposant que l'utilisation du processeur est la métrique de charge initialement spécifiée pour la politique nommée`my-predictive-scaling-policy`). `m2`est la CloudWatch métrique qui contient les valeurs de charge prévues. La syntaxe mathématique de la métrique MAPE est la suivante :

Moyenne de (abs ((Réel - Forecast)/(Réel)))

Visualisez vos mesures de précision et définissez des alarmes

Pour visualiser les données métriques de précision, sélectionnez l'onglet Métriques dans la CloudWatch console. Vous pouvez représenter les données sous forme graphique à partir de là. Pour plus d'informations, consultez la section Ajouter une expression mathématique à un CloudWatch graphique dans le guide de CloudWatch l'utilisateur Amazon.

Vous pouvez définir une alarme sur une métrique que vous surveillez à partir de la section Metrics (Métriques). Pendant que vous êtes sur l'onglet Métriques sous forme de graphique, vous pouvez sélectionner l'icône Créer une alarme sous la colonne Actions. Le Créer une alarme est représentée par une petite cloche. Pour plus d'informations et pour connaître les options de notification, consultez les sections Création CloudWatch d'une alarme basée sur une expression mathématique métrique et Notification aux utilisateurs des modifications apportées aux alarmes dans le guide de l' CloudWatch utilisateur Amazon.

Vous pouvez également utiliser GetMetricDataet effectuer des calculs PutMetricAlarmà l'aide des mathématiques métriques et créer des alarmes en fonction de la sortie.

Remplacer des valeurs de prévision à l'aide d'actions planifiées

Parfois, vous pouvez disposer d'informations supplémentaires sur de futurs besoins de votre application que le calcul prédictif ne peut pas prendre en compte. Par exemple, les calculs prédictifs peuvent sous-estimer la capacité nécessaire pour un événement marketing à venir. Vous pouvez alors utiliser des actions planifiées pour remplacer temporairement la prévision au cours de périodes

ultérieures. Les actions planifiées peuvent être exécutées de manière récurrente, ou à une date et une heure spécifiques en cas de fluctuations ponctuelles de la demande.

Par exemple, vous pouvez créer une action planifiée avec une capacité minimale plus élevée que ce qui est prédit. Au moment de l'exécution, Amazon EC2 Auto Scaling met à jour la capacité minimale de votre groupe Auto Scaling. Étant donné que la mise à l'échelle prédictive optimise la capacité, une action planifiée avec une capacité minimale supérieure aux valeurs prédites est honorée. Cela permet d'éviter que la capacité soit inférieure à celle prévue. Pour cesser de remplacer la prévision, utilisez une deuxième action planifiée afin de rétablir le paramètre d'origine de la capacité minimale.

La procédure suivante présente les étapes à suivre pour remplacer la prévision au cours de périodes ultérieures.

Rubriques

> **Important**
>
> Cette rubrique part du principe que vous essayez de modifier les prévisions pour vous adapter à une capacité supérieure à celle prévue. Si vous devez réduire temporairement la capacité sans interférer avec une politique de dimensionnement prédictif, utilisez plutôt le mode prévision uniquement. En mode prévisions uniquement, la mise à l'échelle prédictive continuera de générer des prévisions, mais elle n'augmentera pas automatiquement la capacité. Vous pouvez ensuite surveiller l'utilisation des ressources et réduire manuellement la taille de votre groupe selon vos besoins. Pour plus d'informations sur le dimensionnement manuel, consultezMise à l'échelle manuelle pour Amazon EC2 Auto Scaling.

Étape 1 : (facultatif) Analyser les données en séries chronologiques

Commencez par analyser les données en séries chronologiques de la prévision. Il s'agit d'une étape facultative, mais elle permet de comprendre les détails de la prévision.

1. Récupérer la prévision

Une fois la prévision créée, vous pouvez interroger une période spécifique au sein de celle-ci. L'objectif de la requête est d'obtenir une vue complète des données en séries chronologiques d'une période spécifique.

Votre requête peut inclure jusqu'à deux jours de données de prévision ultérieures. Si vous utilisez la mise à l'échelle prédictive depuis un certain temps, vous pouvez également accéder à vos données de prévision antérieures. Toutefois, la durée maximale entre le début et la fin est de 30 jours.

Pour obtenir les prévisions à l'aide de la get-predictive-scaling-forecast AWS CLI commande, entrez les paramètres suivants dans la commande :

- Entrez le nom du groupe Auto Scaling dans le paramètre `--auto-scaling-group-name`.
- Entrez le nom de la stratégie dans le paramètre `--policy-name`.
- Entrez l'heure de début dans le paramètre `--start-time` pour ne renvoyer que les données de prévision correspondant à l'heure spécifiée ou ultérieures à celle-ci.
- Entrez l'heure de fin dans le paramètre `--end-time` pour ne renvoyer que les données de prévision correspondant à l'heure spécifiée ou antérieures à celle-ci.

```
aws autoscaling get-predictive-scaling-forecast --auto-scaling-group-name my-asg \
    --policy-name cpu40-predictive-scaling-policy \
    --start-time "2021-05-19T17:00:00Z" \
    --end-time "2021-05-19T23:00:00Z"
```

Si elle aboutit, la commande renvoie des données semblables à l'exemple suivant.

```
{
    "LoadForecast": [
        {
            "Timestamps": [
                "2021-05-19T17:00:00+00:00",
                "2021-05-19T18:00:00+00:00",
                "2021-05-19T19:00:00+00:00",
                "2021-05-19T20:00:00+00:00",
                "2021-05-19T21:00:00+00:00",
                "2021-05-19T22:00:00+00:00",
                "2021-05-19T23:00:00+00:00"
            ],
```

```
            "Values": [
                153.0655799339254,
                128.8288551285919,
                107.1179447150675,
                197.3601844551528,
                626.4039934516954,
                596.9441277518481,
                677.9675713779869
            ],
            "MetricSpecification": {
                "TargetValue": 40.0,
                "PredefinedMetricPairSpecification": {
                    "PredefinedMetricType": "ASGCPUUtilization"
                }
            }
        }
    ],
    "CapacityForecast": {
        "Timestamps": [
            "2021-05-19T17:00:00+00:00",
            "2021-05-19T18:00:00+00:00",
            "2021-05-19T19:00:00+00:00",
            "2021-05-19T20:00:00+00:00",
            "2021-05-19T21:00:00+00:00",
            "2021-05-19T22:00:00+00:00",
            "2021-05-19T23:00:00+00:00"
        ],
        "Values": [
            2.0,
            2.0,
            2.0,
            2.0,
            4.0,
            4.0,
            4.0
        ]
    },
    "UpdateTime": "2021-05-19T01:52:50.118000+00:00"
}
```


La réponse comprend deux prévisions : `LoadForecast` et `CapacityForecast`. `LoadForecast` affiche la prévision de charge horaire. `CapacityForecast` affiche les valeurs

de prévision de la capacité nécessaire sur une base horaire pour gérer la charge prévue tout en maintenant une `TargetValue` de 40 (40 % d'utilisation moyenne du processeur).

2. Identifier la période cible

   Indiquez l'heure ou les heures où la fluctuation de la demande ponctuelle devrait avoir lieu. N'oubliez pas que les dates et les heures indiquées dans la prévision sont basées sur le fuseau horaire UTC.

Étape 2 : créer deux actions planifiées

Créez ensuite deux actions planifiées pour une période spécifique où votre application devra gérer une charge plus élevée que celle prédite. Par exemple, si vous organisez un événement marketing qui va générer du trafic sur votre site pendant une période limitée, vous pouvez planifier une action ponctuelle pour mettre à jour la capacité minimale au début de cet événement. Puis vous pouvez planifier une autre action pour rétablir le paramètre d'origine de la capacité minimale à la fin de l'événement.

Pour créer deux actions planifiées pour des événements ponctuels (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s’ouvre en bas de la page.
3. Sous l'onglet Mise à l'échelle automatique dans Actions planifiées, choisissez Créer une action planifiée.
4. Renseignez les paramètres des actions planifiées suivants :

   a. Dans le champ Nom, attribuez un nom à l'action planifiée.

   b. Dans le champ Min, saisissez la nouvelle capacité minimale de votre groupe Auto Scaling. La valeur Min doit être inférieure ou égale à la taille maximale du groupe. Si la nouvelle valeur Min est supérieure à la taille maximale du groupe, vous devez mettre à jour la valeur Max.

   c. Pour Recurrence (Récurrence), choisissez Once (Une fois).

   d. Dans le champ Fuseau horaire, choisissez un fuseau horaire. Si aucun fuseau horaire n'est choisi, `ETC/UTC` est utilisé par défaut.

e. Définissez une Heure de début spécifique.

5. Choisissez Créer.

   La console affiche les actions planifiées pour le groupe Auto Scaling.

6. Configurez une seconde action planifiée afin de rétablir le paramètre d'origine de la capacité minimale à la fin de l'événement. La mise à l'échelle prédictive ne peut augmenter la capacité que lorsque la valeur Min que vous avez définie est inférieure aux valeurs prédites.

Pour créer deux actions planifiées pour des événements ponctuels (AWS CLI)

Pour AWS CLI créer les actions planifiées, utilisez la commande put-scheduled-update-group-action.

Par exemple, définissons une action planifiée pour maintenir une capacité minimale de trois instances pendant huit heures à partir du 19 mai à 17h00. Les commandes suivantes montrent comment implémenter ce scénario.

La première commande put-scheduled-update-group-action demande à Amazon EC2 Auto Scaling de mettre à jour la capacité minimale du groupe Auto Scaling spécifié à 17 h 00 UTC le 19 mai 2021.

```
aws autoscaling put-scheduled-update-group-action --scheduled-action-name my-event-
start \
  --auto-scaling-group-name my-asg --start-time "2021-05-19T17:00:00Z" --minimum-
capacity 3
```

La deuxième commande indique à Amazon EC2 Auto Scaling de définir la capacité minimale du groupe à 1h00 UTC le 20 mai 2021.

```
aws autoscaling put-scheduled-update-group-action --scheduled-action-name my-event-end
 \
  --auto-scaling-group-name my-asg --start-time "2021-05-20T01:00:00Z" --minimum-
capacity 1
```

Après l'ajout de ces actions planifiées au groupe Auto Scaling, Amazon EC2 Auto Scaling effectue les opérations suivantes :

- À 17h00 UTC le 19 mai 2021, la première action planifiée s'exécute. Si le groupe compte actuellement moins de trois instances, il passe à trois instances. Pendant ce temps et pendant les huit heures suivantes, Amazon EC2 Auto Scaling peut continuer à monter en puissance

si la capacité prévue est supérieure à la capacité réelle ou si une stratégie de mise à l'échelle dynamique est en vigueur.

- À 1h00 UTC le 20 mai 2021, la seconde action planifiée s'exécute. Cette action rétablit le paramètre d'origine de la capacité minimale à la fin de l'événement.

Mise à l'échelle basée sur des planifications récurrentes

Pour remplacer la prévision applicable à la même période chaque semaine, créez deux actions planifiées et fournissez la logique d'heure et de date à l'aide d'une expression cron.

L'expression cron est constituée de cinq champs séparés par des espaces : [Minute] [Heure] [Jour_du_Mois] [Mois_de_Année] [Jour_de_Semaine]. Ces champs peuvent contenir toutes les valeurs autorisées, y compris des caractères spéciaux.

Par exemple, l'expression cron suivante exécute l'action tous les mardis à 6h30. L'astérisque est utilisé comme caractère générique pour représenter toutes les valeurs d'un champ.

```
30 6 * * 2
```

Consultez aussi

Pour plus d'informations sur comment créer, répertorier, modifier et supprimer des actions planifiées, consultez Mise à l'échelle planifiée pour Amazon EC2 Auto Scaling.

Configurations avancées de la politique de mise à l'échelle prédictive à l'aide de métriques personnalisées

Dans une politique de mise à l'échelle prédictive, vous pouvez utiliser des métriques prédéfinies ou personnalisées. Les métriques personnalisées sont utiles lorsque les métriques prédéfinies (CPU, réseau I/O et nombre de requêtes de l'Application Load Balancer) ne décrivent pas suffisamment la charge de votre application.

Lorsque vous créez une politique de dimensionnement prédictif avec des métriques personnalisées, vous pouvez spécifier d'autres CloudWatch métriques fournies par AWS, ou vous pouvez spécifier des métriques que vous définissez et publiez vous-même. Vous pouvez également utiliser les mathématiques des métriques pour agréger et transformer les métriques existantes en une nouvelle série chronologique qui AWS n'est pas automatiquement suivie. Lorsque vous combinez des valeurs

dans vos données, par exemple, en calculant de nouvelles sommes ou moyennes, cela s'appelle l'agrégation. Les données résultantes sont appelées un agrégat.

La section suivante contient les bonnes pratiques et des exemples de construction de la structure JSON pour la politique.

Rubriques

Bonnes pratiques

Les bonnes pratiques suivantes peuvent vous aider à utiliser plus efficacement les métriques personnalisées :

- Pour la spécification de la métrique de charge, la métrique la plus utile est une métrique qui représente la charge d'un groupe Auto Scaling dans son ensemble, indépendamment de la capacité du groupe.
- Pour la spécification de la métrique de mise à l'échelle, la métrique la plus utile pour la mise à l'échelle est une métrique moyenne de débit ou d'utilisation par instance.
- La métrique de mise à l'échelle doit être inversement proportionnelle à la capacité. C'est-à-dire que si le nombre d'instances dans le groupe Auto Scaling augmente, la métrique de mise à l'échelle doit diminuer à peu près dans la même proportion. Pour que la mise à l'échelle prédictive se comporte comme prévu, la métrique de charge et la métrique de mise à l'échelle doivent également présenter une forte corrélation entre elles.
- L'utilisation cible doit correspondre au type de métrique de mise à l'échelle. Pour une configuration de politique qui utilise l'utilisation du CPU, il s'agit d'un pourcentage cible. Pour une configuration de politique qui utilise le débit, tel que le nombre de demandes ou de messages, il s'agit du nombre cible de demandes ou de messages par instance pendant tout intervalle d'une minute.
- Si ces recommandations ne sont pas suivies, les valeurs futures prédites des séries temporelles seront probablement incorrectes. Pour valider que les données sont correctes, vous pouvez visualiser les valeurs prédites dans la console Amazon EC2 Auto Scaling. Sinon, après avoir

créé votre politique de dimensionnement prédictif, inspectez les `CapacityForecast` objets `LoadForecast` et renvoyés par un appel à l'GetPredictiveScalingForecastAPI.

- Nous vous recommandons vivement de configurer la mise à l'échelle prédictive en mode prévision uniquement pour pouvoir évaluer la prévision avant que la mise à l'échelle prédictive ne commence à mettre activement à l'échelle la capacité.

Prérequis

Pour ajouter des métriques personnalisées à votre politique de mise à l'échelle, vous devez disposer des autorisations `cloudwatch:GetMetricData`.

Pour spécifier vos propres indicateurs au lieu des indicateurs AWS fournis, vous devez d'abord les publier sur CloudWatch. Pour plus d'informations, consultez la section Publication de métriques personnalisées dans le guide de CloudWatch l'utilisateur Amazon.

Si vous publiez vos propres métriques, veillez à publier les points de données à une fréquence minimale de cinq minutes. Amazon EC2 Auto Scaling extrait les points de données en CloudWatch fonction de la durée de la période dont il a besoin. Par exemple, la spécification des métriques de charge utilise des métriques horaires pour mesurer la charge de votre application. CloudWatch utilise vos données métriques publiées pour fournir une valeur de données unique pour toute période d'une heure en agrégeant tous les points de données avec des horodatages correspondant à chaque période d'une heure.

Construction du fichier JSON pour les métriques personnalisées

La section suivante contient des exemples de configuration de la mise à l'échelle prédictive pour interroger des données CloudWatch. Il existe deux méthodes pour configurer cette option, qui affecteront le format utilisé pour créer le fichier JSON de votre politique de mise à l'échelle prédictive. Lorsque vous utilisez des mathématiques de métriques, le format du fichier JSON varie davantage en fonction des mathématiques de métriques effectuées.

1. Pour créer une politique qui obtient des données directement à partir d'autres CloudWatch indicateurs fournis par AWS ou sur lesquels vous publiez CloudWatch, voirExemple de politique de mise à l'échelle prédictive avec des métriques de charge et de mise à l'échelle personnalisées (AWS CLI).
2. Pour créer une politique capable d'interroger plusieurs CloudWatch mesures et d'utiliser des expressions mathématiques pour créer de nouvelles séries chronologiques basées sur ces mesures, voirUtiliser des expressions mathématiques de métrique.

Exemple de politique de mise à l'échelle prédictive avec des métriques de charge et de mise à l'échelle personnalisées (AWS CLI)

Pour créer une politique de dimensionnement prédictive avec des métriques de charge et de dimensionnement personnalisées avec le AWS CLI, stockez les arguments pour `--predictive-scaling-configuration` dans un fichier JSON nomméconfig.json.

Vous commencez par ajouter des métriques personnalisées en remplaçant les valeurs remplaçables de l'exemple suivant par celles de vos métriques et de votre utilisation cible.

```
{
  "MetricSpecifications": [
    {
      "TargetValue": 50,
      "CustomizedScalingMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "scaling_metric",
            "MetricStat": {
              "Metric": {
                "MetricName": "MyUtilizationMetric",
                "Namespace": "MyNameSpace",
                "Dimensions": [
                  {
                    "Name": "MyOptionalMetricDimensionName",
                    "Value": "MyOptionalMetricDimensionValue"
                  }
                ]
              },
              "Stat": "Average"
            }
          }
        ]
      },
      "CustomizedLoadMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "load_metric",
            "MetricStat": {
              "Metric": {
                "MetricName": "MyLoadMetric",
                "Namespace": "MyNameSpace",
                "Dimensions": [
```

```
                {
                  "Name": "MyOptionalMetricDimensionName",
                  "Value": "MyOptionalMetricDimensionValue"
                }
              ]
            },
            "Stat": "Sum"
          }
        }
      ]
    }
  ]
}
```

Pour plus d'informations, consultez le MetricDataQuerymanuel Amazon EC2 Auto Scaling API Reference.

> **Note**
>
> Voici quelques ressources supplémentaires qui peuvent vous aider à trouver des noms de métriques, des espaces de noms, des dimensions et des statistiques pour les CloudWatch métriques :
>
> - Pour plus d'informations sur les métriques disponibles pour les AWS services, consultez les AWS services qui publient CloudWatch des métriques dans le guide de CloudWatch l'utilisateur Amazon.
> - Pour obtenir le nom, l'espace de noms et les dimensions exacts (le cas échéant) d'une CloudWatch métrique avec le AWS CLI, consultez list-metrics.

Pour créer cette politique, exécutez la put-scaling-policycommande en utilisant le fichier JSON comme entrée, comme illustré dans l'exemple suivant.

```
aws autoscaling put-scaling-policy --policy-name my-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
```

Si elle aboutit, cette commande renvoie l'Amazon Resource Name (ARN) de la stratégie.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:2f4f5048-d8a8-4d14-
b13a-d1905620f345:autoScalingGroupName/my-asg:policyName/my-predictive-scaling-policy",
  "Alarms": []
}
```

Utiliser des expressions mathématiques de métrique

La section suivante fournit des informations et des exemples de politiques de mise à l'échelle prédictive qui montrent comment vous pouvez utiliser les mathématiques de métriques dans votre politique.

Rubriques

- Comprendre les mathématiques de métrique
- Exemple de politique de mise à l'échelle prédictive qui combine des métriques à l'aide des mathématiques de métriques (AWS CLI)
- Exemple de politique de mise à l'échelle prédictive à utiliser dans un scénario de déploiement bleu/vert (AWS CLI)

Comprendre les mathématiques de métrique

Si vous souhaitez simplement agréger des données métriques existantes, les mathématiques CloudWatch métriques vous évitent les efforts et les coûts liés à la publication d'une autre métrique dans CloudWatch. Vous pouvez utiliser n'importe quelle métrique qui AWS fournit, et vous pouvez également utiliser des métriques que vous définissez dans le cadre de vos applications. Par exemple, vous pourriez vouloir calculer le backlog de la file d'attente Amazon SQS par instance. Vous pouvez le faire en prenant le nombre approximatif de messages disponibles pour la récupération de la file d'attente et en divisant ce nombre par la capacité d'exécution du groupe Auto Scaling.

Pour plus d'informations, consultez la section Utilisation des mathématiques métriques dans le guide de CloudWatch l'utilisateur Amazon.

Si vous choisissez d'utiliser une expression mathématique de métrique dans votre politique de mise à l'échelle prédictive, tenez compte des points suivants :

- Les opérations mathématiques de métrique utilisent les points de données de la combinaison unique de nom de la métrique, d'espace de noms et de paires clé/valeur de dimension des métriques.

- Vous pouvez utiliser n'importe quel opérateur arithmétique (+ - */^), fonction statistique (telle que AVG ou SUM) ou toute autre fonction compatible. CloudWatch
- Vous pouvez utiliser à la fois des métriques et les résultats d'autres expressions mathématiques dans les formules de l'expression mathématique.
- Vos expressions mathématiques de métrique peuvent être composées de différentes agrégations. Cependant, une bonne pratique pour le résultat final de l'agrégation consiste à utiliser `Average` pour la métrique de mise à l'échelle et `Sum` pour la métrique de charge.
- Toutes les expressions utilisées dans une spécification de métrique doivent finalement retourner une seule séries temporelles.

Pour utiliser les mathématiques de métrique, procédez comme suit :

- Choisissez un ou plusieurs CloudWatch indicateurs. Créez ensuite l'expression. Pour plus d'informations, consultez la section Utilisation des mathématiques métriques dans le guide de CloudWatch l'utilisateur Amazon.
- Vérifiez que l'expression mathématique de la métrique est valide à l'aide de la CloudWatch console ou de l' CloudWatch GetMetricDataAPI.

Exemple de politique de mise à l'échelle prédictive qui combine des métriques à l'aide des mathématiques de métriques (AWS CLI)

Parfois, au lieu de spécifier la métrique directement, vous devrez d'abord traiter ses données d'une certaine manière. Par exemple, une application peut extraire le travail d'une file d'attente Amazon SQS et vous souhaitez utiliser le nombre d'éléments dans la file d'attente comme critère de mise à l'échelle prédictive. Le nombre de messages dans la file d'attente ne définit pas uniquement le nombre d'instances dont vous avez besoin. Par conséquent, un travail supplémentaire est nécessaire pour créer une métrique qui peut être utilisée pour calculer le backlog par instance. Pour plus d'informations, consultez Mise à l'échelle basée sur Amazon SQS.

Ce qui suit est un exemple de politique de mise à l'échelle prédictive pour ce scénario. Il spécifie les métriques de mise à l'échelle et de charge qui sont basées sur la métrique `ApproximateNumberOfMessagesVisible` d'Amazon SQS, qui est le nombre de messages disponibles pour la récupération de la file d'attente. Il utilise également la métrique `GroupInServiceInstances` d'Amazon EC2 Auto Scaling et une expression mathématique pour calculer le backlog par instance pour la métrique de mise à l'échelle.

```
aws autoscaling put-scaling-policy --policy-name my-sqs-custom-metrics-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
{
  "MetricSpecifications": [
    {
      "TargetValue": 100,
      "CustomizedScalingMetricSpecification": {
        "MetricDataQueries": [
          {
            "Label": "Get the queue size (the number of messages waiting to be
 processed)",
            "Id": "queue_size",
            "MetricStat": {
              "Metric": {
                "MetricName": "ApproximateNumberOfMessagesVisible",
                "Namespace": "AWS/SQS",
                "Dimensions": [
                  {
                    "Name": "QueueName",
                    "Value": "my-queue"
                  }
                ]
              },
              "Stat": "Sum"
            },
            "ReturnData": false
          },
          {
            "Label": "Get the group size (the number of running instances)",
            "Id": "running_capacity",
            "MetricStat": {
              "Metric": {
                "MetricName": "GroupInServiceInstances",
                "Namespace": "AWS/AutoScaling",
                "Dimensions": [
                  {
                    "Name": "AutoScalingGroupName",
                    "Value": "my-asg"
                  }
                ]
              },
              "Stat": "Sum"
```

```
            },
            "ReturnData": false
          },
          {
            "Label": "Calculate the backlog per instance",
            "Id": "scaling_metric",
            "Expression": "queue_size / running_capacity",
            "ReturnData": true
          }
        ]
      },
      "CustomizedLoadMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "load_metric",
            "MetricStat": {
              "Metric": {
                "MetricName": "ApproximateNumberOfMessagesVisible",
                "Namespace": "AWS/SQS",
                "Dimensions": [
                  {
                    "Name": "QueueName",
                    "Value": "my-queue"
                  }
                ],
              },
              "Stat": "Sum"
            },
            "ReturnData": true
          }
        ]
      }
    }
  ]
}
```

L'exemple renvoie l'ARN de la politique.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:2f4f5048-d8a8-4d14-
b13a-d1905620f345:autoScalingGroupName/my-asg:policyName/my-sqs-custom-metrics-policy",
  "Alarms": []
}
```

Exemple de politique de mise à l'échelle prédictive à utiliser dans un scénario de déploiement bleu/vert (AWS CLI)

Une expression de recherche fournit une option avancée dans laquelle vous pouvez demander une métrique à partir de plusieurs groupes Auto Scaling et effectuer des expressions mathématiques sur eux. Ceci est particulièrement utile pour les déploiements bleu/vert.

> **Note**
>
> Un déploiement bleu/vert est une méthode de déploiement dans laquelle vous créez deux groupes Auto Scaling distincts mais identiques. Seul l'un des groupes reçoit le trafic de production. Le trafic utilisateur est initialement dirigé vers le groupe Auto Scaling précédent (« bleu »), tandis qu'un nouveau groupe (« vert ») est utilisé pour le test et l'évaluation d'une nouvelle version d'une application ou d'un service. Le trafic utilisateur est transféré vers le groupe Auto Scaling vert après qu'un nouveau déploiement ait été testé et accepté. Vous pouvez ensuite supprimer le groupe bleu après le succès du déploiement.

Lorsque de nouveaux groupes Auto Scaling sont créés dans le cadre d'un déploiement bleu/vert, l'historique des métriques de chaque groupe peut être automatiquement inclus dans la politique de mise à l'échelle prédictive sans que vous ayez à modifier ses spécifications de métrique. Pour plus d'informations, consultez la section Utilisation des politiques de dimensionnement prédictif d'EC2 Auto Scaling avec des déploiements bleu/vert sur le Compute Blog. AWS

L'exemple de politique suivant montre comment cela peut être fait. Dans cet exemple, la politique utilise la métrique `CPUUtilization` émise par Amazon EC2. Elle utilise la métrique `GroupInServiceInstances` d'Amazon EC2 Auto Scaling et une expression mathématique pour calculer la valeur de la métrique de mise à l'échelle par instance. Elle spécifie également une métrique de capacité pour obtenir la métrique `GroupInServiceInstances`.

L'expression de recherche trouve la `CPUUtilization` des instances dans plusieurs groupes Auto Scaling en fonction des critères de recherche spécifiés. Si vous créez ultérieurement un nouveau groupe Auto Scaling qui correspond aux mêmes critères de recherche, `CPUUtilization` des instances dans le nouveau groupe Auto Scaling est automatiquement incluse.

```
aws autoscaling put-scaling-policy --policy-name my-blue-green-predictive-scaling-
policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
```

```
{
  "MetricSpecifications": [
    {
      "TargetValue": 25,
      "CustomizedScalingMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "load_sum",
            "Expression": "SUM(SEARCH('{AWS/EC2,AutoScalingGroupName} MetricName=
\"CPUUtilization\" ASG-myapp', 'Sum', 300))",
            "ReturnData": false
          },
          {
            "Id": "capacity_sum",
            "Expression": "SUM(SEARCH('{AWS/AutoScaling,AutoScalingGroupName}
 MetricName=\"GroupInServiceInstances\" ASG-myapp', 'Average', 300))",
            "ReturnData": false
          },
          {
            "Id": "weighted_average",
            "Expression": "load_sum / capacity_sum",
            "ReturnData": true
          }
        ]
      },
      "CustomizedLoadMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "load_sum",
            "Expression": "SUM(SEARCH('{AWS/EC2,AutoScalingGroupName} MetricName=
\"CPUUtilization\" ASG-myapp', 'Sum', 3600))"
          }
        ]
      },
      "CustomizedCapacityMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "capacity_sum",
            "Expression": "SUM(SEARCH('{AWS/AutoScaling,AutoScalingGroupName}
 MetricName=\"GroupInServiceInstances\" ASG-myapp', 'Average', 300))"
          }
        ]
      }
    }
```

```
  ]
}
```

L'exemple renvoie l'ARN de la politique.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:2f4f5048-d8a8-4d14-
b13a-d1905620f345:autoScalingGroupName/my-asg:policyName/my-blue-green-predictive-
scaling-policy",
  "Alarms": []
}
```

Considérations et dépannage

Si un problème survient lors de l'utilisation de métriques personnalisées, nous vous recommandons d'effectuer les opérations suivantes :

- Si un message d'erreur est fourni, lisez le message et résolvez le problème qu'il signale, si possible.
- Si un problème survient lorsque vous essayez d'utiliser une expression de recherche dans un scénario de déploiement bleu/vert, assurez-vous d'abord de comprendre comment créer une expression de recherche qui recherche une correspondance partielle au lieu d'une correspondance exacte. Vérifiez également que votre requête ne trouve que les groupes Auto Scaling qui exécutent l'application spécifique. Pour plus d'informations sur la syntaxe des expressions de recherche, consultez la section Syntaxe des expressions de CloudWatch recherche dans le guide de CloudWatch l'utilisateur Amazon.
- Si vous n'avez pas validé une expression à l'avance, la put-scaling-policycommande la valide lorsque vous créez votre politique de dimensionnement. Cependant, il est possible que cette commande ne parvienne pas à identifier la cause exacte des erreurs détectées. Pour résoudre les problèmes, corrigez les erreurs que vous recevez en réponse à une demande de get-metric-datacommande. Vous pouvez également résoudre les problèmes liés à l'expression depuis la CloudWatch console.
- Lorsque vous affichez vos graphiques de charge et de capacité dans la console, il se peut que le graphique de capacité n'affiche aucune donnée. Pour vous assurer que les graphiques contiennent des données complètes, veillez à activer systématiquement les métriques de groupe pour vos groupes Auto Scaling. Pour plus d'informations, consultez Activer les métriques du groupe Auto Scaling (console).

- La spécification de la métrique de capacité n'est utile que pour les déploiements bleu/vert lorsque vous avez des applications qui s'exécutent dans différents groupes Auto Scaling au cours de leur durée de vie. Cette métrique personnalisée vous permet de fournir la capacité totale de plusieurs groupes Auto Scaling. La mise à l'échelle prédictive l'utilise pour afficher des données historiques dans les graphiques de capacité de la console.
- Vous devez spécifier `false` pour `ReturnData` si `MetricDataQueries` spécifie la fonction SEARCH() seule sans une fonction mathématique comme SUM(). Cela est dû au fait que les expressions de recherche peuvent renvoyer plusieurs séries temporelles et qu'une spécification métrique basée sur une expression ne peut renvoyer qu'une seule séries temporelles.
- Toutes les métriques impliquées dans une expression de recherche doivent avoir la même résolution.

Limites

- Vous pouvez interroger des points de données de 10 métriques au maximum dans une spécification métrique.
- Dans le cadre de cette limite, une expression compte pour une métrique.

Contrôler les instances à scalabilité automatique à résilier pendant une mise à l'échelle horizontale

Amazon EC2 Auto Scaling utilise des politiques de résiliation pour décider de l'ordre de résiliation des instances. Vous pouvez utiliser une politique prédéfinie ou créer une politique personnalisée pour répondre à vos besoins spécifiques. En utilisant une politique personnalisée ou une protection intégrée des instances, vous pouvez également empêcher votre groupe Auto Scaling de mettre fin à des instances qui ne sont pas encore prêtes à le faire.

Table des matières

Quand Amazon EC2 Auto Scaling utilise des politiques de résiliation

Les sections suivantes décrivent les scénarios dans lesquels Amazon EC2 Auto Scaling utilise des politiques de résiliation.

Table des matières

Événements de mise à l'échelle horizontale

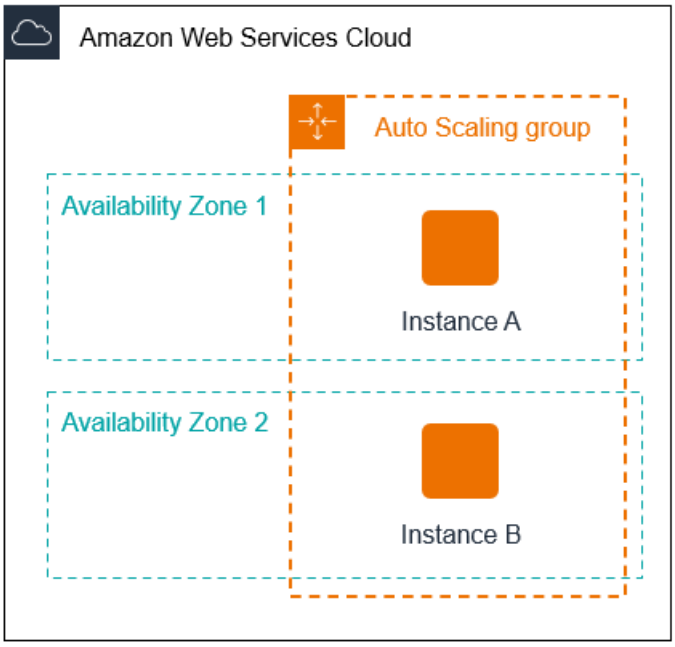
Un événement de mise à l'échelle horizontale se produit lorsqu'une nouvelle valeur pour la capacité souhaitée d'un groupe de mise à l'échelle automatique est inférieure à la capacité actuelle du groupe.

Les événements de mise à l'échelle horizontale se produisent dans les scénarios suivants :

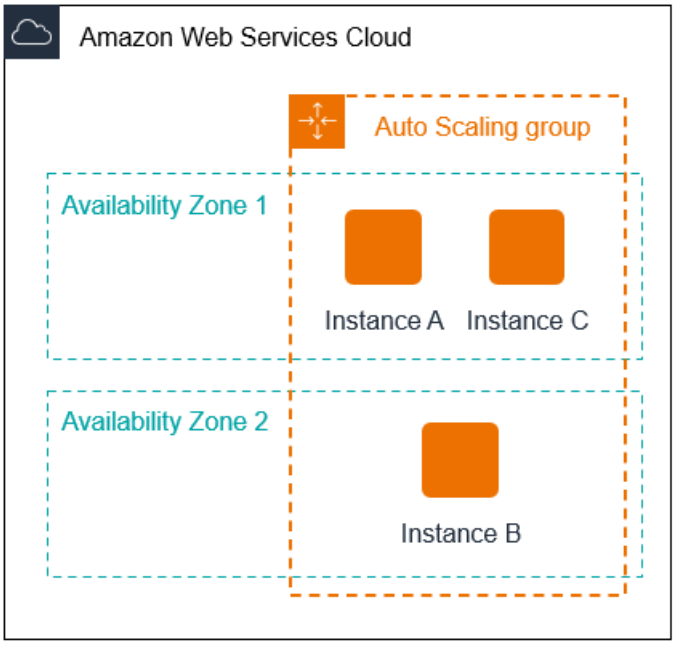
- Lorsque vous utilisez des politiques de mise à l'échelle dynamique et que la taille du groupe diminue à la suite de modifications de la valeur d'une mesure
- Lorsque vous utilisez une mise à l'échelle planifiée et que la taille du groupe diminue à la suite d'une action planifiée
- Lorsque vous réduisez manuellement la taille du groupe

L'exemple suivant montre comment fonctionnent les politiques de résiliation lorsqu'il y a un événement évolutif.

1. Dans cet exemple, le groupe Auto Scaling possède un type d'instance, deux zones de disponibilité et une capacité souhaitée de deux instances. Il dispose également d'une politique de mise à l'échelle dynamique qui ajoute et supprime des instances lorsque l'utilisation des ressources augmente ou diminue. Les deux instances de ce groupe sont réparties entre les deux zones de disponibilité, comme dans le schéma suivant.

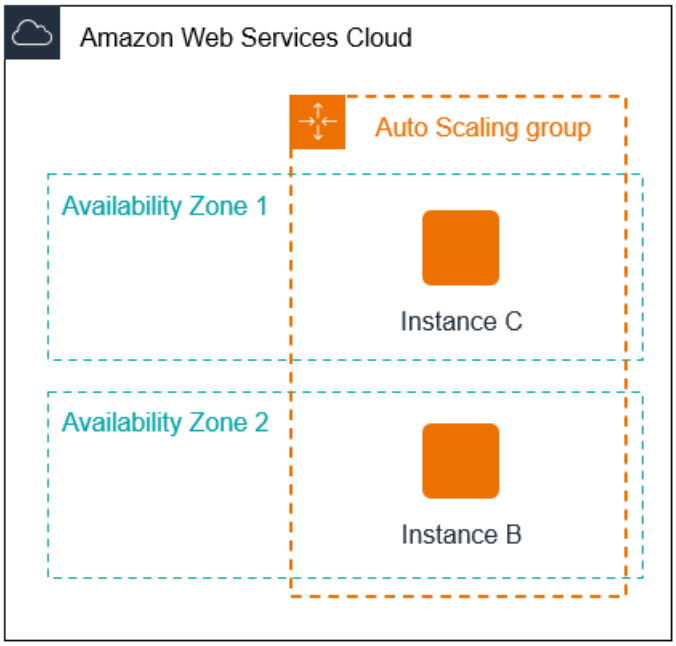


2. Lorsque le groupe Auto Scaling augmente, Amazon EC2 Auto Scaling lance une nouvelle instance. Le groupe Auto Scaling comporte désormais trois instances, réparties entre les deux zones de disponibilité, comme indiqué dans le diagramme suivant.



3. Lorsque le groupe Auto Scaling est mis à l'échelle, Amazon EC2 Auto Scaling résilie l'une des instances.
4. Si vous n'avez pas attribué de politique de mise hors service spécifique au groupe, Amazon EC2 Auto Scaling utilisera la politique de mise hors service par défaut. Il sélectionne la zone de disponibilité avec deux instances et met fin à l'instance qui a été lancée à partir d'une configuration

de lancement, d'un modèle de lancement différent ou de la version la plus ancienne du modèle de lancement actuel. Si les instances ont été lancées à partir du même modèle de lancement et de la même version, Amazon EC2 Auto Scaling sélectionne l'instance la plus proche de l'heure de facturation suivante et y met fin.



Actualisation d'instance

Vous pouvez lancer une actualisation des instances pour mettre à jour les instances de votre groupe Auto Scaling. Au cours d'une actualisation d'instance, Amazon EC2 Auto Scaling résilie les instances du groupe, puis lance les remplacements pour celles qui ont été résiliées. La politique de résiliation du groupe Auto Scaling contrôle les instances remplacées en premier.

Rééquilibrage des zones de disponibilité

Amazon EC2 Auto Scaling équilibre la capacité dans les zones de disponibilité activées pour le groupe Auto Scaling. Cela permet de réduire l'impact d'une panne de la zone de disponibilité. Si la distribution de la capacité entre les zones de disponibilité devient déséquilibrée, Amazon EC2 Auto Scaling rééquilibre le groupe Auto Scaling en lançant des instances dans les zones de disponibilité activées avec le moins d'instances et en résiliant des instances ailleurs. La politique de résiliation détermine quelles instances sont résiliées en premier.

Il existe plusieurs raisons pour lesquelles la distribution des instances entre les zones de disponibilité peut se déséquilibrer.

Suppression d'instances

Si vous détachez des instances de votre groupe Auto Scaling, si vous placez des instances en attente ou si vous mettez fin explicitement à des instances et décrémentez la capacité souhaitée, empêchant le lancement d'instances de remplacement, le groupe peut devenir déséquilibré. Dans ce cas, Amazon EC2 Auto Scaling compense en rééquilibrant les zones de disponibilité.

Utilisation de zones de disponibilité différentes de celles initialement spécifiées

Si vous développez votre groupe Auto Scaling de manière à inclure des zones de disponibilité supplémentaires, ou si vous modifiez les zones de disponibilité utilisées, Amazon EC2 Auto Scaling lance des instances dans les nouvelles zones de disponibilité et résilie des instances dans les autres zones pour vous assurer que votre groupe Auto Scaling est réparti uniformément entre les zones de disponibilité.

Pannes de disponibilité

Les pannes de disponibilité sont rares. Toutefois, si une zone de disponibilité devient indisponible et est restaurée ultérieurement, votre groupe Auto Scaling peut être déséquilibré entre les zones de disponibilité. Amazon EC2 Auto Scaling essaie de rééquilibrer progressivement le groupe et le rééquilibrage peut résilier les instances dans d'autres zones.

Par exemple, imaginons que vous avez un groupe Auto Scaling avec un type d'instance, deux zones de disponibilité et une capacité souhaitée de deux instances. Dans l'hypothèse où une zone de disponibilité échoue, Amazon EC2 Auto Scaling lance automatiquement une nouvelle instance dans la zone de disponibilité saine pour remplacer celle de la zone de disponibilité défectueuse. Ainsi, lorsque la zone de disponibilité défectueuse redevient saine ultérieurement, Amazon EC2 Auto Scaling lance automatiquement une nouvelle instance dans cette zone, ce qui met fin à une instance dans la zone non affectée.

Note

Lors du rééquilibrage, Amazon EC2 Auto Scaling lance de nouvelles instances avant de résilier les anciennes, afin que le rééquilibrage ne compromette pas les performances ou la disponibilité de l'application.
Comme Amazon EC2 Auto Scaling tente de lancer de nouvelles instances avant de résilier les anciennes, le fait d'atteindre ou de s'approcher de la capacité maximale spécifiée peut entraver ou arrêter les activités de rééquilibrage. Pour éviter ce problème, le système peut temporairement dépasser la capacité maximum spécifiée d'un groupe d'une marge de 10 pour cent (ou d'une marge de 1 instance, la plus importante des deux) pendant l'activité

de rééquilibrage. La marge est étendue uniquement si le groupe atteint ou s'approche de la capacité maximum et nécessite un rééquilibrage, à cause d'une demande de modification du zonage par l'utilisateur ou pour compenser des problèmes de zone de disponibilité. L'extension dure uniquement le temps de rééquilibrer le groupe.

Configurer les politiques de résiliation pour Amazon EC2 Auto Scaling

Une politique de résiliation fournit les critères qu'Amazon EC2 Auto Scaling suit pour mettre fin aux instances dans un ordre spécifique.

Par défaut, Amazon EC2 Auto Scaling utilise une politique de résiliation conçue pour mettre fin d'abord aux instances qui utilisent des configurations obsolètes. Vous pouvez modifier la politique de résiliation afin de contrôler les instances les plus importantes à résilier en premier.

Lorsqu'Amazon EC2 Auto Scaling met fin à des instances, il essaie de maintenir l'équilibre entre les zones de disponibilité activées pour votre groupe Auto Scaling. Le maintien de l'équilibre zonal prime sur la politique de résiliation. Si une zone de disponibilité possède plus d'instances que d'autres, Amazon EC2 Auto Scaling applique d'abord la politique de résiliation à la zone déséquilibrée. Si les zones de disponibilité sont équilibrées, la politique de résiliation s'applique à toutes les zones.

Rubriques

Comment fonctionne la politique de résiliation par défaut

Lorsqu'Amazon EC2 Auto Scaling doit mettre fin à une instance, il identifie d'abord la zone de disponibilité (ou les zones) qui compte le plus d'instances et au moins une instance qui n'est pas protégée contre le scaling in. Il procède ensuite à l'évaluation des instances non protégées dans la zone de disponibilité identifiée comme suit :

Instances utilisant des configurations obsolètes

- Pour les groupes qui utilisent un modèle de lancement : déterminez si l'une des instances utilise des configurations obsolètes, en hiérarchisant dans cet ordre :

1. Vérifiez d'abord les instances lancées avec une configuration de lancement.
2. Vérifiez ensuite si les instances ont été lancées à l'aide d'un modèle de lancement différent du modèle de lancement actuel.
3. Enfin, vérifiez les instances utilisant la version la plus ancienne du modèle de lancement actuel.

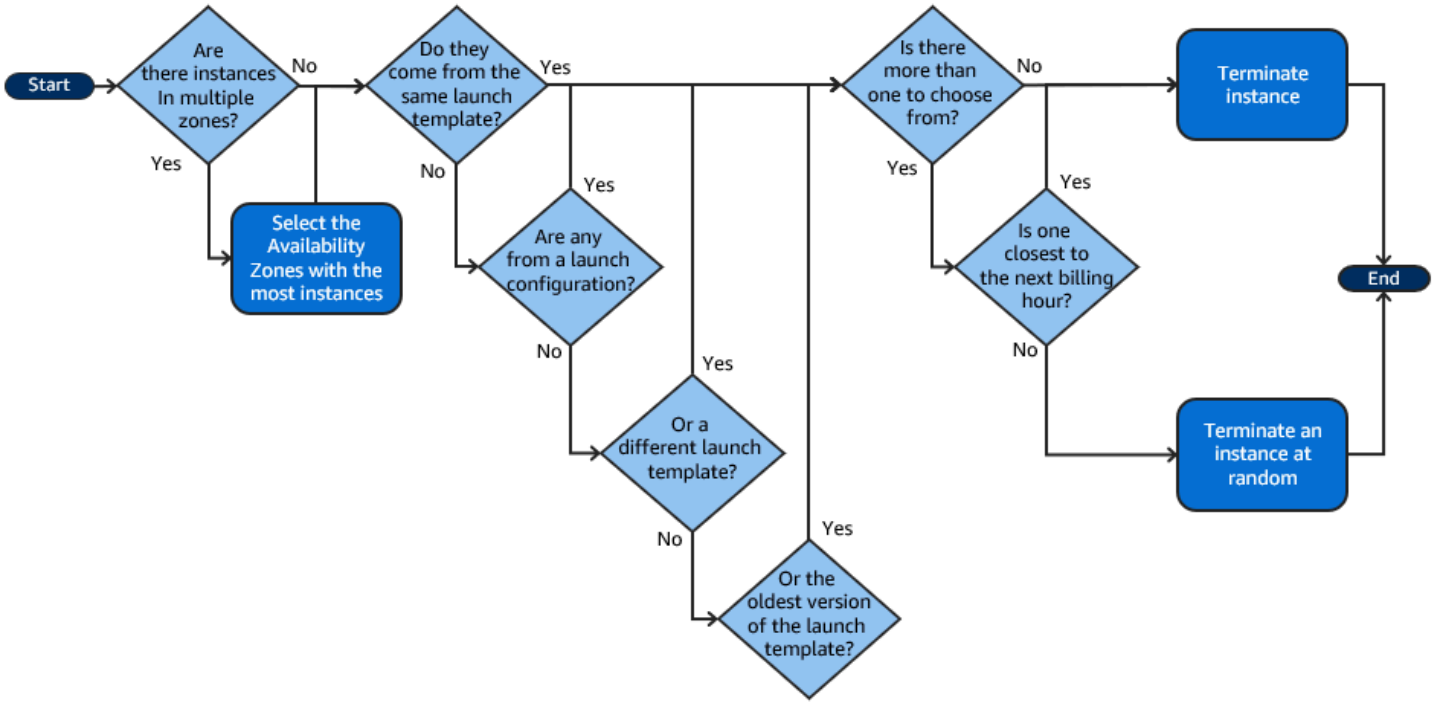
- Pour les groupes utilisant une configuration de lancement : déterminez si l'une des instances utilise la configuration de lancement la plus ancienne.

Si aucune instance avec des configurations obsolètes n'est trouvée, ou s'il existe plusieurs instances parmi lesquelles choisir, Amazon EC2 Auto Scaling prend en compte le prochain critère selon lequel les instances approchent de leur prochaine heure de facturation.

Instances approchant de l'heure de facturation suivante

Déterminez si l'une des instances répondant aux critères précédents est la plus proche de l'heure de facturation suivante. Si plusieurs instances sont également proches, arrêtez-en une au hasard. Cela vous permet d'optimiser l'utilisation de vos instances facturées à l'heure. Cependant, la majeure partie de l'utilisation de l'EC2 est désormais facturée à la seconde, de sorte que cette optimisation présente moins d'avantages. Pour plus d'informations, consultez Tarification Amazon EC2.

L'organigramme suivant illustre le fonctionnement de la politique de résiliation par défaut pour les groupes qui utilisent un modèle de lancement.



Politique de résiliation par défaut et groupes d'instances mixtes

Amazon EC2 Auto Scaling applique des critères supplémentaires lors de la résiliation d'instances dans des groupes d'instances mixtes.

Lorsqu'Amazon EC2 Auto Scaling doit mettre fin à une instance, il identifie d'abord quelle option d'achat (ponctuelle ou à la demande) doit être résiliée en fonction des paramètres du groupe. Cela garantit que le groupe évolue vers le ratio spécifié d'instances ponctuelles et à la demande au fil du temps.

Il applique ensuite la politique de résiliation indépendamment au sein de chaque zone de disponibilité. Il détermine quelle instance ponctuelle ou à la demande dans quelle zone de disponibilité mettre fin afin de maintenir l'équilibre des zones de disponibilité. La même logique s'applique à un groupe d'instances mixte avec des pondérations définies pour les types d'instances.

Dans chaque zone, la politique de résiliation par défaut fonctionne comme suit pour déterminer quelle instance non protégée peut être résiliée dans le cadre de l'option d'achat identifiée :

1. Déterminez si l'une des instances peut être interrompue afin d'améliorer l'alignement avec la stratégie d'allocation spécifiée pour le groupe Auto Scaling. Si aucune instance n'est identifiée pour l'optimisation, ou s'il existe plusieurs instances parmi lesquelles choisir, l'évaluation se poursuit.
2. Déterminez si l'une des instances utilise des configurations obsolètes, en hiérarchisant dans cet ordre :
   a. Vérifiez d'abord les instances lancées avec une configuration de lancement.
   b. Vérifiez ensuite si les instances ont été lancées à l'aide d'un modèle de lancement différent du modèle de lancement actuel.
   c. Enfin, vérifiez les instances utilisant la version la plus ancienne du modèle de lancement actuel.

   Si aucune instance avec des configurations obsolètes n'est trouvée, ou s'il existe plusieurs instances parmi lesquelles choisir, l'évaluation se poursuit.
3. Déterminez si l'une des instances est la plus proche de l'heure de facturation suivante. Si plusieurs instances sont également proches, choisissez-en une au hasard.

Politiques de résiliation prédéfinies

Vous avez le choix entre les politiques de résiliation prédéfinies suivantes :

- **`Default`**— Mettez fin aux instances conformément à la politique de résiliation par défaut.
- **`AllocationStrategy`**— Mettez fin aux instances du groupe Auto Scaling pour aligner les instances restantes sur la stratégie d'allocation pour le type d'instance en cours de résiliation (instance ponctuelle ou instance à la demande). Cette politique est utile lorsque vous préférez les types d'instances ayant changé. Si la politique d'allocation Spot est `lowest-price`, vous pouvez progressivement rééquilibrer la distribution d'instances Spot entre vos groupes Spot ayant le prix le plus bas. Si la politique d'allocation Spot est `capacity-optimized`, vous pouvez progressivement rééquilibrer la distribution d'instances Spot entre vos groupes Spot ayant la capacité Spot disponible.la plus élevée. Vous pouvez également remplacer progressivement les instances à la demande ayant un type de priorité inférieur par les instances à la demande ayant un type de priorité supérieur.
- **`OldestLaunchTemplate`**— Mettez fin aux instances dont le modèle de lancement est le plus ancien. Avec cette politique, les instances qui utilisent le modèle de lancement non courant sont mises hors service en premier, suivies des instances qui utilisent la plus ancienne version du modèle de lancement actuel. Cette politique est utile lorsque vous mettez à jour un groupe et supprimez les instances d'une configuration précédente.
- **`OldestLaunchConfiguration`**— Mettez fin aux instances dont la configuration de lancement est la plus ancienne. Cette politique est utile lorsque vous mettez à jour un groupe et supprimez les instances d'une configuration précédente. Avec cette politique, les instances qui utilisent la configuration de lancement non courant sont mises hors service en premier.
- **`ClosestToNextInstanceHour`**— Résiliez les instances les plus proches de l'heure de facturation suivante. Cette politique vous permet d'optimiser l'utilisation de vos instances qui ont un tarif horaire.
- **`NewestInstance`**— Mettez fin à l'instance la plus récente du groupe. Cette politique est utile lorsque vous testez une nouvelle configuration de lancement mais ne souhaitez pas la garder en production.
- **`OldestInstance`**— Mettez fin à la plus ancienne instance du groupe. Cette option est utile lorsque vous mettez à niveau les instances du groupe Auto Scaling vers un nouveau type d'instance EC2. Vous pouvez petit à petit remplacer les anciennes instances par des nouvelles.

> **Note**
>
> Amazon EC2 Auto Scaling équilibre toujours les instances entre les zones de disponibilité en premier, quelle que soit la politique de résiliation utilisée. Par conséquent, vous pouvez rencontrer des situations dans lesquelles certaines instances plus récentes sont résiliées avant les instances plus anciennes. Par exemple, lorsqu'une zone de disponibilité est

ajoutée plus récemment ou lorsqu'une zone de disponibilité possède plus d'instances que celles qui sont utilisées par le groupe.

Modifier la politique de résiliation d'un groupe Auto Scaling

Pour modifier la politique de résiliation de votre groupe Auto Scaling, appliquez l'une des méthodes suivantes.

Console

Vous ne pouvez pas modifier la politique de résiliation lorsque vous créez initialement un groupe Auto Scaling dans la console Amazon EC2 Auto Scaling. La politique de résiliation par défaut est utilisée automatiquement. Une fois votre groupe Auto Scaling créé, vous pouvez remplacer la politique par défaut par une autre politique de résiliation ou par plusieurs politiques de résiliation répertoriées dans l'ordre dans lequel elles doivent s'appliquer.

Pour modifier la politique de résiliation d'un groupe Auto Scaling

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Sous l'onglet Détails, choisissez Configurations avancées, Modifier.
4. Pour Termination policies (Politiques de résiliation), choisissez une ou plusieurs politiques de résiliation. Si vous choisissez plusieurs politiques, placez-les dans l'ordre dans lequel vous souhaitez qu'elles soient évaluées.

   Vous pouvez éventuellement choisir Custom termination policy (Politique de résiliation personnalisée), puis choisir une fonction Lambda qui répond à vos besoins. Si vous avez créé des versions et des alias pour votre fonction Lambda, vous pouvez choisir une version ou un alias dans la liste déroulante Version/Alias. Pour utiliser la version non publiée de votre fonction Lambda, laissez Version/Alias sur sa valeur par défaut. Pour plus d'informations, consultez Créer une politique de résiliation personnalisée avec Lambda.

   > **Note**
   >
   > Lorsque vous utilisez plusieurs politiques, leur ordre doit être défini correctement :

- Si vous utilisez la politique Par défaut, elle doit être la dernière de la liste.
- Si vous utilisez une Politique de résiliation personnalisée, elle doit être la première politique de la liste.

5. Choisissez Mettre à jour.

AWS CLI

La politique de résiliation par défaut est utilisée automatiquement sauf si une politique différente est spécifiée.

Pour modifier la politique de résiliation d'un groupe Auto Scaling

Utilisez l'une des commandes suivantes :

- create-auto-scaling-group
- update-auto-scaling-group

Vous pouvez utiliser individuellement les politiques de résiliation ou les combiner dans une liste de politiques. Par exemple, utilisez la commande suivante pour mettre à jour un groupe Auto Scaling afin d'utiliser la politique `OldestLaunchConfiguration` en premier lieu, puis la politique `ClosestToNextInstanceHour`.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg --
termination-policies "OldestLaunchConfiguration" "ClosestToNextInstanceHour"
```

Si vous utilisez la politique de résiliation `Default`, mettez-la en fin de liste. Par exemple, `--termination-policies "OldestLaunchConfiguration" "Default"`.

Pour utiliser une politique de résiliation personnalisée, vous devez d'abord créer votre politique de résiliation à l'aide de AWS Lambda. Pour spécifier la fonction Lambda à utiliser comme politique de résiliation, mettez-la en fin de liste. Par exemple, `--termination-policies "arn:aws:lambda:us-west-2:123456789012:function:HelloFunction:prod" "OldestLaunchConfiguration"`. Pour plus d'informations, consultez Créer une politique de résiliation personnalisée avec Lambda.

Créer une politique de résiliation personnalisée avec Lambda

Amazon EC2 Auto Scaling utilise des politiques de résiliation pour hiérarchiser les instances à résilier en premier lorsque vous diminuez la taille de votre groupe Auto Scaling (appelé Mise à l'échelle horizontale). Votre groupe Auto Scaling utilise une politique de résiliation par défaut, mais vous pouvez éventuellement choisir ou créer vos propres politiques de résiliation. Pour plus d'informations sur le choix d'une politique de résiliation prédéfinie, consultez Configurer les politiques de résiliation pour Amazon EC2 Auto Scaling.

Dans cette rubrique, vous allez apprendre à créer une politique de résiliation personnalisée à l'aide d'une fonction AWS Lambda qu'Amazon EC2 Auto Scaling appelle en réponse à certains événements. La fonction Lambda que vous créez traite les informations contenues dans les données d'entrée envoyées par Amazon EC2 Auto Scaling et renvoie une liste d'instances prêtes à être résiliées.

Une politique de résiliation personnalisée offre un meilleur contrôle sur les instances qui sont résiliées et à quel moment. Par exemple, lorsque votre groupe Auto Scaling est dimensionné, Amazon EC2 Auto Scaling ne peut pas déterminer si certaines charges de travail en cours d'exécution ne doivent pas être perturbées. Avec une fonction Lambda, vous pouvez valider la demande de résiliation et attendre que la charge de travail soit terminée avant de renvoyer l'ID d'instance à Amazon EC2 Auto Scaling pour la résiliation.

Table des matières

Données d'entrée

Amazon EC2 Auto Scaling génère une charge utile JSON pour les événements de mise à l'échelle horizontale, et le fait également lorsque les instances sont sur le point d'être résiliées en raison de la durée de vie maximale de l'instance ou des fonctions d'actualisation de l'instance. Il génère également une charge utile JSON pour les événements de mise à l'échelle horizontale qu'il peut initier lors du rééquilibrage de votre groupe entre les zones de disponibilité.

Cette charge utile contient des informations sur la capacité qu'Amazon EC2 Auto Scaling doit résilier, une liste des instances qu'il suggère pour la résiliation et l'événement qui a déclenché la résiliation.

Voici un exemple de charge utile :

```
{
  "AutoScalingGroupARN": "arn:aws:autoscaling:us-east-1:<account-
id>:autoScalingGroup:d4738357-2d40-4038-ae7e-b00ae0227003:autoScalingGroupName/my-asg",
  "AutoScalingGroupName": "my-asg",
  "CapacityToTerminate": [
    {
      "AvailabilityZone": "us-east-1b",
      "Capacity": 2,
      "InstanceMarketOption": "on-demand"
    },
    {
      "AvailabilityZone": "us-east-1b",
      "Capacity": 1,
      "InstanceMarketOption": "spot"
    },
    {
      "AvailabilityZone": "us-east-1c",
      "Capacity": 3,
      "InstanceMarketOption": "on-demand"
    }
  ],
  "Instances": [
    {
      "AvailabilityZone": "us-east-1b",
      "InstanceId": "i-0056faf8da3e1f75d",
      "InstanceType": "t2.nano",
      "InstanceMarketOption": "on-demand"
    },
    {
      "AvailabilityZone": "us-east-1c",
      "InstanceId": "i-02e1c69383a3ed501",
      "InstanceType": "t2.nano",
      "InstanceMarketOption": "on-demand"
    },
    {
      "AvailabilityZone": "us-east-1c",
      "InstanceId": "i-036bc44b6092c01c7",
      "InstanceType": "t2.nano",
      "InstanceMarketOption": "on-demand"
```

```
    },
    ...
  ],
  "Cause": "SCALE_IN"
}
```

La charge utile inclut le nom du groupe Auto Scaling, son nom Amazon Resource Name (ARN) et les éléments suivants :

- `CapacityToTerminate` décrit la quantité de votre capacité Spot ou à la demande définie pour être résiliée dans une zone de disponibilité donnée.
- `Instances` représente les instances qu'Amazon EC2 Auto Scaling suggère pour la résiliation en fonction des informations contenues dans `CapacityToTerminate`.
- `Cause` décrit l'événement qui a provoqué la résiliation : `SCALE_IN`, `INSTANCE_REFRESH`, `MAX_INSTANCE_LIFETIME` ou `REBALANCE`.

Les informations suivantes décrivent les facteurs les plus significatifs dans la façon dont Amazon EC2 Auto Scaling génère les `Instances` dans les données d'entrée :

- Le maintien de l'équilibre entre les zones de disponibilité est prioritaire lorsqu'une instance est résiliée en raison d'événements de mise à l'échelle horizontale et de résiliations basées sur l'actualisation d'instance. Par conséquent, si une zone de disponibilité compte plus d'instances que les autres zones de disponibilité utilisées par le groupe, les données d'entrée contiennent des instances qui ne peuvent être résiliées qu'à partir de la zone de disponibilité déséquilibrée. Si les zones de disponibilité utilisées par le groupe sont équilibrées, les données en entrée contiennent des instances de toutes les zones de disponibilité pour le groupe.
- Lors de l'utilisation d'une politique d'instances mixtes, le maintien de l'équilibre de vos capacités Spot et à la demande en fonction des pourcentages souhaités pour chaque option d'achat est également prioritaire. Nous identifions d'abord lequel des deux types (Spot ou à la demande) doit être résilié. Nous identifions ensuite les instances (dans le cadre de l'option d'achat identifiée) dans lesquelles nous pouvons résilier les zones de disponibilité qui permettront d'équilibrer les zones de disponibilité.

Données de réponse

Les données d'entrée et les données de réponse fonctionnent ensemble pour affiner la liste des instances à résilier.

Avec l'entrée donnée, la réponse de votre fonction Lambda devrait ressembler à l'exemple suivant :

```
{
  "InstanceIDs": [
    "i-02e1c69383a3ed501",
    "i-036bc44b6092c01c7",
    ...
  ]
}
```

Dans la réponse, `InstanceIDs` représentent les instances qui sont prêtes à être résiliées.

Vous pouvez également renvoyer un ensemble différent d'instances prêtes à être résiliées, ce qui remplace les instances dans les données d'entrée. Si aucune instance n'est prête à être résiliée lorsque votre fonction Lambda est invoquée, vous pouvez également choisir de ne pas renvoyer d'instances.

Si aucune instance n'est prête à être mise hors service, la réponse de votre fonction Lambda devrait ressembler à l'exemple suivant :

```
{
  "InstanceIDs": [ ]
}
```

Considérations

Tenez compte des éléments suivants lors de l'utilisation d'une politique de résiliation personnalisée :

- Le renvoi d'une instance en premier dans les données de réponse ne garantit pas sa résiliation. Si le nombre d'instances renvoyées lors de l'appel de votre fonction Lambda est supérieur au nombre requis, Amazon EC2 Auto Scaling évalue chaque instance en fonction des autres politiques de résiliation que vous avez spécifiées pour votre groupe Auto Scaling. Lorsqu'il existe plusieurs politiques de résiliation, il tente d'appliquer la politique de résiliation suivante dans la liste, et s'il y a plus d'instances que nécessaire pour la résiliation, il passe à la politique de résiliation suivante, et ainsi de suite. Si aucune autre politique de résiliation n'est spécifiée, la politique de résiliation par défaut est utilisée pour déterminer les instances à résilier.
- Si aucune instance n'est renvoyée ou si votre fonction Lambda expire, Amazon EC2 Auto Scaling attend peu de temps avant d'appeler à nouveau votre fonction. Pour tout événement de mise à l'échelle horizontale, il continue d'essayer tant que la capacité souhaitée du groupe est inférieure

à sa capacité actuelle. Pour les résiliations basées sur l'actualisation d'instance, il effectue de nouvelles tentatives pendant une heure. Après cela, s'il ne parvient toujours pas à résilier les instances, l'opération d'actualisation de l'instance échoue. Avec la durée de vie maximale de l'instance, Amazon EC2 Auto Scaling continue d'essayer de résilier l'instance qui est identifiée comme dépassant sa durée de vie maximale.

- Parce que votre fonction est retentée à plusieurs reprises, assurez-vous de tester et de corriger toutes les erreurs permanentes dans votre code avant d'utiliser une fonction Lambda comme politique de résiliation personnalisée.
- Si vous remplacez les données en entrée par votre propre liste d'instances à résilier et que la résiliation de ces instances déséquilibre les zones de disponibilité, Amazon EC2 Auto Scaling rééquilibre progressivement la distribution de la capacité entre les zones de disponibilité. Tout d'abord, il appelle votre fonction Lambda pour voir si des instances sont prêtes à être résiliées afin de déterminer s'il faut commencer le rééquilibrage. Si des instances sont prêtes à être résiliées, elles lancent d'abord de nouvelles instances. Lorsque le lancement des instances se termine, il détecte alors que la capacité actuelle de votre groupe est supérieure à la capacité désirée et lance un événement évolutif.
- Une politique de résiliation personnalisée n'affecte pas votre capacité à utiliser également une protection contre la mise à l'échelle horizontale des tâches pour empêcher la résiliation de certaines instances. Pour plus d'informations, consultez Utiliser la protection de la taille d'instance.

Créer la fonction Lambda

Commencez par créer la fonction Lambda afin que vous puissiez spécifier son Amazon Resource Name (ARN) dans les politiques de résiliation de votre groupe Auto Scaling.

Pour créer une fonction Lambda (console)

1. Ouvrez la page Functions (Fonctions) sur la console Lambda.
2. Dans la barre de navigation située en haut de l'écran, choisissez la même région que celle utilisée lorsque vous avez créé le groupe Auto Scaling.
3. Sélectionnez Create function (Créer une fonction), puis Author from scratch (Créer à partir de zéro).
4. Sous Informations de base, dans Nom de fonction, entrez le nom de votre fonction.
5. Sélectionnez Create function (Créer une fonction). Vous retournez au code et à la configuration de la fonction.

6. Avec votre fonction toujours ouverte dans la console, sous Function code (Code de fonction), collez votre code dans l'éditeur.
7. Choisissez Deploy (Déployer).
8. Vous pouvez également créer une version publiée de la fonction Lambda en choisissant l'option Versions, puis Publish new version (Publier une nouvelle version). Pour en savoir plus sur la gestion des versions dans Lambda, veuillez consulter Versions de fonctions Lambda dans le Guide du développeur AWS Lambda .
9. Si vous avez choisi de publier une version, choisissez l'option Alias si vous voulez associer un alias à cette version de la fonction Lambda. Pour plus d'informations sur l'utilisation des alias Lambda, consultez Alias de fonction Lambda dans le Guide du développeur AWS Lambda .
10. Ensuite, choisissez l'onglet Configuration, puis Permissions (Autorisations).
11. Faites défiler l'écran jusqu'à Resource-based policy (Politique basée sur des ressources) puis choisissez Add permissions (Ajouter des autorisations). Une politique basée sur des ressources est utilisée pour accorder des autorisations pour appeler votre fonction au principal qui est spécifié dans la politique. Dans ce cas, le principal sera le rôle lié au service Amazon EC2 Auto Scaling associée au groupe Auto Scaling.
12. Dans Policy statement (Déclaration de politique), configurez vos autorisations :

    a. Sélectionnez Compte AWS.
    b. Pour Principal , saisissez l'ARN du rôle lié au service appelant, par exemple, **`arn:aws:iam::<aws-account-id>:role/aws-service-role/autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling`**.
    c. Pour Action, choisissez lambda : InvokeFunction.
    d. Pour Statement ID (ID de l'instruction), saisissez un ID d'instruction unique, tel que **`AllowInvokeByAutoScaling`**.
    e. Choisissez Save (Enregistrer).
13. Une fois que vous avez suivi ces instructions, l'étape suivante consiste à spécifier l'ARN de votre fonction dans les politiques de résiliation de votre groupe Auto Scaling. Pour plus d'informations, consultez Modifier la politique de résiliation d'un groupe Auto Scaling.

> **Note**
>
> Pour des exemples que vous pouvez utiliser comme référence pour développer votre fonction Lambda, consultez le GitHub référentiel Amazon EC2 Auto Scaling.

Limites

- Vous ne pouvez spécifier qu'une seule fonction Lambda dans les politiques de résiliation pour un groupe Auto Scaling. Si plusieurs politiques de résiliation sont spécifiées, la fonction Lambda doit d'abord être spécifiée.
- Vous pouvez référencer votre fonction Lambda en utilisant soit un ARN non qualifié (sans suffixe), soit un ARN qualifié qui a une version ou un alias comme suffixe. Si un ARN non qualifié est utilisé (par exemple, `function:my-function`), votre politique basée sur des ressources doit être créée sur la version non publiée de votre fonction. Si un ARN qualifié est utilisé (par exemple, `function:my-function:1` ou `function:my-function:prod`), votre politique basée sur les ressources doit être créée sur cette version publiée spécifique de votre fonction.
- Vous ne pouvez pas utiliser un ARN qualifié avec le suffixe `$LATEST`. Si vous essayez d'ajouter une politique de résiliation personnalisée qui fait référence à un ARN qualifié avec le suffixe `$LATEST`, cela entraînera une erreur.
- Le nombre d'instances fournies dans les données d'entrée est limité à 30 000 instances. Si plus de 30 000 instances peuvent être résiliées, les données en entrée incluent `"HasMoreInstances": true`pour indiquer que le nombre maximal d'instances renvoyées sont renvoyées.
- La durée maximale d'exécution de votre fonction Lambda est de deux secondes (2 000 millisecondes). Il est recommandé de définir la valeur de délai d'expiration de votre fonction Lambda en fonction du temps d'exécution prévu. Les fonctions Lambda ont un délai d'attente par défaut de trois secondes, mais cela peut être réduit.
- Si votre temps d'exécution dépasse la limite de 2 secondes, toute action évolutive sera suspendue jusqu'à ce que le temps d'exécution tombe en dessous de ce seuil. Pour les fonctions Lambda dont les temps d'exécution sont constamment plus longs, trouvez un moyen de réduire le temps d'exécution, par exemple en mettant en cache les résultats afin qu'ils puissent être récupérés lors des appels Lambda suivants.

Utiliser la protection de la taille d'instance

La protection intégrée des instances vous permet de contrôler les instances auxquelles Amazon EC2 Auto Scaling peut mettre fin. Un cas d'utilisation courant de cette fonctionnalité est le dimensionnement des charges de travail basées sur des conteneurs. Pour plus d’informations, consultez Concevez vos applications sur Amazon EC2 Auto Scaling pour gérer de manière optimale la résiliation des instances.

Par défaut, la protection contre le scale-in de l'instance est désactivée lorsque vous créez un groupe Auto Scaling. Cela signifie qu'Amazon EC2 Auto Scaling peut mettre fin à n'importe quelle instance du groupe.

Vous pouvez protéger les instances dès leur lancement en activant le paramètre de protection contre la mise à l'échelle horizontale d'instance sur votre groupe Auto Scaling. La protection contre la diminution de la taille d'instance démarre lorsque le statut de l'instance est `InService`. Ensuite, pour contrôler quelles instances peuvent être résiliées, désactivez le paramètre de protection contre la mise à l'échelle horizontale sur les instances individuelles du groupe Auto Scaling. Ce faisant, vous pouvez continuer à protéger certaines instances contre les résiliations indésirables.

Rubriques

Considérations

Les points suivants doivent être pris en compte lors de l'utilisation de la protection évolutive des instances :

- Si toutes les instances d'un groupe Auto Scaling sont protégées de la mise à l'échelle horizontale et qu'un tel événement se produit, sa capacité souhaitée est diminuée. Cependant, le groupe Auto Scaling ne peut pas résilier le nombre requis d'instances tant que les paramètres de protection contre la mise à l'échelle horizontale des instances sont activés. Dans le AWS Management Console, l'historique des activités du groupe Auto Scaling inclut le message suivant si toutes les instances d'un groupe Auto Scaling sont protégées contre la mise à l'échelle lorsqu'un événement de scale-in se produit : `Could not scale to desired capacity because all remaining instances are protected from scale-in.`
- Si vous détachez une instance qui est protégée de la mise à l'échelle horizontale, son paramètre de protection contre la diminution de la taille d'instance est perdu. Lorsque vous attachez de nouveau l'instance au groupe, elle hérite du paramètre actuel de protection contre la diminution de la taille d'instance du groupe. Lorsque Amazon EC2 Auto Scaling lance une nouvelle instance ou déplace une instance d'un groupe chaud vers le groupe Auto Scaling, l'instance hérite du paramètre de protection de mise à l'échelle horizontale de l'instance du groupe Auto Scaling.
- La protection contre la mise à l'échelle horizontale d'instance ne protège pas les instances Auto Scaling des actions suivantes :

- Remplacement lié à la surveillance de l'état si une instance échoue à des surveillances de l'état. Pour plus d'informations, consultez Surveillance de l'état des instances dans un groupe Auto Scaling.
- Interruptions d'instances Spot Une instance Spot est mise hors service lorsque la capacité n'est plus disponible ou lorsque son prix dépasse votre prix maximum.
- La réservation d'un bloc de capacité prend fin. Amazon EC2 récupère les instances Capacity Block même si elles sont protégées contre toute extension.
- Résiliation manuelle par le biais de la `terminate-instance-in-auto-scaling-group` commande. Pour plus d'informations, consultez Résilier une instance de votre groupe Auto Scaling (AWS CLI).
- Résiliation manuelle via la console Amazon EC2, les commandes CLI et les opérations d'API. Pour protéger les instances Auto Scaling d'une résiliation manuelle, activez la protection de la résiliation Amazon EC2. (Cela n'empêche pas Amazon EC2 Auto Scaling de mettre fin à des instances ou de les arrêter manuellement par le biais de la `terminate-instance-in-auto-scaling-group` commande.) Pour plus d'informations sur l'activation de la protection de résiliation Amazon EC2 dans un modèle de lancement, consultez. Créer un modèle de lancement à l'aide de paramètres avancés

Modifier la protection intégrée pour un groupe Auto Scaling

Vous pouvez activer ou désactiver le paramètre de protection contre la mise à l'échelle horizontale d'instance pour un groupe Auto Scaling. Lorsque vous l'activez, la protection évolutive est activée pour toutes les nouvelles instances lancées par le groupe.

L'activation ou la désactivation de ce paramètre pour un groupe Auto Scaling n'affecte pas les instances existantes.

Console

Pour activer la protection évolutive pour un nouveau groupe Auto Scaling

Lorsque vous créez le groupe Auto Scaling, sur la page Configurer la taille du groupe et les politiques de dimensionnement, sous Protection évolutive des instances, cochez la case Activer la protection évolutive des instances.

Pour activer ou désactiver la protection évolutive pour un groupe existant

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Sous l'onglet Détails, choisissez Configurations avancées, Modifier.
4. Pour la protection intégrée de l'instance, cochez ou décochez la case Activer la protection intégrée de l'instance pour activer ou désactiver cette option selon les besoins.
5. Choisissez Mettre à jour.

AWS CLI

Pour activer la protection évolutive pour un nouveau groupe Auto Scaling

Utilisez la commande create-auto-scaling-group suivante pour activer la protection de mise à l'échelle horizontale d'instance.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg --new-
instances-protected-from-scale-in ...
```

Pour activer la protection évolutive d'un groupe existant

Utilisez la commande update-auto-scaling-group suivante pour activer la protection d'instance pour le groupe Auto Scaling spécifié.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg --new-
instances-protected-from-scale-in
```

Pour désactiver la protection intégrée pour un groupe existant

Utilisez la commande suivante pour désactiver la protection contre la mise à l'échelle horizontale des instances pour le groupe spécifié.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg --no-new-
instances-protected-from-scale-in
```

Modifier la protection évolutive d'une instance

Par défaut, une instance récupère le paramètre de protection contre la diminution de la taille d'instance de son groupe Auto Scaling. Toutefois, vous pouvez activer ou désactiver la protection évolutive des instances individuelles après leur lancement.

Console

Pour activer ou désactiver la protection évolutive pour une instance

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Sous l'onglet Gestion des instances dans Instances, sélectionnez une instance.
4. Pour activer la protection contre la mise à l'échelle horizontale des instances, choisissez Actions, Set Scale In Protection (Définir la protection contre la mise à l'échelle horizontale). Lorsque vous y êtes invité, choisissez Set Scale In Protection (Définir la protection de l'échelle).
5. Pour désactiver la protection contre la mise à l'échelle horizontale des instances, choisissez Actions, Remove scale in protection (Supprimer la protection contre la mise à l'échelle horizontale). Lorsque vous y êtes invité, choisissez Remove Scale In Protection (Supprimer la protection contre la mise à l'échelle horizontale).

AWS CLI

Pour activer la protection évolutive d'une instance

Utilisez la commande set-instance-protection suivante pour activer la protection contre la mise à l'échelle horizontale d'instance pour l'instance spécifiée.

```
aws autoscaling set-instance-protection --instance-ids i-5f2e8a0d --auto-scaling-
group-name my-asg --protected-from-scale-in
```

Pour désactiver la protection évolutive pour une instance

Utilisez la commande suivante pour désactiver la protection contre la mise à l'échelle horizontale des instances pour l'instance spécifiée.

```
aws autoscaling set-instance-protection --instance-ids i-5f2e8a0d --auto-scaling-
group-name my-asg --no-protected-from-scale-in
```

> **Note**
>
> N'oubliez pas que la protection évolutive des instances ne garantit pas que les instances ne seront pas résiliées en cas d'erreur humaine, par exemple si quelqu'un met fin manuellement à une instance à l'aide de la console Amazon EC2 ou. AWS CLI Pour protéger votre instance contre une résiliation accidentelle, vous pouvez utiliser la protection contre la résiliation Amazon EC2. Toutefois, même si la protection de fin d'instance et la protection d'échelle d'instance sont activées, les données enregistrées dans le stockage d'instance peuvent être perdues si une surveillance de l'état d'intégrité détermine qu'une instance est défectueuse ou si le groupe lui-même est supprimé accidentellement. Comme pour tout environnement, la meilleure pratique consiste à sauvegarder vos données fréquemment ou chaque fois qu'elles sont adaptées à vos besoins en matière de continuité d'activité.

Concevez vos applications sur Amazon EC2 Auto Scaling pour gérer de manière optimale la résiliation des instances

Cette rubrique décrit les différentes approches que vous pouvez adopter si des applications s’exécutent sur des instances qui, idéalement, ne devraient pas être résiliées de façon inattendue lorsqu’Amazon EC2 Auto Scaling répond à un événement de mise à l’échelle horizontale.

Supposons, par exemple, que vous ayez une file d’attente Amazon SQS qui collecte les messages entrants pour des tâches de longue durée. Lorsqu’un nouveau message arrive, une instance du groupe Auto Scaling récupère le message et commence à le traiter. Le traitement de chaque message prend 3 heures. À mesure que le nombre de messages augmente, de nouvelles instances sont automatiquement ajoutées au groupe Auto Scaling. À mesure que le nombre de messages diminue, les instances existantes sont automatiquement supprimées. Dans ce cas, Amazon EC2 Auto Scaling doit décider quelle instance résilier. Par défaut, il est possible qu’Amazon EC2 Auto Scaling mette fin à une instance après 2,9 heures de traitement d’un travail de 3 heures, plutôt qu’une instance actuellement inactive. Pour éviter les problèmes liés à des résiliations inattendues lors de l’utilisation d’Amazon EC2 Auto Scaling, vous devez concevoir votre application de manière à répondre à ce scénario.

Vous pouvez utiliser les fonctionnalités suivantes pour empêcher votre groupe Auto Scaling de résilier des instances qui ne sont pas encore prêtes à être résiliées ou de résilier des instances trop rapidement pour leur permettre de terminer les tâches qui leur sont assignées. Ces trois fonctionnalités peuvent être utilisées conjointement ou séparément.

Table des matières

> **Important**
>
> Lorsque vous concevez vos applications sur Amazon EC2 Auto Scaling pour gérer de manière optimale la résiliation des instances, gardez ces points à l'esprit.
>
> - Si une instance est défectueuse, Amazon EC2 Auto Scaling la remplacera quelle que soit la fonctionnalité que vous utilisez (sauf si vous suspendez le processus `ReplaceUnhealthy`). Vous pouvez utiliser un hook de cycle de vie pour permettre à l'application de s'arrêter de manière optimale ou pour copier les données que vous devez récupérer avant de résilier l'instance.
> - Il n'est pas garanti qu'un hook de cycle de vie de résiliation s'exécute ou se termine avant la résiliation d'une instance. En cas d'échec, Amazon EC2 Auto Scaling résilie quand même l'instance.

Protection contre la mise à l'échelle horizontale d'instance

Vous pouvez utiliser la protection contre la mise à l'échelle horizontale des instances dans de nombreuses situations où la résiliation d'instances est une action critique qui doit être refusée par défaut et autorisée de manière explicite uniquement pour des instances spécifiques. Par exemple, lors de l'exécution de charges de travail conteneurisées, il est courant de vouloir protéger toutes les instances et de supprimer la protection uniquement pour les instances sans tâches en cours ou planifiées. Des services tels qu'Amazon ECS ont intégré à leurs produits des intégrations avec protection contre la mise à l'échelle horizontale d'instance.

Vous pouvez activer la protection contre la mise à l'échelle horizontale dans le groupe Auto Scaling pour appliquer la protection contre la mise à l'échelle horizontale aux instances lors de leur création

et l'activer pour les instances existantes. Lorsqu'une instance n'a plus de travail à effectuer, elle peut désactiver la protection. L'instance peut continuer à rechercher de nouvelles tâches et réactiver la protection lorsque de nouvelles tâches sont attribuées.

Les applications peuvent définir la protection soit à partir d'un plan de contrôle centralisé qui détermine si une instance est résiliable ou non, soit à partir des instances elles-mêmes. Toutefois, une flotte importante peut rencontrer des problèmes de limitation si un grand nombre d'instances activent continuellement leur protection contre la mise à l'échelle horizontale.

Pour plus d'informations, consultez Utiliser la protection de la taille d'instance.

Politique de résiliation personnalisée

À l'instar de la protection contre la mise à l'échelle horizontale d'instance, une politique de résiliation personnalisée vous aide à empêcher votre groupe Auto Scaling de résilier des instances spécifiques.

Par défaut, votre groupe Auto Scaling utilise une politique de résiliation par défaut pour déterminer quelles instances il résilie en premier. Si vous souhaitez mieux contrôler quelles instances sont résiliées en premier, vous pouvez implémenter votre propre politique de résiliation personnalisée à l'aide d'une fonction Lambda. Amazon EC2 Auto Scaling appelle la fonction chaque fois qu'il doit décider de l'instance à résilier. Cela ne met fin qu'à une instance renvoyée par la fonction. Si la fonction est erronée, expire ou produit une liste vide, Amazon EC2 Auto Scaling ne met pas fin aux instances.

Une politique de résiliation personnalisée est utile si l'on sait qu'une instance est suffisamment redondante ou sous-utilisée pour pouvoir être résiliée. Pour ce faire, vous devez mettre en œuvre votre application avec un plan de contrôle qui surveille la charge de travail au sein du groupe. Ainsi, si une instance traite encore des tâches, la fonction Lambda sait qu'il ne faut pas l'inclure.

Pour plus d'informations, consultez Créer une politique de résiliation personnalisée avec Lambda.

Hooks de cycle de vie de résiliation

Un hook de cycle de vie de résiliation prolonge la durée de vie d'une instance déjà sélectionnée pour être résiliée. Cela permet de disposer de plus de temps pour traiter tous les messages ou demandes actuellement affectés à l'instance, ou pour enregistrer la progression et transférer le travail vers une autre instance.

Pour de nombreuses charges de travail, un hook de cycle de vie peut être suffisant pour arrêter de manière optimale une application sur une instance sélectionnée pour être résiliée. Il s'agit d'une

approche basée sur le meilleur effort qui ne peut pas être utilisée pour empêcher la résiliation en cas de panne.

Pour utiliser un hook de cycle de vie, vous devez savoir quand une instance est sélectionnée pour être résiliée. Vous pouvez le savoir de deux manières :

Option	Description	Utiliser en priorité pour	Lien vers la documentation
À l'intérieur de l'instance	Le service de métadonnées d'instance (IMDS) est un point de terminaison sécurisé dans lequel vous pouvez interroger le statut d'une instance directement depuis celle-ci. Si les métadonnées sont renvoyées avec `Terminated` , il est prévu de mettre fin à votre instance.	Applications dans lesquelles vous devez effectuer une action sur l'instance avant qu'elle ne soit résiliée.	Récupérer l'état du cycle de vie cible
En dehors de l'instance	Lorsqu'une instance est résiliée, une notification d'événement est générée. Vous pouvez créer des règles à l'aide d'Amazon EventBridge, Amazon SQS ou Amazon SNS pour capturer ces événements et appeler une réponse, par exemple avec une fonction Lambda.	Applications qui doivent agir en dehors de l'instance.	Configurer une cible de notification

Pour utiliser un hook de cycle de vie, vous devez aussi savoir quand une instance est prête à être entièrement résiliée. Amazon EC2 Auto Scaling ne demandera pas à Amazon EC2 de mettre fin à l'instance tant qu'elle n'aura pas reçu CompleteLifecycleun appel Action ou que le délai d'expiration n'aura pas expiré, selon la première éventualité.

Par défaut, une instance peut continuer à fonctionner pendant une heure (délai d'attente des pulsations) en raison d'un hook de cycle de vie. Vous pouvez configurer le délai d'expiration par défaut si une heure n'est pas suffisante pour terminer l'action du cycle de vie. Lorsqu'une action du cycle de vie est réellement en cours, vous pouvez prolonger le délai d'expiration à l'aide d'appels RecordLifecycleActionHeartbeatd'API.

Pour plus d'informations, consultez Hooks de cycle de vie Amazon EC2 Auto Scaling.

Suspendre et reprendre les processus Amazon EC2 Auto Scaling

Cette rubrique décrit comment suspendre puis reprendre un ou plusieurs processus pour votre groupe Auto Scaling afin de désactiver temporairement certaines opérations.

La suspension des processus peut être utile lorsque vous devez étudier ou résoudre un problème sans que des politiques de dimensionnement ou des actions planifiées n'interfèrent. Cela permet également d'empêcher Amazon EC2 Auto Scaling de marquer les instances comme étant défectueuses et de les remplacer lorsque vous apportez des modifications à votre groupe Auto Scaling.

Rubriques

> **Note**
>
> En plus des suspensions que vous lancez, Amazon EC2 Auto Scaling peut également suspendre des processus pour des groupes Auto Scaling qui, de façon répétée, ne parviennent pas à lancer les instances. Cette action est appelée suspension administrative. Une suspension administrative s'applique généralement aux groupes Auto Scaling qui ont tenté de lancer des instances pendant plus de 24 heures sans succès. Vous pouvez reprendre des processus suspendus par Amazon EC2 Auto Scaling pour raisons administratives.

Types de processus

La fonction Suspend-resume (Suspendre–reprendre) prend en charge les processus suivants :

- `Launch`— Ajoute des instances au groupe Auto Scaling lorsque le groupe s'agrandit ou lorsqu'Amazon EC2 Auto Scaling choisit de lancer des instances pour d'autres raisons, par exemple lorsqu'il ajoute des instances à un pool de chaleur.
- `Terminate`— Supprime des instances du groupe Auto Scaling lorsque le groupe prend de l'ampleur ou lorsqu'Amazon EC2 Auto Scaling choisit de mettre fin à des instances pour d'autres raisons, par exemple lorsqu'une instance est résiliée pour avoir dépassé sa durée de vie maximale ou pour avoir échoué à un bilan de santé.
- `AddToLoadBalancer`— Ajoute des instances au groupe cible de l'équilibreur de charge attaché ou au Classic Load Balancer lors de leur lancement. Pour plus d'informations, consultez Utiliser Elastic Load Balancing pour répartir le trafic sur les instances dans votre groupe Auto Scaling..
- `AlarmNotification`— Accepte les notifications provenant d' CloudWatchalarmes associées à des politiques de dimensionnement dynamiques. Pour plus d'informations, consultez Mise à l'échelle dynamique pour Amazon EC2 Auto Scaling.
- `AZRebalance`— Équilibre le nombre d'instances EC2 du groupe de manière égale dans toutes les zones de disponibilité spécifiées lorsque le groupe devient déséquilibré, par exemple lorsqu'une zone de disponibilité précédemment indisponible revient à un état sain. Pour plus d'informations, consultez Activités de rééquilibrage.
- `HealthCheck`— Vérifie l'état des instances et marque une instance comme étant défectueuse si Amazon EC2 ou Elastic Load Balancing indique à Amazon EC2 Auto Scaling que l'instance n'est pas saine. Ce processus peut remplacer l'état de santé d'une instance défini manuellement. Pour plus d'informations, consultez Surveillance de l'état des instances dans un groupe Auto Scaling.
- `InstanceRefresh`— Résout et remplace les instances à l'aide de la fonction d'actualisation des instances. Pour plus d'informations, consultez Utiliser une actualisation d'instance pour mettre à jour les instances d'un groupe Auto Scaling.
- `ReplaceUnhealthy`— Met fin aux instances marquées comme défectueuses, puis crée de nouvelles instances pour les remplacer. Pour plus d'informations, consultez Surveillance de l'état des instances dans un groupe Auto Scaling.
- `ScheduledActions`— Exécute les actions de dimensionnement planifiées que vous créez ou qui sont créées pour vous lorsque vous créez un plan de AWS Auto Scaling dimensionnement et activez le dimensionnement prédictif. Pour plus d'informations, consultez Mise à l'échelle planifiée pour Amazon EC2 Auto Scaling.

Considérations

Tenez compte des éléments suivants avant de suspendre les processus :

- La suspension vous `AlarmNotification` permet d'arrêter temporairement les politiques de suivi des cibles, d'étapes et de dimensionnement simple du groupe sans supprimer les politiques de dimensionnement ni les CloudWatch alarmes associées. Pour arrêter temporairement les politiques de mise à l'échelle individuelles, reportez-vous à la rubrique Désactiver une politique de mise à l'échelle pour un groupe Auto Scaling.
- Vous pouvez choisir de suspendre les `ReplaceUnhealthy` processus `HealthCheck` et pour redémarrer les instances sans qu'Amazon EC2 Auto Scaling ne mette fin aux instances sur la base de ses tests de santé. Toutefois, si vous avez besoin d'Amazon EC2 Auto Scaling pour continuer à effectuer des contrôles de santé sur les instances restantes, utilisez plutôt la fonction de veille. Pour plus d'informations, consultez Supprimer temporairement des instances du groupe Auto Scaling.
- Si vous suspendez les processus `Launch` et `Terminate`, ou `AZRebalance`, puis que vous apportez des modifications à votre groupe Auto Scaling, par exemple en détachant des instances ou en modifiant les zones de disponibilité spécifiées, votre groupe peut devenir déséquilibré entre les zones de disponibilité. Si cela se produit, après avoir repris les processus suspendus, Amazon EC2 Auto Scaling redistribue progressivement les instances de manière uniforme entre les zones de disponibilité.
- Si vous suspendez le `Terminate` processus, vous pouvez toujours forcer la fermeture des instances en utilisant la commande delete-auto-scaling-group avec l'option force delete.
- La suspension du `Terminate` processus ne s'applique qu'aux instances qui sont actuellement dans `InService` cet état. Cela n'empêche pas la mise hors service d'instances situées dans d'autres états, par exemple`Pending`, ou d'instances qui ne redémarrent pas correctement depuis le mode veille.
- Le `RemoveFromLoadBalancerLowPriority` processus peut être ignoré lorsqu'il est présent dans des appels destinés à décrire des groupes Auto Scaling à l'aide des SDK AWS CLI ou. Ce processus est obsolète et n'est conservé qu'à des fins de rétrocompatibilité.

Suspendre des processus

Pour suspendre un processus pour un groupe Auto Scaling, appliquez l'une des méthodes suivantes :

Console

Pour suspendre un processus

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Sous l'onglet Détails, choisissez Configurations avancées, Modifier.
4. Pour Suspended Processes (Processus suspendus), sélectionnez le processus à suspendre.
5. Choisissez Mettre à jour.

AWS CLI

Utilisez la commande suspend-processes (suspendre-reprendre) suivante pour suspendre des processus individuels.

```
aws autoscaling suspend-processes --auto-scaling-group-name my-asg --scaling-
processes HealthCheck ReplaceUnhealthy
```

Pour suspendre tous les processus, omettez l'option `--scaling-processes`, comme suit.

```
aws autoscaling suspend-processes --auto-scaling-group-name my-asg
```

Processus de CV

Pour reprendre un processus suspendu pour un groupe Auto Scaling, appliquez l'une des méthodes suivantes :

Console

Pour reprendre un processus suspendu

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

Un volet fractionné s'ouvre en bas de la page.

3. Sous l'onglet Détails, choisissez Configurations avancées, Modifier.
4. Pour Suspended Processes (Processus suspendus), sélectionnez le processus à suspendre.
5. Choisissez Mettre à jour.

AWS CLI

Pour reprendre un processus suspendu, utilisez la commande resume-processes suivante.

```
aws autoscaling resume-processes --auto-scaling-group-name my-asg --scaling-
processes HealthCheck
```

Pour reprendre tous les processus suspendus, omettez l'option `--scaling-processes`, comme suit.

```
aws autoscaling resume-processes --auto-scaling-group-name my-asg
```

Comment les processus suspendus affectent les autres processus

Les sections suivantes décrivent ce qui se passe lorsque différents processus sont suspendus individuellement.

Rubriques

Launchest suspendu

- `AlarmNotification` est toujours actif, mais votre groupe Auto Scaling ne peut pas lancer d'activités de montée en puissance pour les alarmes en violation.
- `ScheduledActions` est actif, mais votre groupe Auto Scaling ne peut pas lancer d'activités de montée en puissance pour les actions planifiées qui se produisent.
- `AZRebalance` cesse de rééquilibrer le groupe.
- `ReplaceUnhealthy` continue de mettre fin aux instances malsaines, mais ne lance pas de remplacements. Lorsque vous reprenez le processus `Launch`, Amazon EC2 Auto Scaling remplace immédiatement toutes les instances qu'il a interrompues pendant la période où `Launch` était suspendu.
- `InstanceRefresh` ne remplace pas les instances.

Terminateest suspendu

- `AlarmNotification` est toujours actif, mais votre groupe Auto Scaling ne peut pas lancer d'activités de mise à l'échelle horizontale pour les alarmes en violation.
- `ScheduledActions` est actif, mais votre groupe Auto Scaling ne peut pas lancer d'activités de mise à l'échelle horizontale pour les actions planifiées qui se produisent.
- `AZRebalance` est toujours actif mais ne fonctionne pas correctement. Il peut lancer de nouvelles instances sans résilier les anciennes. Cela peut entraîner une augmentation de votre groupe Auto Scaling jusqu'à une taille supérieure de 10 % à sa taille maximum, ce qui est autorisé temporairement pendant les activités de rééquilibrage. Votre groupe Auto Scaling peut rester au-dessus de sa taille maximum jusqu'à ce que vous repreniez le processus `Terminate`.
- Le processus `ReplaceUnhealthy` est inactif, mais pas le processus `HealthCheck`. Lorsque `Terminate` reprend, l'exécution du processus `ReplaceUnhealthy` démarre immédiatement. Si toutes les instances ont été marquées comme non saines alors que `Terminate` était suspendu, elles seront immédiatement remplacées.
- `InstanceRefresh` ne remplace pas les instances.

AddToLoadBalancerest suspendu

- Amazon EC2 Auto Scaling lance les instances mais ne les ajoute pas au groupe cible de l'équilibreur de charge ou au Classic Load Balancer. Lorsque vous reprenez le processus `AddToLoadBalancer`, il reprend l'ajout d'instances à l'équilibreur de charge lorsqu'elles sont

lancées. Cependant, il n'ajoute pas les instances lancées pendant la suspension du processus. Vous devez enregistrer ces instances manuellement.

AlarmNotificationest suspendu

- Amazon EC2 Auto Scaling n'invoque pas de politiques de dimensionnement lorsqu'un seuil CloudWatch d'alarme est dépassé. Lorsque vous reprenez le processus `AlarmNotification`, Amazon EC2 Auto Scaling prend en compte les politiques avec des seuils d'alarme qui sont actuellement dépassés.

AZRebalanceest suspendu

- Amazon EC2 Auto Scaling ne tente pas de redistribuer les instances après certains événements. Cependant, si un événement d'évolutivité horizontale ou de mise à l'échelle horizontale se produit, le processus de mise à l'échelle tente toujours d'équilibrer les zones de disponibilité. Par exemple, pendant l'augmentation de la taille des instances, il lance l'instance dans la zone de disponibilité avec le moins d'instances. Si cela entraîne un déséquilibre du groupe pendant la suspension et la reprise du processus `AZRebalance`, Amazon EC2 Auto Scaling tente de rééquilibrer le groupe. Le service commence par appeler `Launch`, puis `Terminate`.

HealthCheckest suspendu

- Amazon EC2 Auto Scaling arrête de marquer les instances identifiées non saines suite à des surveillances de l'état EC2 et Elastic Load Balancing. Vos surveillances de l'état personnalisées continuent de fonctionner correctement. Après avoir suspendu `HealthCheck`, vous pouvez définir manuellement l'état des instances de votre groupe et demander au processus `ReplaceUnhealthy` de les remplacer.

InstanceRefreshest suspendu

- Amazon EC2 Auto Scaling arrête de remplacer les instances suite à une actualisation d'instance. Si une actualisation d'instance est en cours, cette opération interrompt l'opération sans l'annuler.

`ReplaceUnhealthy`est suspendu

- Amazon EC2 Auto Scaling arrête de remplacer les instances marquées comme non saines. Les instances dont les surveillances de l'état EC2 ou Elastic Load Balancing échouent sont toujours marquées comme non saines. Dès que vous reprenez le processus `ReplaceUnhealthy`, Amazon EC2 Auto Scaling remplace les instances qui ont été marquées comme non saines pendant la suspension du processus. Le processus `ReplaceUnhealthy` appelle `Terminate` d'abord puis `Launch`.

`ScheduledActions`est suspendu

- Amazon EC2 Auto Scaling n'exécute pas les actions planifiées qui sont programmées pour s'exécuter pendant la période de suspension. Lorsque vous reprenez `ScheduledActions`, Amazon EC2 Auto Scaling ne prend en compte que les actions planifiées dont l'heure prévue n'est pas encore passée.

Considérations supplémentaires

De plus, lorsque `Launch` ou `Terminate` sont suspendus, les fonctions suivantes peuvent ne pas fonctionner correctement :

- Durée de vie maximale des instances : en cas de suspension `Launch` ou `Terminate` de suspension, la fonctionnalité de durée de vie maximale des instances ne peut remplacer aucune instance.
- Interruptions des instances Spot : si l'instance `Terminate` est suspendue et que votre groupe Auto Scaling possède des instances Spot, elles peuvent toujours être résiliées si la capacité Spot n'est plus disponible. Pendant la suspension de `Launch`, Amazon EC2 Auto Scaling ne peut pas lancer d'instances de remplacement à partir d'un autre groupe d'instances Spot ou à partir du même groupe d'instances Spot lorsqu'il est disponible.
- Rééquilibrage des capacités : s'il `Terminate` est suspendu et que vous utilisez le rééquilibrage des capacités pour gérer les interruptions des instances Spot, le service Amazon EC2 Spot peut toujours mettre fin aux instances si la capacité Spot n'est plus disponible. Si `Launch` est suspendu, Amazon EC2 Auto Scaling ne peut pas lancer d'instances de remplacement à partir d'un autre groupe d'Instances Spot ou du même groupe d'Instances Spot lorsqu'il est à nouveau disponible.

- Attacher et détacher des instances : lorsque `Launch` celles-ci `Terminate` sont suspendues, vous pouvez détacher les instances qui sont attachées à votre groupe Auto Scaling, mais en `Launch` cas de suspension, vous ne pouvez pas associer de nouvelles instances au groupe.
- Instances en veille : lorsque `Launch` vous `Terminate` êtes suspendue, vous pouvez mettre une instance dans l'`Standby`état, mais pendant qu'elle `Launch` est suspendue, vous ne pouvez pas remettre en service une instance dans `Standby` cet état.

Surveillez vos groupes Amazon EC2 Auto Scaling

La surveillance joue un rôle important dans le maintien de la fiabilité, de la disponibilité et des performances d'Amazon EC2 Auto Scaling et de vos AWS Cloud solutions. AWS fournit les outils de surveillance suivants pour surveiller Amazon EC2 Auto Scaling, signaler un problème et prendre des mesures automatiques le cas échéant :

Surveillance de l'état

Amazon EC2 Auto Scaling effectue périodiquement des surveillances de l'état sur les instances de votre groupe Auto Scaling. Si une instance échoue à la surveillance de l'état, elle est marquée comme défectueuse et est résiliée pendant qu'Amazon EC2 Auto Scaling lance une nouvelle instance pour la remplacer. Pour plus d'informations, consultez Surveillance de l'état des instances dans un groupe Auto Scaling.

AWS Health Dashboard

AWS Health Dashboard Affiche des informations et fournit également des notifications qui sont invoquées en cas de modification de l'état des AWS ressources. Les informations sont présentées de deux manières : sur un tableau de bord qui montre les événements récents et à venir organisés par catégorie, et dans un journal des événements complet qui contient tous les événements des 90 derniers jours. Pour plus d'informations, consultez AWS Health Dashboard notifications pour Amazon EC2 Auto Scaling.

CloudTrail

Avec AWS CloudTrail, vous pouvez suivre les appels passés à l'API Amazon EC2 Auto Scaling par ou en votre nom. Compte AWS CloudTrail stocke les informations dans des fichiers journaux du compartiment Amazon S3 que vous spécifiez. Vous pouvez utiliser ces fichiers journaux pour surveiller l'activité de vos groupes Auto Scaling. Les journaux incluent les demandes envoyées, les adresses IP sources d'où émanent les demandes, l'identité du demandeur, le moment où la demande a été faite, etc. Pour plus d'informations, consultez Enregistrez les appels d'API Amazon EC2 Auto Scaling avec AWS CloudTrail.

> **Collecte des journaux de vos instances Amazon EC2**
>
> Vous pouvez l'utiliser CloudWatch pour collecter des journaux à partir des systèmes d'exploitation de vos instances EC2. Pour plus d'informations, consultez Collecter des métriques et des journaux à partir d'instances Amazon EC2 et de serveurs locaux avec l'

CloudWatch agent et Afficher les données de journal envoyées à CloudWatch Logs dans le guide de l'utilisateur Amazon CloudWatch .
Pour plus d'informations sur les autres AWS services qui peuvent vous aider à enregistrer et à collecter des données relatives à vos charges de travail, consultez le guide de journalisation et de surveillance destiné aux propriétaires d'applications dans le guide AWS prescriptif.

Amazon CloudWatch

Amazon vous CloudWatch aide à analyser les journaux et, en temps réel, à surveiller les indicateurs de vos AWS ressources et de vos applications hébergées. Vous pouvez collecter et suivre les métriques, créer des tableaux de bord personnalisés, et définir des alarmes qui vous informent ou prennent des mesures lorsqu'une métrique spécifique atteint un seuil que vous spécifiez. Par exemple, vous pouvez être averti lorsque l'activité réseau est soudainement supérieure ou inférieure à la valeur attendue d'une métrique. Pour plus d'informations sur l'utilisation de ce service en vue de suivre les métriques de vos groupes Auto Scaling et de vos instances, consultez Surveillez CloudWatch les métriques de vos groupes et instances Auto Scaling.

CloudWatch suit également les métriques d'utilisation des AWS API pour Amazon EC2 Auto Scaling. Vous pouvez utiliser ces métriques pour configurer des alarmes qui vous alertent lorsque le volume d'appels de votre API dépasse un seuil que vous avez défini. Pour plus d'informations, consultez les statistiques AWS d'utilisation dans le guide de CloudWatch l'utilisateur Amazon.

AWS Compute Optimizer

Compute Optimizer fournit des recommandations relatives aux instances Amazon EC2 qui peuvent vous aider à décider de passer à un nouveau type d'instance. Il analyse si un type d'instance de groupe Auto Scaling convient et génère des recommandations afin de réduire le coût et d'améliorer les performances de vos charges de travail. Pour plus d'informations, consultez AWS Compute Optimizer À utiliser pour obtenir des recommandations pour le type d'instance d'un groupe Auto Scaling.

Amazon EventBridge

Amazon EventBridge est un service de bus d'événements sans serveur qui permet de connecter facilement vos applications à des données provenant de diverses sources. EventBridge fournit un flux de données en temps réel à partir de vos propres applications, applications SaaS (Software-as-a-Service) AWS et services et achemine ces données vers des cibles telles que Lambda. Cela

vous permet de surveiller les événements qui se produisent dans les services et de créer des architectures basées sur les événements. Pour plus d'informations, consultez EventBridge À utiliser pour gérer les événements Auto Scaling.

AWS Security Hub

Utilisez AWS Security Hub pour surveiller votre utilisation d'Amazon EC2 Auto Scaling, car cela est lié aux bonnes pratiques de sécurité. Security Hub utilise des contrôles de sécurité de détection pour évaluer les configurations des ressources et les normes de sécurité afin de vous aider à vous conformer à divers cadres de conformité. Pour plus d'informations sur l'utilisation de Security Hub pour évaluer les ressources Amazon EC2 Auto Scaling, consultez Contrôles Amazon EC2 Auto Scaling dans le Guide de l'utilisateur AWS Security Hub .

Amazon Simple Notification Service

Vous pouvez configurer les groupes Auto Scaling pour envoyer des notifications Amazon SNS quand Amazon EC2 Auto Scaling lance ou résilie des instances. Pour plus d'informations, voir Options de notification Amazon SNS pour Amazon EC2 Auto Scaling.

Surveillance de l'état des instances dans un groupe Auto Scaling

Amazon EC2 Auto Scaling surveille en permanence l'état de santé des instances d'un groupe Auto Scaling afin de maintenir la capacité souhaitée.

Toutes les instances d'un groupe Auto Scaling commencent par un `Healthy` statut. Les instances sont supposées être saines, sauf si Amazon EC2 Auto Scaling reçoit une notification indiquant qu'elles sont défectueuses. Il peut recevoir des notifications de différentes sources lorsqu'une instance devient défectueuse et doit être remplacée. Ces sources sont notamment les suivantes :

- Amazon EC2
- Elastic Load Balancing
- VPC Lattice
- Contrôles de santé personnalisés que vous définissez

Lorsqu'Amazon EC2 Auto Scaling détermine qu'une `InService` instance n'est pas saine, il la remplace par une nouvelle instance afin de maintenir la capacité souhaitée du groupe. La nouvelle instance se lance à l'aide des paramètres actuels du groupe Auto Scaling et du modèle de lancement associé ou de la configuration du lancement.

Des instances défectueuses peuvent également se produire lorsqu'une instance se ferme de manière inattendue, par exemple à la suite d'une interruption d'une instance Spot ou d'une résiliation manuelle par un utilisateur. Là encore, Amazon EC2 Auto Scaling lancera automatiquement une instance de remplacement dans ces cas afin de maintenir la capacité souhaitée.

Table des matières

À propos des surveillances de l’état de votre groupe Auto Scaling

Cette rubrique fournit une vue d'ensemble des types de bilans de santé disponibles et décrit les principales considérations relatives à l'intégration des contrôles de santé Amazon EC2 Auto Scaling à vos applications.

Table des matières

Type de surveillance de l'état

Amazon EC2 Auto Scaling peut déterminer l'état de santé d'une `InService` instance en utilisant un ou plusieurs des tests de santé suivants :

Type de surveillance de l'état	Ce qu'il vérifie
Vérifications de statut Amazon EC2 et événements planifiés	• Vérifie que l'instance est en cours d'exécution. • Vérifie les problèmes matériels ou logiciels sous-jacents susceptibles d'affecter l'instance. Il s'agit du type de surveillance de l'état par défaut pour un groupe Auto Scaling.
Surveillances de l'état Elastic Load Balancing	• Vérifie si l'équilibreur de charge indique que l'instance est saine, confirmant ainsi si l'instance est disponible pour traiter les demandes. Pour exécuter ce type de contrôle de santé, vous devez l'activer pour votre groupe Auto Scaling.
Surveillances de l’état de VPC Lattice	• Vérifie si VPC Lattice indique que l'instance est saine, confirmant ainsi si l'instance est disponible pour traiter les demandes. Pour exécuter ce type de contrôle de santé, vous devez l'activer pour votre groupe Auto Scaling.
Surveillances d'état personnalisées	• Vérifie tout autre problème susceptible d'indiquer des problèmes de santé de l'instance, conformément à vos bilans de santé personnalisés.

Surveillance de l'état Amazon EC2

Après le lancement d’une instance, elle est attachée au groupe Auto Scaling et entre dans l’état `InService`. Pour plus d'informations sur les différents états de cycle de vie des instances dans un groupe Auto Scaling, consultez Cycle de vie d'une instance Amazon EC2 Auto Scaling.

Amazon EC2 Auto Scaling vérifie périodiquement l'état de toutes les instances du groupe Auto Scaling pour s'assurer qu'elles fonctionnent et sont en bon état.

Contrôles des statuts

Par défaut, Amazon EC2 Auto Scaling utilise les résultats des vérifications de statut de l'instance Amazon EC2 et les vérifications de statut du système pour déterminer l'état d'une instance. Si l'instance se trouve dans un état Amazon EC2 autre que `running`, ou si son état pour les vérifications d'état passe à `impaired`, Amazon EC2 Auto Scaling considère que l'instance est défectueuse et la remplace. Même quand l'instance se trouve dans l'un des états suivants :

- `stopping`
- `stopped`
- `shutting-down`
- `terminated`

Les contrôles de statut Amazon EC2 ne nécessitent aucune configuration spéciale et sont toujours activés. Pour plus d'informations, consultez la section Types de vérifications de statut dans le guide de l'utilisateur Amazon EC2.

> **Important**
>
> Amazon EC2 Auto Scaling laisse les vérifications d'état échouer occasionnellement, sans prendre aucune mesure. Lorsqu'une vérification de statut échoue, Amazon EC2 Auto Scaling attend quelques minutes pour AWS résoudre le problème. Il ne marque pas immédiatement une instance comme `Unhealthy` lorsque son état pour les contrôles d'état devient `impaired`.
>
> Cependant, si Amazon EC2 Auto Scaling détecte qu'une instance n'est plus dans l'état `running`, cette situation est traitée comme un échec immédiat. Dans ce cas, il marque immédiatement l'instance comme telle `Unhealthy` et la remplace.

Événements planifiés

Amazon EC2 peut parfois planifier des événements sur vos instances pour qu'ils soient exécutés après un horodatage particulier. Pour plus d'informations, consultez la section Événements planifiés pour vos instances dans le guide de l'utilisateur Amazon EC2.

Si l'une de vos instances est affectée par un événement planifié, Amazon EC2 Auto Scaling considère que l'instance est défectueuse et la remplace. L'instance ne commence à s'arrêter que lorsque la date et l'heure spécifiées dans l'horodatage sont atteintes.

Surveillances de l'état Elastic Load Balancing

Lorsque vous activez les contrôles de santé Elastic Load Balancing pour votre groupe Auto Scaling, Amazon EC2 Auto Scaling peut utiliser les résultats de ces tests pour déterminer l'état de santé d'une instance.

Avant de pouvoir activer les contrôles de santé Elastic Load Balancing pour votre groupe Auto Scaling, vous devez configurer un équilibreur de charge Elastic Load Balancing et configurer un bilan de santé pour celui-ci afin de déterminer si vos instances sont saines. Pour plus d'informations, consultez Préparez-vous à associer un équilibreur de charge Elastic Load Balancing à votre groupe Auto Scaling.

Une fois que vous avez attaché l'équilibreur de charge à votre groupe Auto Scaling, voici ce qui se produit :

- Amazon EC2 Auto Scaling enregistre les instances du groupe Auto Scaling auprès de l'équilibreur de charge.
- Une fois qu'une instance a terminé son enregistrement, elle entre dans l'état `InService` et devient disponible pour une utilisation avec l'équilibreur de charge.

Par défaut, Amazon EC2 Auto Scaling ignore les résultats des surveillances de l'état Elastic Load Balancing. Une fois que vous avez activé ces contrôles de santé pour votre groupe Auto Scaling, lorsqu'Elastic Load Balancing signale une instance enregistrée comme telle`Unhealthy`, Amazon EC2 Auto Scaling marque l'instance `Unhealthy` lors de son prochain contrôle de santé périodique et la remplace.

Si le drainage de la connexion (délai d'annulation d'enregistrement) est activé pour votre équilibreur de charge, Amazon EC2 Auto Scaling attend soit la fin des demandes à la volée, soit l'expiration du délai maximal avant de mettre fin aux instances défectueuses.

> **Note**
>
> Pour savoir comment connecter l'équilibreur de charge et activer les contrôles de santé d'Elastic Load Balancing pour votre groupe Auto Scaling, consultezAssociez un équilibreur de charge Elastic Load Balancing à votre groupe Auto Scaling.
> Lorsque vous activez les contrôles de santé d'Elastic Load Balancing pour un groupe, Amazon EC2 Auto Scaling peut remplacer les instances signalées par Elastic Load Balancing comme étant défectueuses, mais uniquement une fois que l'équilibreur de charge est dans

cet état. `InService` Pour plus d'informations, consultez Vérifier l'état d'attachement de votre équilibreur de charge.

Surveillances de l'état de VPC Lattice

Par défaut, Amazon EC2 Auto Scaling ignore les résultats des surveillances de l'état de VPC Lattice. Vous pouvez éventuellement activer ces contrôles de santé pour votre groupe Auto Scaling. Après cette opération, lorsque VPC Lattice signale une instance enregistrée comme étant `Unhealthy`, Amazon EC2 Auto Scaling marque l'instance comme étant `Unhealthy` lors de sa prochaine surveillance de l'état périodique et la remplace. Le processus d'enregistrement des instances puis de vérification de leur état est le même que celui des surveillances de l'état Elastic Load Balancing.

> **Note**
>
> Pour savoir comment associer le groupe cible VPC Lattice et activer les contrôles de santé VPC Lattice pour votre groupe Auto Scaling, consultez. Attacher un groupe cible VPC Lattice à votre groupe Auto Scaling
>
> Lorsque vous activez les contrôles d'état de VPC Lattice pour un groupe, Amazon EC2 Auto Scaling peut remplacer les instances signalées par VPC Lattice comme étant défectueuses, mais uniquement une fois que le groupe cible est dans cet état. `InService` Pour plus d'informations, consultez Vérifier l'état d'attachement de votre groupe cible VPC Lattice.

Comment Amazon EC2 Auto Scaling minimise les temps d'arrêt

Par défaut, les nouvelles instances sont mises en service en même temps que les instances existantes sont résiliées, ce qui peut empêcher l'acceptation de nouvelles demandes tant que les nouvelles instances ne sont pas pleinement opérationnelles.

Si Amazon EC2 Auto Scaling détermine que des instances ne sont plus en cours d'exécution (ou si elles ont été `Unhealthy` marquées par la commande set-instance-health), il les remplace immédiatement. Toutefois, si d'autres instances sont jugées défectueuses, Amazon EC2 Auto Scaling utilise l'approche suivante pour récupérer les pannes. Cette approche minimise les temps d'arrêt qui pourraient survenir en raison de problèmes temporaires ou de surveillances d'état mal configurées.

- Si une activité de dimensionnement est en cours et que la capacité de votre groupe Auto Scaling est inférieure de 10 % ou plus à la capacité souhaitée, Amazon EC2 Auto Scaling attend l'activité de dimensionnement en cours avant de remplacer les instances défectueuses.
- Lors de la montée en puissance, Amazon EC2 Auto Scaling attend que les instances passent une surveillance de l'état initiale. Il attend également la fin de la préparation de l'instance par défaut pour s'assurer que les nouvelles instances sont prêtes.
- Une fois que le préchauffage des instances est terminé et que le groupe a atteint plus de 90 % de sa capacité souhaitée, Amazon EC2 Auto Scaling remplace les instances défectueuses comme suit :
  - Amazon EC2 Auto Scaling ne remplace que 10 % de la capacité souhaitée du groupe à la fois. Il le fait jusqu'à ce que toutes les instances défectueuses soient remplacées.
  - Lorsqu'il remplace des instances, il attend que les nouvelles instances passent une surveillance de l'état initiale. Il attend également la fin de la préparation de l'instance par défaut avant de poursuivre.

> **Note**
>
> Si la taille d'un groupe Auto Scaling est suffisamment petite pour que la valeur résultante de 10 % soit inférieure à un, Amazon EC2 Auto Scaling remplace les instances défectueuses une par une. Cela pourrait entraîner un certain temps d'arrêt pour le groupe.
> En outre, si toutes les instances d'un groupe Auto Scaling sont signalées comme défectueuses par les surveillances de l'état d'Elastic Load Balancing et que l'équilibreur de charge se trouve dans l'état `InService`, Amazon EC2 Auto Scaling pourrait marquer moins d'instances comme étant défectueuses à la fois. Le nombre d'instances remplacées à la fois peut ainsi être bien inférieur aux 10 % appliqués dans d'autres scénarios. Cela vous laisse le temps de résoudre le problème sans qu'Amazon EC2 Auto Scaling ne mette automatiquement fin à l'ensemble du groupe.

Contrôles de santé pour les cas dans une piscine chaude

Amazon EC2 Auto Scaling effectue également des contrôles de santé sur les instances d'un pool chaud. Pour plus d'informations, consultez Afficher le statut de surveillance de l'état et les motifs des échecs de surveillances de l'état.

Considérations relatives à la surveillance de l'état

Les points suivants sont à prendre en compte lors de l'utilisation des tests de santé Amazon EC2 Auto Scaling.

- Si vous avez besoin que quelque chose se produise sur l'instance en cours de résiliation ou sur l'instance en cours de démarrage, vous pouvez utiliser des hooks de cycle de vie. Ces hooks vous permettent d'effectuer une action personnalisée quand Amazon EC2 Auto Scaling lance des instances ou les résilie. Pour plus d'informations, consultez Hooks de cycle de vie Amazon EC2 Auto Scaling.
- Amazon EC2 Auto Scaling ne fournit pas de moyen de supprimer les contrôles d'état d'Amazon EC2 et les événements planifiés de ses surveillances de l'état. Si vous ne voulez pas que les instances soient remplacées, nous vous recommandons de suspendre le processus `ReplaceUnhealthy` et `HealthCheck` pour les groupes Auto Scaling individuels. Pour plus d'informations, consultez Suspendre et reprendre les processus Amazon EC2 Auto Scaling.
- Pour rétablir manuellement l'état de santé d'une instance défectueuse à `Healthy`, vous pouvez essayer d'utiliser la commande set-instance-health. Si vous obtenez une erreur, c'est probablement parce que la résiliation de l'instance est déjà en cours. En règle générale, redéfinir l'état de santé d'une instance en état `Healthy` avec la commande set-instance-health n'est utile que dans les cas où soit le processus `ReplaceUnhealthy`, soit le processus `Terminate`, est suspendu.
- Si vous devez dépanner une instance sans interférer avec les tests de santé, vous pouvez la mettre en `Standby` état. Amazon EC2 Auto Scaling n'effectue aucun contrôle de santé sur les instances en `Standby` état tant que vous ne les avez pas remises en service. Pour plus d'informations, consultez Supprimer temporairement des instances du groupe Auto Scaling.
- Lorsque l'instance est résiliée, les adresses IP Elastic associées sont dissociées et ne sont pas automatiquement associées à la nouvelle instance. Vous devez manuellement associer les adresses IP Elastic à la nouvelle instance, ou le faire automatiquement avec une solution basée sur le hook de cycle de vie. Pour plus d'informations, veuillez consulter la rubrique Adresses IP Elastic dans le Guide de l'utilisateur Amazon EC2.
- De la même façon, lorsque l'instance est résiliée, ses volumes EBS attachés sont détachés (ou supprimés selon l'attribut `DeleteOnTermination` du volume). Vous devez attacher manuellement ces volumes EBS à la nouvelle instance, ou le faire automatiquement avec une solution basée sur le hook de cycle de vie. Pour plus d'informations, consultez la section Attacher un volume Amazon EBS à une instance dans le guide de l'utilisateur Amazon EBS.

Surveillances d'état personnalisées

Si vous le souhaitez, vous pouvez également exécuter des tâches de détection d'état personnalisées sur les instances de votre groupe Auto Scaling et définir l'état d'une instance comme étant `Unhealthy` si la tâche échoue. Cela étend vos surveillances d'état en utilisant une combinaison de surveillances d'état personnalisées, de surveillances d'état Amazon EC2 et de surveillances d'état Elastic Load Balancing, si elles sont activées.

Vous pouvez envoyer les informations sur l'état de l'instance directement à Amazon EC2 Auto Scaling en utilisant la AWS CLI ou un kit SDK. Les exemples suivants montrent comment utiliser le AWS CLI pour configurer l'état de santé d'une instance, puis vérifier l'état de santé de l'instance.

Utilisez la commande set-instance-health suivante pour définir l'état de l'instance spécifiée sur **`Unhealthy`**.

```
aws autoscaling set-instance-health --instance-id i-1234567890abcdef0 --health-
status Unhealthy
```

Par défaut, cette commande respecte la période de grâce de la surveillance de l'état. Toutefois, vous pouvez remplacer ce comportement et ne pas respecter la période de grâce en incluant l'option `--no-should-respect-grace-period`.

Utilisez la commande describe-auto-scaling-groups suivante pour vérifier que l'état de l'instance est `Unhealthy`.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-names my-asg
```

Voici un exemple de réponse qui montre que l'état de santé de l'instance est `Unhealthy` et qu'elle est en cours de mise hors service.

```
{
    "AutoScalingGroups": [
        {
            ....
            "Instances": [
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
```

```
                    "LaunchTemplateId": "lt-1234567890abcdef0"
                },
                "InstanceId": "i-1234567890abcdef0",
                "InstanceType": "t2.micro",
                "HealthStatus": "Unhealthy",
                "LifecycleState": "Terminating"
            },
            ...
        ]
    }
]
}
```

Définir la période de grâce de la surveillance de l'état pour un groupe Auto Scaling

Lorsqu'une surveillance de l'état Amazon EC2 Auto Scaling détermine qu'une instance `InService` est défectueuse, il la remplace par une nouvelle instance. La période de grâce de la surveillance de l'état indique la durée minimale (en secondes) nécessaire pour maintenir une nouvelle instance en service avant de la résilier si elle devait être défectueuse.

On peut citer comme cas d'utilisation la nécessité pour Amazon EC2 Auto Scaling de ne pas prendre de mesure si les surveillances de l'état Elastic Load Balancing échouent et que la cause réside dans le fait que l'instance est toujours en cours d'initialisation. Les surveillances de l'état Elastic Load Balancing s'exécutent en parallèle, à partir du moment où l'instance est enregistrée auprès de l'équilibreur de charge. La période de grâce empêche Amazon EC2 Auto Scaling de marquer vos instances nouvellement lancées `Unhealthy` et de les résilier inutilement si elles ne passent pas ces tests de santé immédiatement après leur entrée dans l'État. `InService`

Dans la console, par défaut, la période de grâce de la surveillance de l'état est de 300 secondes lorsque vous créez un groupe Auto Scaling. Sa valeur par défaut est de 0 seconde lorsque vous créez un groupe Auto Scaling à l'aide du AWS CLI ou d'un SDK. La valeur 0 désactive la période de grâce de la surveillance de l'état.

Lorsque cette valeur est trop élevée, l'efficacité des surveillances de l'état Amazon EC2 Auto Scaling est réduite. Si vous utilisez des Hooks de cycle de vie pour le lancement de l'instance, vous pouvez définir la valeur de la période de grâce de la surveillance de l'état sur 0. Grâce aux Hooks de cycle de vie, Amazon EC2 Auto Scaling permet de s'assurer que les instances sont toujours initialisées avant de passer à l'état `InService`. Pour plus d'informations, consultez Hooks de cycle de vie Amazon EC2 Auto Scaling.

La période de grâce s'applique aux instances suivantes :

- Instances récemment lancées
- Instances remises en service après avoir été en veille
- Instances que vous attachez manuellement au groupe

> **Important**
>
> Pendant la période de grâce de la surveillance de l'état, si Amazon EC2 Auto Scaling détecte qu'une instance n'est plus à l'état `running` Amazon EC2, il marque l'instance comme `Unhealthy` et la remplace. Par exemple, si vous arrêtez une instance dans un groupe Auto Scaling, elle est marquée comme `Unhealthy` et remplacée.

Définir la période de grâce de la surveillance de l'état pour un groupe

Vous pouvez définir la période de grâce de la surveillance de l'état pour des groupes Auto Scaling nouveaux et existants.

Console

Pour modifier le délai de grâce du bilan de santé d'un nouveau groupe

Lorsque vous créez le groupe Auto Scaling, entrez la durée (en secondes) sur la page Configurer les options avancées, Health checks, Health check grace period. Il s'agit de la durée pendant laquelle Amazon EC2 Auto Scaling doit attendre avant de vérifier l'état de santé d'une instance après son entrée dans cet état. `InService`

AWS CLI

Pour modifier le délai de grâce du bilan de santé d'un nouveau groupe

Ajoutez l'option `--health-check-grace-period` à la commande create-auto-scaling-group. L'exemple suivant configure la période de grâce de la surveillance de l'état avec une valeur de *60* secondes pour un nouveau groupe Auto Scaling nommé *my-asg*.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-grace-period 60 ...
```

Console

Pour modifier le délai de grâce du bilan de santé d'un groupe existant

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Dans la barre de navigation située en haut de l'écran, choisissez l' Région AWS dans laquelle vous avez créé votre groupe Auto Scaling.
3. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
4. Sous l'onglet Détails choisissez Vérifications de l'états, Modifier.
5. Dans le champ Health check grace period (Période de grâce de la surveillance de l'état), saisissez le délai en secondes. Il s'agit de la durée pendant laquelle Amazon EC2 Auto Scaling doit attendre avant de vérifier l'état de santé d'une instance après son entrée dans cet état. `InService`
6. Choisissez Mettre à jour.

AWS CLI

Pour modifier le délai de grâce du bilan de santé d'un groupe existant

Ajoutez l'option `--health-check-grace-period` à la commande update-auto-scaling-group. L'exemple suivant configure la période de grâce de la surveillance de l'état avec une valeur de *120* secondes pour un groupe Auto Scaling existant nommé *my-asg*.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-grace-period 120
```

> **Note**
>
> Nous vous recommandons vivement de définir également le temps de préparation d'instance par défaut de votre groupe Auto Scaling. Pour plus d'informations, consultez Définir la préparation par défaut d'instance d'un groupe Auto Scaling.

Afficher le motif des échecs d’une surveillance de l’état

À l’aide de la procédure suivante, vous pouvez consulter les informations relatives à toutes les instances remplacées à la suite d’une surveillance de l’état.

Par défaut, Amazon EC2 Auto Scaling crée une nouvelle activité de mise à l’échelle pour résilier l’instance défectueuse, puis la résilie. Pendant que l'instance est résiliée, une autre activité de mise à l'échelle lance une nouvelle instance. Vous pouvez modifier ce comportement pour commencer à lancer une nouvelle instance dès que possible en utilisant une politique de maintenance des instances. Pour plus d’informations, consultez Politiques de maintenance des instances.

Console

Affichage de la raison des échecs du bilan de santé

1. Ouvrez la console Amazon EC2 à l’adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard du groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page Auto Scaling groups (Groupes Auto Scaling).
3. Sous l'onglet Activity (Activité) sous Activity history (Historique des activités), la colonne Status (État) indique si votre groupe Auto Scaling a réussi à lancer ou à résilier des instances.

   Si des instances malsaines sont interrompues, la colonne Cause indique la date et l'heure de l'interruption et le motif de l'échec de la surveillance de l'état. Par exemple, `At 2022-05-14T20:11:53Z an instance was taken out of service in response to a user health-check`. Ce message indique qu'un contrôle de santé personnalisé a indiqué que l'instance était défectueuse.

   Pour obtenir de l'aide en cas d'échec du bilan de santé, consultezRésoudre les problèmes liés aux instances défectueuses dans Amazon EC2 Auto Scaling.

AWS CLI

Affichage de la raison des échecs du bilan de santé

Utilisez la commande describe-scaling-activities suivante.

```
aws autoscaling describe-scaling-activities --auto-scaling-group-name my-asg
```

Voici un exemple de réponse `Cause` contenant la raison de l'échec du bilan de santé.

```
{
  "Activities": [
    {
      "ActivityId": "4c65e23d-a35a-4e7d-b6e4-2eaa8753dc12",
      "AutoScalingGroupName": "my-asg",
      "Description": "Terminating EC2 instance: i-04925c838b6438f14",
      "Cause": "At 2021-04-01T21:48:35Z an instance was taken out of service in
 response to a user health-check.",
      "StartTime": "2021-04-01T21:48:35.859Z",
      "EndTime": "2021-04-01T21:49:18Z",
      "StatusCode": "Successful",
      "Progress": 100,
      "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":\"us-
west-2a\"...}",
      "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:283179a2-
f3ce-423d-93f6-66bb518232f7:autoScalingGroupName/my-asg"
    },
...
  ]
}
```

Pour obtenir une description des champs de la sortie, consultez Activité dans la Référence de l'API Amazon EC2 Auto Scaling.

Pour décrire les activités de dimensionnement après la suppression du groupe Auto Scaling, ajoutez l'`--include-deleted-groups`option à la commande describe-scaling-activities.

Résoudre les problèmes liés aux instances défectueuses dans Amazon EC2 Auto Scaling

Vous trouverez ci-dessous les messages d'erreur renvoyés par Amazon EC2 Auto Scaling, les causes potentielles et les mesures que vous pouvez prendre pour résoudre les problèmes.

Pour récupérer un message d'erreur, consultez Afficher le motif des échecs d'une surveillance de l'état.

Messages d’erreur

Une instance a été mise hors service en réponse à un échec de surveillance de l'état de l'instance EC2

Problème : les instances Auto Scaling échouent aux surveillances de l'état d'Amazon EC2.

Cause 1 : Si des problèmes amènent Amazon EC2 à considérer que les instances de votre groupe Auto Scaling sont altérées, Amazon EC2 Auto Scaling remplace automatiquement les instances dans le cadre de ses bilans de santé.

Solution 1 : Lorsqu'une vérification de l'état d'une instance échoue, vous devez généralement résoudre le problème vous-même en modifiant la configuration de l'instance jusqu'à ce que votre application ne présente plus aucun problème. Pour résoudre ce problème, procédez comme suit :

1. Créez manuellement une instance Amazon EC2 qui ne fait pas partie du groupe Auto Scaling et examinez le problème. Pour obtenir de l'aide générale sur les instances défectueuses, consultez Résoudre les problèmes liés aux instances dont les vérifications de statut ont échoué dans le guide de l'utilisateur Amazon EC2 et Résolution des problèmes liés aux instances Windows dans le guide de l'utilisateur Amazon EC2.
2. Après avoir confirmé que votre instance a été lancée avec succès et qu'elle est saine, déployez une nouvelle configuration d'instance sans erreur dans le groupe Auto Scaling.
3. Supprimez l'instance que vous avez créée pour éviter les frais continus de votre instance AWS .

Une instance a été mise hors service en réponse à une surveillance de l'état EC2 qui indiquait qu'elle avait été résiliée ou arrêtée

Problème : les instances Auto Scaling qui ont été arrêtées, redémarrées ou résiliées sont remplacées.

Cause 1 : un utilisateur a arrêté, redémarré ou résilié l'instance manuellement.

Solution 1 : si vous devez arrêter ou redémarrer les instances de votre groupe Auto Scaling, nous vous recommandons de les mettre d'abord en veille. Pour plus d'informations, consultez Supprimer temporairement des instances du groupe Auto Scaling.

Cause 2 : Amazon EC2 Auto Scaling tente de remplacer les instances Spot après que le service Amazon EC2 Spot interrompe les instances, car le prix Spot augmente au-dessus de votre prix maximum ou capacité n'est plus disponible.

Solution 2 : il n'y a aucune garantie qu'une instance Spot existe pour répondre à la demande à un moment donné. Cependant, vous pouvez essayer l'une des actions suivantes :

- Utilisez un prix maximum Spot plus élevé (éventuellement prix à la demande). En fixant votre prix maximum plus élevé, cela donne au service Amazon EC2 Spot plus de chances de lancer et de maintenir la quantité de capacité requise.
- Augmentez le nombre de pools de capacités différents à partir desquels vous pouvez lancer des instances en exécutant plusieurs types d'instances dans plusieurs zones de disponibilité. Pour plus d'informations, consultez Groupes Auto Scaling combinant plusieurs types d'instances et options d'achat.
- Si vous utilisez plusieurs types d'instance, envisagez d'activer la fonction de Rééquilibrage de capacité. Ceci est utile si vous souhaitez que le service Amazon EC2 Ponctuel tente de lancer une nouvelle instance Spot avant qu'une instance en cours d'exécution ne soit résiliée. Pour plus d'informations, consultez Utiliser le rééquilibrage de la capacité pour gérer les interruptions Spot Amazon EC2.

Cause 3 : avec les blocs de capacité, Amazon EC2 met fin à toutes les instances encore en cours d'exécution 30 minutes avant l'heure de fin du bloc de capacité. Cette interruption abrupte pousse votre groupe Auto Scaling à essayer de lancer de nouvelles instances pour maintenir la capacité souhaitée, alors même que le bloc de capacité touche à sa fin.

Solution 3 : pour résoudre ce problème, essayez ce qui suit :

- Diminuez la capacité souhaitée du groupe Auto Scaling pour l'empêcher d'essayer de lancer de nouvelles instances. Pour plus d'informations, consultez Mise à l'échelle manuelle pour Amazon EC2 Auto Scaling.
- Assurez-vous de redimensionner votre groupe Auto Scaling 30 minutes avant l'heure de fin du Capacity Block afin de ne pas rencontrer cette erreur fréquemment. Assurez-vous que

tous les hooks du cycle de vie sont terminés 30 minutes avant la fin du bloc de capacité. Pour plus d'informations, consultez Utilisation Capacity Blocks pour les charges de travail liées à l'apprentissage automatique.

Une instance a été mise hors service en réponse à une défaillance de la surveillance de l'état du système ELB

Problème : les instances Auto Scaling peuvent réussir aux surveillances de l'état EC2. Cependant, elles peuvent échouer aux surveillances de l'état Elastic Load Balancing pour les groupes cibles ou les équilibreurs de charge classiques auprès desquels le groupe Auto Scaling est enregistré.

Cause 1 : si votre groupe Auto Scaling s'appuie sur les tests de santé fournis par Elastic Load Balancing, Amazon EC2 Auto Scaling détermine l'état de santé de vos instances en vérifiant les résultats des contrôles d'état EC2 et des tests de santé d'Elastic Load Balancing. L'équilibreur de charge effectue des surveillances de l'état en envoyant une requête à chaque instance et en attendant la réponse correcte, ou en établissant une connexion avec l'instance. Une instance peut ne pas réussir la surveillance de l'état Elastic Load Balancing, parce qu'une application s'exécutant sur l'instance connaît des problèmes faisant que l'équilibreur de charge considère l'instance comme étant hors service.

Solution 1 : pour réussir les surveillances de l'état Elastic Load Balancing :

- Vérifiez que les paramètres de surveillance de l'état de vos groupes cibles sont correctement configurés. Vous définissez des paramètres de surveillance de l'état de votre équilibreur de charge pour chaque groupe cible. Pour plus d'informations, consultez Configuration des contrôles de santé pour les cibles.
- Notez les codes de réussite attendus par l'équilibreur de charge et si votre application est configurée correctement pour renvoyer ces codes lorsque la surveillance de l'état est concluante.
- Vérifiez que les groupes de sécurité de votre équilibreur de charge et de votre groupe Auto Scaling sont correctement configurés.
- Vérifiez que l'équilibreur de charge est configuré dans les mêmes zones de disponibilité que votre groupe Auto Scaling.

Solution 2 : mettez à jour le groupe Auto Scaling pour désactiver les surveillances de l'état Elastic Load Balancing. Pour obtenir des instructions sur la façon de désactiver ces contrôles de santé, consultezAssociez un équilibreur de charge Elastic Load Balancing à votre groupe Auto Scaling.

Cause 2 : il y a une discordance entre la période de grâce de surveillance de l'état et l'heure de démarrage de l'instance.

Solution 3 : modifier le délai de grâce du bilan de santé de votre groupe Auto Scaling. Définissez la période de grâce sur une période suffisamment longue pour prendre en charge le nombre de tests de santé réussis consécutifs requis avant qu'Elastic Load Balancing considère qu'une instance nouvellement lancée est saine. Pour plus d'informations, consultez Définir la période de grâce de la surveillance de l'état pour un groupe Auto Scaling.

Ressources supplémentaires

Si vous rencontrez un autre problème, consultez les AWS re:Post articles suivants pour obtenir une aide supplémentaire en matière de résolution des problèmes :

- Pourquoi Amazon EC2 Auto Scaling a-t-il résilié une instance ?
- Pourquoi Amazon EC2 Auto Scaling ne résilie-t-il pas une instance en mauvais état ?

AWS Health Dashboard notifications pour Amazon EC2 Auto Scaling

Vous fournissez AWS Health Dashboard une assistance pour les notifications provenant d'Amazon EC2 Auto Scaling. Ces notifications vous permettent de mieux cerner les problèmes de performances ou de disponibilité des ressources susceptibles d'affecter vos applications. Seuls les événements spécifiques aux groupes de sécurité et aux modèles de lancement manquants sont actuellement disponibles.

Cela AWS Health Dashboard fait partie du AWS Health service. Il ne nécessite aucune configuration et peut être affiché par n'importe quel utilisateur authentifié dans votre compte. Pour plus d'informations, consultez Commencer à utiliser votre AWS Health tableau de bord.

Si vous recevez un message similaire aux messages suivants, il doit être traité comme une alarme pour agir.

Exemple : l'absence d'un groupe de sécurité empêche la montée en puissance du groupe Auto Scaling

```
Hello,
```

```
 At 2020-01-11 04:00 UTC, we detected an issue with your Auto Scaling group [ARN] in
Compte AWS 123456789012.

 A security group associated with this Auto Scaling group cannot be found. Each time
a
 scale out operation is performed, it will be prevented until you make a change that
 fixes the issue.

 We recommend that you review and update your Auto Scaling group configuration to
change
 the launch template or launch configuration that depends on the unavailable security
 group.

 Sincerely,
 Amazon Web Services
```

Exemple : l'absence d'un modèle de lancement empêche la montée en puissance du groupe Auto Scaling

```
 Hello,

 At 2021-05-11 04:00 UTC, we detected an issue with your Auto Scaling group [ARN] in
Compte AWS 123456789012.

 The launch template associated with this Auto Scaling group cannot be found. Each
time
 a scale out operation is performed, it will be prevented until you make a change
that
 fixes the issue.

 We recommend that you review and update your Auto Scaling group configuration and
 specify an existing launch template to use.

 Sincerely,
 Amazon Web Services
```

Surveillez CloudWatch les métriques de vos groupes et instances Auto Scaling

Les métriques sont le concept fondamental d'Amazon CloudWatch. Une métrique représente un ensemble chronologique de points de données publiés sur CloudWatch. Envisagez une métrique

comme une variable à surveiller et les points de données comme les valeurs de cette variable au fil du temps. Vous pouvez utiliser ces métriques pour vérifier que le système fonctionne comme prévu.

Les métriques Amazon EC2 Auto Scaling qui recueillent des informations relatives aux groupes Auto Scaling sont dans l'espace de noms `AWS/AutoScaling`. Les métriques d'instance Amazon EC2 qui collectent des données sur le processeur et d'autres données d'utilisation des instances Auto Scaling se trouvent dans`AWS/EC2`espace de noms.

La console Amazon EC2 Auto Scaling affiche une série de graphiques pour les métriques de groupe et les métriques d'instance agrégées pour le groupe. Selon vos besoins, vous préférerez peut-être accéder aux données de vos groupes et instances Auto Scaling depuis Amazon CloudWatch plutôt que depuis la console Amazon EC2 Auto Scaling.

Pour plus d'informations, consultez le guide de CloudWatch l'utilisateur Amazon.

Table des matières

Afficher des graphiques de surveillance dans la console Amazon EC2 Auto Scaling

Dans la section Amazon EC2 Auto Scaling de la console Amazon EC2, vous pouvez minute-by-minute suivre la progression de chaque groupe Auto Scaling à l'aide de métriques. CloudWatch

Vous pouvez surveiller les types de métriques suivants :

- Métriques Auto Scaling : les métriques Auto Scaling ne sont activées que lorsque vous les activez. Pour plus d'informations, consultez Activer les métriques du groupe Auto Scaling (console). Lorsque les métriques Auto Scaling sont activées, les graphiques de surveillance affichent les données publiées à une granularité d'une minute pour les métriques Auto Scaling.
- Métriques EC2 — Les métriques de l'instance Amazon EC2 sont toujours activées. Lorsque la surveillance détaillée est activée, les graphiques de surveillance affichent les données publiées à la minute pour les métriques de l'instance. Pour plus d'informations, consultez Configurer la surveillance pour les instances à scalabilité automatique.

Pour afficher les graphiques de surveillance à l'aide de la console Amazon EC2 Auto Scaling

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case en regard du groupe Auto Scaling pour lequel vous souhaitez afficher les métriques.

   Un volet fractionné s'ouvre dans la partie inférieure de la page Auto Scaling groups (Groupes Auto Scaling).
3. Sélectionnez l'onglet Monitoring (Surveillance).

   Amazon EC2 Auto Scaling affiche les graphiques de surveillance pour les métriques Auto Scaling.
4. Pour afficher les graphiques de surveillance des métriques d'instance agrégées du groupe, sélectionnez EC2.

Actions sur le graphique

- Survolez un point de données pour afficher une fenêtre contextuelle de données pour une heure spécifique en UTC.
- Pour agrandir un graphique, choisissez Enlarge (Agrandir) dans l'outil menu (les trois points verticaux) en haut à droite du graphique. Vous pouvez également choisir l'icône Maximiser en haut d'un graphique.
- Ajustez la durée d'affichage des données dans le graphique en sélectionnant l'une des valeurs de durée prédéfinies. Si le graphique est agrandi, vous pouvez choisir Custom (Personnalisé) pour définir votre propre durée.
- Choisissez Refresh (Actualiser) dans le menu outil pour mettre à jour les données d'un graphique.
- Faites glisser votre curseur sur les données du graphique pour sélectionner une plage spécifique. Vous pouvez alors choisir Apply time range (Appliquer la plage de temps) dans l'outil de menu.
- Choisissez Afficher les journaux dans le menu pour afficher les flux de journaux associés (le cas échéant) dans la CloudWatch console.
- Pour afficher un graphique dans CloudWatch, choisissez Afficher dans les métriques dans le menu. Cela vous amène à la CloudWatch page correspondant à ce graphique. Là, vous pouvez afficher plus d'informations ou accéder aux informations historiques pour mieux comprendre l'évolution de votre groupe Auto Scaling sur une période prolongée.

Graphique des métriques pour vos groupes Auto Scaling

Après avoir créé un groupe Auto Scaling, vous pouvez ouvrir la console Amazon EC2 Auto Scaling et afficher les graphiques de surveillance pour le groupe dans l'onglet Surveillance.

Dans la section Auto Scaling, le graphique des métriques comprend les métriques suivantes. Ces métriques fournissent des mesures qui peuvent être des indicateurs d'un problème potentiel, comme le nombre d'instances en cours de résiliation ou le nombre d'instances en attente. La section CloudWatch Métriques Amazon pour Amazon EC2 Auto Scaling fournit les définitions de ces métriques.

Nom d'affichage	CloudWatch nom de la métrique
Taille minimale du groupe	`GroupMinSize`
Taille maximale du groupe	`GroupMaxSize`
Capacité souhaitée	`GroupDesiredCapacity`
Instances en service	`GroupInServiceInstances`
Instances en attente	`GroupPendingInstances`
Instances en veille	`GroupStandbyInstances`
Instances résiliées	`GroupTerminatingInstances`
Total des instances	`GroupTotalInstances`

Dans la section EC2, vous trouverez le graphique des métriques suivantes, basé sur les principales mesures de performance de vos instances Amazon EC2. Ces métriques EC2 sont un agrégat de métriques pour toutes les instances du groupe. Vous trouverez les définitions de ces mesures dans la section Répertorier les CloudWatch mesures disponibles pour vos instances dans le guide de l'utilisateur Amazon EC2.

Nom d'affichage	CloudWatch nom de la métrique
Utilisation de l’UC	`CPUUtilization`

Nom d'affichage	CloudWatch nom de la métrique
Lectures sur disque	`DiskReadBytes`
Opérations de lecture sur disque	`DiskReadOps`
Ecritures sur disque	`DiskWriteBytes`
Opérations d'écriture sur disque	`DiskWriteOps`
Entrée réseau	`NetworkIn`
Sortie réseau	`NetworkOut`
Status Check Failed (Any) [Échec du contrôle de statut (tous)]	`StatusCheckFailed`
Status Check Failed (Instance) [Échec du contrôle de statut (instance)]	`StatusCheckFailed_Instance`
Status Check Failed (System) [Échec du contrôle de statut (système)]	`StatusCheckFailed_System`

En outre, certaines métriques sont disponibles pour des cas d'utilisation spécifiques dans le graphique de métriques Auto Scaling.

Les métriques suivantes sont utiles si votre groupe utilise des pondérations déterminant la contribution, en nombre d'unités, de chaque instance à la capacité souhaitée du groupe. La section CloudWatch Métriques Amazon pour Amazon EC2 Auto Scaling fournit les définitions de ces métriques.

Nom d'affichage	CloudWatch nom de la métrique
Unités de capacité en service	`GroupInServiceCapacity`
Unités de capacité en attente	`GroupPendingCapacity`
Unités de capacité de sauvegarde	`GroupStandbyCapacity`
Unités de capacité résiliées	`GroupTerminatingCapacity`
Unités de capacité totales	`GroupTotalCapacity`

Les statistiques suivantes sont utiles si votre groupe utilise la fonctionnalité Groupe chaud. La section CloudWatch Métriques Amazon pour Amazon EC2 Auto Scaling fournit les définitions de ces métriques.

Nom d'affichage	CloudWatch nom de la métrique
Taille minimale du groupe chaud	`WarmPoolMinSize`
Capacité souhaitée du groupe chaud	`WarmPoolDesiredCapacity`
Unités de capacité en attente du groupe chaud	`WarmPoolPendingCapacity`
Unités de capacité en cours de résiliation du groupe chaud	`WarmPoolTerminatingCapacity`
Unités de capacité réchauffé es du groupe chaud	`WarmPoolWarmedCapacity`
Total des unités de capacité lancées du groupe chaud	`WarmPoolTotalCapacity`

Nom d'affichage	CloudWatch nom de la métrique
Capacité souhaitée du groupe et du groupe chaud	`GroupAndWarmPoolDesiredCapacity`
Total des unités de capacité lancées du groupe et du groupe chaud	`GroupAndWarmPoolTotalCapacity`

Ressources connexes

- Pour surveiller les métriques par instance, consultez Graph les métriques de vos instances dans le guide de l'utilisateur Amazon EC2.
- CloudWatch les tableaux de bord sont des pages d'accueil personnalisables dans la CloudWatch console. Vous pouvez utiliser ces pages pour surveiller vos ressources dans une seule vue, y compris les ressources réparties sur différentes régions. Vous pouvez utiliser CloudWatch des tableaux de bord pour créer des vues personnalisées des mesures et des alarmes relatives à vos AWS ressources. Pour plus d'informations, consultez le guide de CloudWatch l'utilisateur Amazon.

CloudWatch Métriques Amazon pour Amazon EC2 Auto Scaling

Amazon EC2 Auto Scaling publie les métriques suivantes dans l'espace de noms `AWS/AutoScaling`. Les métriques de groupe Auto Scaling réellement mises à disposition dépendront de l'activation ou non des métriques de groupe et des métriques de groupe que vous avez activées. Les métriques de groupe sont disponibles à une granularité d'une minute sans frais supplémentaires, mais vous devez les activer.

Lorsque vous activez les métriques de groupe Auto Scaling, Amazon EC2 Auto Scaling envoie des échantillons de données par minute, dans CloudWatch la mesure du possible. Dans de rares cas, en CloudWatch cas d'interruption de service, les données ne sont pas remplies pour combler les lacunes dans l'historique des indicateurs de groupe.

Table des matières

- Activer les métriques du groupe Auto Scaling (console)
- Activer les métriques du groupe Auto Scaling (AWS CLI)

Métriques du groupe Auto Scaling

Avec ces métriques, vous bénéficiez d'une visibilité quasi-continue de l'historique de votre groupe Auto Scaling, comme l'évolution de la taille du groupe au fil du temps.

Métrique	Description
`GroupMinSize`	Taille minimale du groupe Auto Scaling. Critères de reporting : signalé si la collecte de métriques est activée.
`GroupMaxSize`	Taille maximale du groupe Auto Scaling. Critères de reporting : signalé si la collecte de métriques est activée.
`GroupDesiredCapacity`	Nombre d'instances que le groupe Auto Scaling tente de gérer. Critères de reporting : signalé si la collecte de métriques est activée.
`GroupInServiceInstances`	Nombre d'instances qui sont en cours d'exécution dans le cadre du groupe Auto Scaling. Cette métrique n'inclut pas les instances qui sont en suspens ou en cours de résiliation. Critères de reporting : signalé si la collecte de métriques est activée.
`GroupPendingInstances`	Nombre d'instances qui sont en suspens. Une instance en suspens n'est pas encore en service. Cette métrique n'inclut pas les instances qui sont en service ou en cours de résiliation. Critères de reporting : signalé si la collecte de métriques est activée.

Métrique	Description
`GroupStandbyInstances`	Nombre d'instances qui sont à l'état `Standby`. Les instances qui sont dans cet état sont toujours en cours d'exécution, mais ne sont pas activement en service. Critères de reporting : signalé si la collecte de métriques est activée.
`GroupTerminatingInstances`	Nombre d'instances qui sont en cours de résiliation. Cette métrique n'inclut pas les instances qui sont en service ou en suspens. Critères de reporting : signalé si la collecte de métriques est activée.
`GroupTotalInstances`	Nombre total d'instances dans le groupe Auto Scaling. Cette métrique identifie le nombre d'instances qui sont en service, en suspens et en cours de résiliation. Critères de reporting : signalé si la collecte de métriques est activée.

Lorsque vous configurez un groupe d’instances mixtes pour mesurer la capacité souhaitée dans différentes unités, par exemple en attribuant des pondérations basées sur le nombre de vCPU de chaque type d’instance, les métriques suivantes comptent le nombre d’unités utilisées par votre groupe Auto Scaling. Si vous n’avez pas configuré de groupe d’instances mixtes pour mesurer la capacité souhaitée dans différentes unités, les métriques suivantes sont renseignées, mais elles sont égales aux métriques définies dans le tableau précédent. Pour plus d’informations, consultez Présentation de la configuration.

Métrique	Description
`GroupInServiceCapacity`	Nombre d'unités de capacité exécutées dans le cadre du groupe Auto Scaling.

Métrique	Description
	Critères de reporting : signalé si la collecte de métriques est activée.
`GroupPendingCapacity`	Nombre d'unités de capacité en attente. Critères de reporting : signalé si la collecte de métriques est activée.
`GroupStandbyCapacity`	Nombre d'unités de capacité qui sont dans un état `Standby`. Critères de reporting : signalé si la collecte de métriques est activée.
`GroupTerminatingCapacity`	Nombre d'unités de capacité en cours de terminaison. Critères de reporting : signalé si la collecte de métriques est activée.
`GroupTotalCapacity`	Nombre total d'unités de capacité dans le groupe Auto Scaling. Critères de reporting : signalé si la collecte de métriques est activée.

Amazon EC2 Auto Scaling fournit également les métriques suivantes pour les groupes Auto Scaling disposant d'un groupe d'instances pré-initialisées. Pour plus d'informations, consultez Groupes d'instances pré-initialisées pour Amazon EC2 Auto Scaling.

Métrique	Description
`WarmPoolMinSize`	La taille minimale du groupe chaud. Critères de reporting : signalé si la collecte de métriques est activée.
`WarmPoolDesiredCapacity`	La quantité de capacité qu'Amazon EC2 Auto Scaling tente de maintenir dans le groupe chaud.

Métrique	Description
	Cela équivaut à la taille maximale du groupe Auto Scaling moins sa capacité souhaitée ou, si elle est définie, à la capacité maximale préparée du groupe Auto Scaling moins sa capacité souhaitée. Toutefois, lorsque la taille minimale du groupe chaud est égale ou supérieure à la différence entre la taille maximale (ou, si elle est définie, la capacité maximale préparée) et la capacité souhaitée du groupe Auto Scaling, alors la capacité souhaitée du groupe chaud sera équivalente à la `WarmPoolMinSize` . Critères de reporting : signalé si la collecte de métriques est activée.
`WarmPoolPendingCapacity`	La quantité de capacité dans le groupe chaud qui est en attente. Cette métrique n'inclut pas les instances qui sont en cours d'exécution, arrêtées ou en cours de résiliation. Critères de reporting : signalé si la collecte de métriques est activée.
`WarmPoolTerminatingCapacity`	La quantité de capacité dans le groupe chaud qui est en cours de résiliation. Cette métrique n'inclut pas les instances qui sont en cours d'exécution, arrêtées ou en attente. Critères de reporting : signalé si la collecte de métriques est activée.
`WarmPoolWarmedCapacity`	La quantité de capacité disponible pour entrer dans le groupe Auto Scaling pendant la montée en puissance. Cette métrique n'inclut pas les instances qui sont en suspens ou en cours de résiliation. Critères de reporting : signalé si la collecte de métriques est activée.

Métrique	Description
`WarmPoolTotalCapacity`	La capacité totale du groupe chaud, y compris les instances qui sont en cours d'exécution, arrêtées, en attente ou en cours de résiliation. Critères de reporting : signalé si la collecte de métriques est activée.
`GroupAndWarmPoolDesiredCapacity`	La capacité souhaitée du groupe Auto Scaling et du groupe chaud combinés. Critères de reporting : signalé si la collecte de métriques est activée.
`GroupAndWarmPoolTotalCapacity`	La capacité totale du groupe Auto Scaling et du groupe chaud combinés. Ceci inclut les instances qui sont en cours d'exécution, arrêtées, en attente, en cours de résiliation ou en service. Critères de reporting : signalé si la collecte de métriques est activée.

Dimensions pour les métriques du groupe Auto Scaling

Vous pouvez utiliser les dimensions suivantes pour affiner les métriques répertoriées dans les tableaux précédents.

Dimension	Description
`AutoScalingGroupName`	Filtre sur le nom d'un groupe Auto Scaling.

Métriques et dimensions de mise à l'échelle

Le nom d'espace `AWS/AutoScaling` comprend les métriques suivantes pour une mise à l'échelle prédictive.

Les mesures sont disponibles avec une résolution d'une heure.

Vous pouvez évaluer la précision des prévisions en comparant les valeurs prévues aux valeurs réelles. Pour plus d'informations sur l'évaluation de la précision des prévisions avec ces métriques, consultezSurveillez les mesures de dimensionnement prédictives avec CloudWatch.

Métrique	Description	Dimensions
`PredictiveScalingLoadForecast`	La quantité de charge qui devrait être générée par votre application. Le`Average`,`Minimum`, et`Maximum`les statistiques sont utiles, mais`Sum`la statistique ne l'est pas. Critères de déclaration : Reporté après la création de la prévision initiale.	`AutoScalingGroupName` , `PolicyName` , `PairIndex`
`PredictiveScalingCapacityForecast`	La quantité prévue de capacité nécessaire pour répondre à la demande des applications. Ceci est basé sur la prévision de charge et le niveau d'utilisation cible auxquels vous souhaitez maintenir vos instances Auto Scaling. Les statistiques `Average`, `Minimum` et `Maximum` statistiques sont utiles, mais la statistique `Sum` ne l'est pas. Critères de déclaration : Reporté après la création de la prévision initiale.	`AutoScalingGroupName` , `PolicyName`
`PredictiveScalingMetricPairCorrelation`	Corrélation entre la métrique de mise à l'échelle et la moyenne par instance de la métrique de charge. La mise à l'échelle prédictive suppose une corrélation élevée. Par conséquent, si vous observez une valeur faible pour cette métrique, il est préférable de ne pas utiliser de paire de métriques.	`AutoScalingGroupName` , `PolicyName` , `PairIndex`

Métrique	Description	Dimensions
	Les statistiques `Average`, `Minimum` et `Maximum` statistiques sont utiles, mais la statistique `Sum` ne l'est pas. Critères de déclaration : Reporté après la création de la prévision initiale.	

> **Note**
>
> Le`PairIndex`renvoie des informations associées à l'index de la paire de métriques de mise à l'échelle de charge telle qu'attribuée par Amazon EC2 Auto Scaling. Actuellement, la seule valeur valide est `0`.

Activer les métriques du groupe Auto Scaling (console)

Pour activer les métriques du groupe

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Sous l'onglet Surveillance, cochez la case Activer de l'option Collecte des métriques du groupe Auto Scaling en haut de la page, sous Auto Scaling.

Pour désactiver les métriques du groupe

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Sélectionnez votre groupe Auto Scaling.
3. Sous l'onglet Surveillance, décochez la case Activer de l'option Collecte des métriques du groupe Auto Scaling.

Activer les métriques du groupe Auto Scaling (AWS CLI)

Pour activer les métriques du groupe Auto Scaling

Activez une ou plusieurs métriques de groupe avec la commande enable-metrics-collection. Par exemple, la commande suivante active une métrique unique pour le groupe Auto Scaling indiqué.

```
aws autoscaling enable-metrics-collection --auto-scaling-group-name my-asg \
  --metrics GroupDesiredCapacity --granularity "1Minute"
```

Si vous omettez l'option `--metrics`, toutes les métriques sont activées.

```
aws autoscaling enable-metrics-collection --auto-scaling-group-name my-asg \
  --granularity "1Minute"
```

Pour désactiver les métriques du groupe Auto Scaling

Utilisez la commande disable-metrics-collection pour désactiver toutes les métriques de groupe.

```
aws autoscaling disable-metrics-collection --auto-scaling-group-name my-asg
```

Configurer la surveillance pour les instances à scalabilité automatique

Amazon EC2 recueille et traite les données brutes des instances en métriques lisibles et disponibles presque en temps réel qui décrivent le processeur et d'autres données d'utilisation pour votre groupe Auto Scaling. Vous pouvez configurer l'intervalle de surveillance de ces mesures en choisissant une granularité d'une minute ou de cinq minutes.

La surveillance de l'instance est activée chaque fois qu'une instance est lancée, soit par le moyen d'une surveillance basique (avec une granularité de cinq minutes), soit par surveillance détaillée (avec une granularité d'une minute). Pour la surveillance détaillée, des frais supplémentaires s'appliquent. Pour plus d'informations, consultez la section CloudWatch Tarification Amazon et surveillance de vos instances CloudWatch à l'aide du guide de l'utilisateur Amazon EC2.

Avant de créer un groupe Auto Scaling, vous devez créer une configuration ou un modèle de lancement qui autorise le type de surveillance approprié vers votre application. Si vous ajoutez une stratégie de mesure à votre groupe, nous vous recommandons vivement d'utiliser la surveillance détaillée pour obtenir les données de mesure pour les instances EC2 avec une minute de mesure, car cela permet de réagir plus rapidement aux changements de charge.

Table des matières

Activer la surveillance détaillée (console)

Par défaut, la surveillance de base est activée lorsque vous utilisez le AWS Management Console pour créer un modèle de lancement ou une configuration de lancement.

Pour activer la surveillance détaillée dans un modèle de lancement

Lorsque vous créez le modèle de lancement en utilisant AWS Management Console, dans la section Détails avancés, pour une CloudWatchsurveillance détaillée, choisissez Activer. Sinon, la surveillance de base est activée. Pour plus d'informations, consultez Créer un modèle de lancement à l'aide de paramètres avancés.

Pour activer la surveillance détaillée dans une configuration de lancement

Lorsque vous créez la configuration de lancement à l'aide de AWS Management Console, dans la section Configuration supplémentaire, sélectionnez Activer la surveillance détaillée de l'instance EC2 dans CloudWatch. Sinon, la surveillance de base est activée. Pour plus d'informations, consultez Créez une configuration de lancement.

Activer la surveillance détaillée (AWS CLI)

Par défaut, la surveillance basique est activée lorsque vous créez un modèle de lancement à l'aide de l'interface AWS CLI. La surveillance détaillée est activée par défaut lorsque vous créez une configuration de lancement à l'aide de l'interface AWS CLI.

Pour activer la surveillance détaillée dans un modèle de lancement

Pour les modèles de lancement, utilisez la commande create-launch-template et transmettez un fichier JSON contenant les informations de création du modèle de lancement. Définissez l'attribut de surveillance sur `"Monitoring":{"Enabled":true}` pour activer la surveillance détaillée ou sur `"Monitoring":{"Enabled":false}` pour activer la surveillance de base.

Pour activer la surveillance détaillée dans une configuration de lancement

Pour les configurations de lancement, utilisez la commande create-launch-configuration avec l'option `--instance-monitoring`. Définissez cette option sur `true` pour activer la surveillance détaillée ou sur `false` pour activer la surveillance de base.

```
--instance-monitoring Enabled=true
```

Basculer entre la surveillance basique et la surveillance détaillée (ou inversement)

Pour modifier le type de surveillance activé sur les nouvelles instances EC2, mettez à jour le modèle de lancement ou le groupe Auto Scaling de manière à utiliser un nouveau modèle de lancement ou une nouvelle configuration de lancement. Les instances existantes continuent d'utiliser le type de surveillance précédemment activé. Pour procéder à la mise à jour de toutes les instances, résiliez-les afin qu'elles soient remplacées par votre groupe Auto Scaling, ou mettez-les à jour individuellement à l'aide des commandes monitor-instances et unmonitor-instances.

> **Note**
>
> Grâce aux fonctions d'actualisation d'instance et de durée de vie maximale, vous pouvez également remplacer toutes les instances du groupe Auto Scaling pour lancer de nouvelles instances qui utilisent les nouveaux paramètres. Pour plus d'informations, consultez Recycler les instances de votre groupe Auto Scaling.

Lorsque vous passez de la surveillance basique à la surveillance détaillée (ou inversement) :

Si des CloudWatch alarmes sont associées aux politiques de dimensionnement par étapes ou à des politiques de dimensionnement simples pour votre groupe Auto Scaling, utilisez la commande put-metric-alarm pour mettre à jour chaque alarme. Faites correspondre chaque période avec son type de surveillance (300 secondes pour une surveillance de base et 60 secondes pour une surveillance détaillée). Si vous passez d'une surveillance détaillée à une surveillance basique sans mettre à jour les alarmes pour qu'elles correspondent à la période de cinq minutes, les alarmes continuent de vérifier les statistiques toutes les minutes. Elles risquent de ne trouver aucune donnée disponible pendant près de quatre périodes sur cinq.

Collectez des métriques supplémentaires à l'aide de l' CloudWatch agent

Pour collecter des métriques au niveau du système d'exploitation, telles que la mémoire disponible et utilisée, vous devez installer l' CloudWatch agent. Des frais supplémentaires peuvent s'appliquer.

Vous pouvez utiliser l' CloudWatch agent pour collecter à la fois les métriques du système et les fichiers journaux à partir des instances Amazon EC2. Pour plus d'informations, consultez la section Mesures collectées par l' CloudWatch agent dans le guide de CloudWatch l'utilisateur Amazon.

Enregistrez les appels d'API Amazon EC2 Auto Scaling avec AWS CloudTrail

Amazon EC2 Auto Scaling est intégré à AWS CloudTrail un service qui fournit un enregistrement des actions entreprises par un utilisateur, un rôle ou un service à l'aide d'Amazon EC2 Auto Scaling. CloudTrail capture tous les appels d'API pour Amazon EC2 Auto Scaling sous forme d'événements. Les appels capturés incluent les appels depuis la console Amazon EC2 Auto Scaling et les appels de code vers l'API Amazon EC2 Auto Scaling.

Si vous créez un suivi, vous pouvez activer la diffusion continue d' CloudTrail événements vers un compartiment Amazon S3, y compris des événements pour Amazon EC2 Auto Scaling. Si vous ne configurez pas de suivi, vous pouvez toujours consulter les événements les plus récents dans la CloudTrail console dans Historique des événements. À l'aide des informations collectées par CloudTrail, vous pouvez déterminer la demande envoyée à Amazon EC2 Auto Scaling, l'adresse IP à partir de laquelle la demande a été effectuée, l'auteur de la demande, la date à laquelle elle a été faite, ainsi que des informations supplémentaires.

Pour en savoir plus CloudTrail, consultez le guide de AWS CloudTrail l'utilisateur.

Informations sur Amazon EC2 Auto Scaling dans CloudTrail

CloudTrail est activé sur votre compte Amazon Web Services lorsque vous créez le compte. Lorsqu'une activité se produit dans Amazon EC2 Auto Scaling, cette activité est enregistrée dans un CloudTrail événement avec d'autres événements Amazon Web Services dans l'historique des événements. Vous pouvez afficher, rechercher et télécharger les événements récents dans votre compte Amazon Web Services. Pour plus d'informations, consultez la section Affichage des événements avec l'historique des CloudTrail événements.

Pour un enregistrement continu des événements dans votre compte Amazon Web Services, y compris les événements relatifs à Amazon EC2 Auto Scaling, créez un journal d'activité. Un suivi permet CloudTrail de fournir des fichiers journaux à un compartiment Amazon S3. Par défaut, lorsque vous créez un journal de suivi dans la console, il s’applique à toutes les régions . Le journal d'activité consigne les événements de toutes les régions dans la partition Amazon Web Services

et livre les fichiers journaux au compartiment Amazon S3 de votre choix. En outre, vous pouvez configurer d'autres Amazon Web Services pour analyser plus en détail les données d'événements collectées dans les CloudTrail journaux et agir en conséquence. Pour plus d’informations, consultez les ressources suivantes :

- Présentation de la création d’un journal de suivi
- CloudTrail services et intégrations pris en charge
- Configuration des notifications Amazon SNS pour CloudTrail
- Réception de fichiers CloudTrail journaux de plusieurs régions et réception de fichiers CloudTrail journaux de plusieurs comptes

Toutes les actions Amazon EC2 Auto Scaling sont enregistrées CloudTrail et documentées dans le manuel Amazon EC2 Auto Scaling API Reference. Par exemple, les appels aux CreateLaunchConfigurationDescribeAutoScalingGroup, et UpdateAutoScalingGrouples actions génèrent des entrées dans les fichiers CloudTrail journaux.

Chaque événement ou entrée de journal contient des informations sur la personne ayant initié la demande. Les informations relatives à l’identité permettent de déterminer les éléments suivants :

- Si la demande a été faite avec les informations d'identification de l'utilisateur root ou AWS Identity and Access Management (IAM).
- Si la demande a été effectuée avec les informations d’identification de sécurité temporaires d’un rôle ou d’un utilisateur fédéré.
- Si la demande a été effectuée par un autre service .

Pour plus d'informations, consultez l'CloudTrail`userIdentity`élément.

Présenter des entrées des fichiers journaux Amazon EC2 Auto Scaling

Un suivi est une configuration qui permet de transmettre des événements sous forme de fichiers journaux à un compartiment Amazon S3 que vous spécifiez. CloudTrail les fichiers journaux contiennent une ou plusieurs entrées de journal. Un événement représente une demande unique provenant de n'importe quelle source et inclut des informations sur l'action demandée, la date et l'heure de l'action, les paramètres de la demande, etc. CloudTrail les fichiers journaux ne constituent pas une trace ordonnée des appels d'API publics, ils n'apparaissent donc pas dans un ordre spécifique.

L'exemple suivant montre une entrée de CloudTrail journal illustrant l'CreateLaunchConfigurationaction.


```
{
    "eventVersion": "1.05",
    "userIdentity": {
        "type": "Root",
        "principalId": "123456789012",
        "arn": "arn:aws:iam::123456789012:root",
        "accountId": "123456789012",
        "accessKeyId": "AKIAIOSFODNN7EXAMPLE",
        "sessionContext": {
            "attributes": {
                "mfaAuthenticated": "false",
                "creationDate": "2018-08-21T17:05:42Z"
            }
        }
    },
    "eventTime": "2018-08-21T17:07:49Z",
    "eventSource": "autoscaling.amazonaws.com",
    "eventName": "CreateLaunchConfiguration",
    "awsRegion": "us-west-2",
    "sourceIPAddress": "192.0.2.0",
    "userAgent": "Coral/Jakarta",
    "requestParameters": {
        "ebsOptimized": false,
        "instanceMonitoring": {
            "enabled": false
        },
        "instanceType": "t2.micro",
        "keyName": "EC2-key-pair-oregon",
        "blockDeviceMappings": [
            {
                "deviceName": "/dev/xvda",
                "ebs": {
                    "deleteOnTermination": true,
                    "volumeSize": 8,
                    "snapshotId": "snap-01676e0a2c3c7de9e",
                    "volumeType": "gp2"
                }
            }
        ],
        "launchConfigurationName": "launch_configuration_1",
```

```
        "imageId": "ami-6cd6f714d79675a5",
        "securityGroups": [
            "sg-00c429965fd921483"
        ]
    },
    "responseElements": null,
    "requestID": "0737e2ea-fb2d-11e3-bfd8-99133058e7bb",
    "eventID": "3fcfb182-98f8-4744-bd45-b38835ab61cb",
    "eventType": "AwsApiCall",
    "recipientAccountId": "123456789012"
}
```


Ressources connexes

Avec CloudWatch Logs, vous pouvez surveiller et recevoir des alertes pour des événements spécifiques capturés par CloudTrail. Les événements envoyés à CloudWatch Logs sont ceux configurés pour être enregistrés par votre parcours. Assurez-vous donc d'avoir configuré votre ou vos sentiers pour enregistrer les types d'événements que vous souhaitez surveiller. CloudWatch Les journaux peuvent surveiller les informations contenues dans les fichiers journaux et vous avertir lorsque certains seuils sont atteints. Vous pouvez également archiver vos données de journaux dans une solution de stockage hautement durable. Pour plus d'informations, consultez le guide de l'utilisateur Amazon CloudWatch Logs et la rubrique Surveillance des fichiers CloudTrail CloudWatch journaux avec Amazon Logs du guide de AWS CloudTrail l'utilisateur.

Options de notification Amazon SNS pour Amazon EC2 Auto Scaling

Vous pouvez configurer votre groupe Auto Scaling pour qu'il vous informe des événements importants qui affectent votre application. Grâce aux notifications, vous pouvez également supprimer le sondage et vous ne rencontrerez pas l'`RequestLimitExceeded`erreur qui résulte parfois d'un sondage.

Il existe deux manières de recevoir des notifications concernant Amazon EC2 Auto Scaling :

- Amazon Simple Notification Service : Amazon SNS peut vous avertir lorsque votre groupe Auto Scaling lance ou met fin à des instances. Vous pouvez seulement activer ou désactiver les notifications Amazon SNS. Pour plus d’informations, consultez Amazon SNS et Amazon EC2 Auto Scaling.

- Amazon EventBridge : EventBridge fournit des notifications plus avancées, basées sur des événements, correspondant à des critères spécifiques et envoyées à diverses cibles, y compris Amazon SNS. EventBridge peut également surveiller un plus large éventail d'événements Auto Scaling pour une surveillance plus précise. Pour plus d'informations, consultez EventBridge À utiliser pour gérer les événements Auto Scaling.

Vous pouvez également effectuer une action personnalisée lorsqu'une instance entre dans un état d'attente lors du lancement ou de la cessation en utilisant des hooks et des services de cycle de vie tels qu' EventBridgeAmazon SNS et Amazon SQS. Les hooks du cycle de vie peuvent également permettre à une nouvelle instance de disposer de plus de temps pour exécuter un script spécifié dans les données utilisateur avant qu'Amazon EC2 Auto Scaling n'ajoute l'instance au groupe. Pour plus d'informations, consultez Hooks de cycle de vie Amazon EC2 Auto Scaling.

Amazon SNS et Amazon EC2 Auto Scaling

Cette section explique comment utiliser Amazon SNS pour surveiller le moment où votre groupe Auto Scaling lance ou met fin à des instances.

Par exemple, si vous configurez un groupe Auto Scaling pour qu'il utilise le type de notification `autoscaling: EC2_INSTANCE_TERMINATE`, et que ce groupe résilie une instance, il envoie une notification par e-mail. Cet e-mail contient les détails de l'instance résiliée, comme son ID et la raison de sa résiliation.

Notez qu'au fur et à mesure qu'Amazon EC2 Auto Scaling ajoute ou supprime des instances du groupe, des notifications concernant ces modifications vous sont envoyées, avec une notification envoyée par instance. Toutefois, l'envoi de ces notifications se fait dans la mesure du possible, et vos instances peuvent toujours échouer après la notification initiale, par exemple en cas d'échec d'un bilan de santé ultérieur. Ainsi, même si Amazon EC2 Auto Scaling vous en informe dans un premier temps, une instance peut tout de même échouer ultérieurement. Notez que vous pouvez configurer la durée d'attente d'Amazon EC2 Auto Scaling après le lancement d'une instance avant d'effectuer le premier bilan de santé. Pour plus d'informations, consultez Définir la période de grâce de la surveillance de l'état pour un groupe Auto Scaling.

Pour plus d'informations sur Amazon SNS en général, consultez le guide du développeur Amazon Simple Notification Service.

Table des matières

Notifications SNS

Amazon EC2 Auto Scaling prend en charge l'envoi de notifications Amazon SNS lorsque les événements suivants se produisent.

Événement	Description
`autoscaling:EC2_INSTANCE_LAUNCH`	Lancement de l'instance réussi
`autoscaling:EC2_INSTANCE_LAUNCH_ERROR`	Échec du lancement de l'instance
`autoscaling:EC2_INSTANCE_TERMINATE`	Mise hors service de l'instance réussie
`autoscaling:EC2_INSTANCE_TERMINATE_ERROR`	Échec de la mise hors service de l'instance

Le message comprend les informations suivantes :

- `Event` : événement.
- `AccountId` : ID du compte Amazon Web Services.
- `AutoScalingGroupName` : nom du groupe Auto Scaling.
- `AutoScalingGroupARN` : ARN du groupe Auto Scaling.
- `EC2InstanceId` : ID de l'instance EC2.

Par exemple :

```
Service: AWS Auto Scaling
Time: 2016-09-30T19:00:36.414Z
RequestId: 4e6156f4-a9e2-4bda-a7fd-33f2ae528958
Event: autoscaling:EC2_INSTANCE_LAUNCH
AccountId: 123456789012
AutoScalingGroupName: my-asg
AutoScalingGroupARN: arn:aws:autoscaling:region:123456789012:autoScalingGroup...
ActivityId: 4e6156f4-a9e2-4bda-a7fd-33f2ae528958
Description: Launching a new EC2 instance: i-0598c7d356eba48d7
Cause: At 2016-09-30T18:59:38Z a user request update of AutoScalingGroup constraints
 to ...
StartTime: 2016-09-30T19:00:04.445Z
EndTime: 2016-09-30T19:00:36.414Z
StatusCode: InProgress
StatusMessage:
Progress: 50
EC2InstanceId: i-0598c7d356eba48d7
Details: {"Subnet ID":"subnet-id","Availability Zone":"zone"}
Origin: AutoScalingGroup
Destination: EC2
```

Configuration des notifications Amazon SNS pour Amazon EC2 Auto Scaling

Pour utiliser Amazon SNS afin d'envoyer des notifications par e-mail, vous devez d'abord créer une rubrique, puis abonner les adresses e-mail requises à cette rubrique.

Créer une rubrique Amazon SNS

Une rubrique SNS est un point d'accès logique, un canal de communication utilisé par le groupe Auto Scaling pour envoyer les notifications. Pour créer une rubrique, donnez-lui un nom.

Lorsque vous créez un nom de rubrique, celui-ci doit répondre aux critères suivants :

- Contenir 1 à 256 caractères
- Contenir des lettres majuscules et minuscules ASCII, des chiffres, des traits de soulignement ou de traits d'union

Pour plus d'informations, consultez Création d'une rubrique Amazon SNS dans le Guide du développeur Amazon Simple Notification Service.

Abonner à la rubrique Amazon SNS

Pour recevoir les notifications que votre groupe Auto Scaling envoie à la rubrique, vous devez abonner un point de terminaison à cette dernière. Dans cette procédure, sous Point de terminaison, spécifiez l'adresse e-mail à laquelle vous souhaitez recevoir les notifications envoyées par Amazon EC2 Auto Scaling.

Pour plus d'informations, consultez Abonnement à une rubrique Amazon SNS dans le Guide du développeur Amazon Simple Notification Service.

Confirmer votre abonnement Amazon SNS

Amazon SNS envoie un e-mail de confirmation à l'adresse que vous avez spécifiée à l'étape précédente.

Avant de passer à l'étape suivante, ouvrez l'e-mail envoyé par AWS Notifications et cliquez sur le lien de confirmation de l'abonnement.

Vous recevrez un message d'accusé de réception de. AWS Amazon SNS est maintenant configuré pour recevoir des notifications et envoyer la notification par e-mail à l'adresse spécifiée.

Configurer le groupe Auto Scaling pour qu'il envoie des notifications

Vous pouvez configurer le groupe Auto Scaling pour qu'il envoie des notifications à Amazon SNS lorsqu'un événement de mise à l'échelle, comme le lancement ou la résiliation d'instances, se produit. Amazon SNS envoie une notification contenant des informations sur les instances à l'adresse e-mail spécifiée.

Pour configurer des notifications Amazon SNS pour votre groupe Auto Scaling (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre dans la partie inférieure de la page avec des informations sur le groupe sélectionné.
3. Sous l'onglet Activité, choisissez Notifications d'activité et Créer une notification.
4. Dans le volet Create notifications (Créer des notifications), procédez comme suit :

   a. Dans le champ Rubrique SNS, sélectionnez votre rubrique SNS.

b. Dans le champ Types d'événements, sélectionnez les événements pour lesquels vous souhaitez envoyer des notifications.

c. Choisissez Créer.

Pour configurer des notifications Amazon SNS pour votre groupe Auto Scaling (AWS CLI)

Utilisez la commande put-notification-configuration suivante.

```
aws autoscaling put-notification-configuration --auto-scaling-group-name my-
asg --topic-arn arn --notification-types "autoscaling:EC2_INSTANCE_LAUNCH"
 "autoscaling:EC2_INSTANCE_TERMINATE"
```

Tester la notification

Pour générer une notification pour un événement de lancement, mettez à jour le groupe Auto Scaling en augmentant de 1 la capacité souhaitée du groupe Auto Scaling. Vous recevez une notification quelques minutes après le lancement de l'instance.

Pour modifier la capacité souhaitée (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre dans la partie inférieure de la page des Auto Scaling groups (groupes Auto Scaling) avec des informations sur le groupe sélectionné.
3. Sous l'onglet Details (Détails) choisissez Group details (Détails du groupe), Edit (Modifier).
4. Pour Desired capacity (Capacité désirée), augmentez la valeur actuelle de 1. Si cette valeur dépasse la Maximum capacity (capacité maximale), vous devez également augmenter la valeur de la Maximum capacité (capacité maximale) de 1.
5. Choisissez Mettre à jour.
6. Après quelques minutes, vous recevrez une notification pour l'événement. Si vous n'avez pas besoin de l'instance supplémentaire que vous avez lancée pour ce test, vous pouvez réduire la Desired capacity (capacité souhaitée) de 1. Après quelques minutes, vous recevrez une notification pour l'événement.

Supprimer une configuration de notification

Si vous n'utilisez plus une configuration de notification Amazon EC2 Auto Scaling, vous pouvez la supprimer.

Pour supprimer une configuration de notification Amazon EC2 Auto Scaling (console)

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Sélectionnez votre groupe Auto Scaling.
3. Sous l'onglet Activité, cochez la case située en regard de la notification que vous souhaitez supprimer, puis choisissez Actions, Supprimer.

Pour supprimer une configuration de notification Amazon EC2 Auto Scaling (AWS CLI)

Utilisez la commande delete-notification-configuration suivante.

```
aws autoscaling delete-notification-configuration --auto-scaling-group-name my-asg --
topic-arn arn
```

Pour plus d'informations sur la suppression de la rubrique Amazon SNS et de tous les abonnements associés à un groupe Auto Scaling, consultez Suppression d'un abonnement et d'une rubrique Amazon SNS dans le Guide du développeur Amazon Simple Notification Service.

Stratégie de clé pour une rubrique Amazon SNS chiffrée

La rubrique Amazon SNS que vous indiquez peut être chiffrée à l'aide d'une clé gérée par le client, créée avec le AWS Key Management Service. Pour autoriser Amazon EC2 Auto Scaling à publier sur des rubriques chiffrées, vous devez d'abord créer votre clé KMS, puis ajouter l'instruction suivante à la stratégie de la clé KMS. Remplacez l'ARN en exemple par l'ARN du rôle approprié lié à un service qui est autorisé à accéder à la clé. Pour plus d'informations, consultez Configurer les autorisations AWS KMS dans le Guide du développeur Amazon Simple Notification Service.

Dans cet exemple, la déclaration de politique donne au rôle lié au service nommé AWSServiceRoleForAutoScalingles autorisations d'utiliser la clé gérée par le client. Pour en savoir plus sur le rôle lié à un service Amazon EC2 Auto Scaling, consultez Rôles liés à un service pour Amazon EC2 Auto Scaling.

```
{
```

```
  "Sid": "Allow service-linked role use of the customer managed key",
  "Effect": "Allow",
  "Principal": {
    "AWS": "arn:aws:iam::123456789012:role/aws-service-role/autoscaling.amazonaws.com/
AWSServiceRoleForAutoScaling"
  },
  "Action": [
    "kms:GenerateDataKey*",
    "kms:Decrypt"
  ],
  "Resource": "*"
}
```

Les clés de condition `aws:SourceArn` et `aws:SourceAccount` ne sont pas prises en charge dans les stratégies de clé qui autorisent Amazon EC2 Auto Scaling à publier sur des rubriques chiffrées.

AWS services intégrés à Amazon EC2 Auto Scaling

Amazon EC2 Auto Scaling peut être intégré à d'autres AWS services. Consultez les options d'intégration suivantes pour en savoir plus sur la façon dont chaque service fonctionne avec Amazon EC2 Auto Scaling.

Rubriques

Utiliser le rééquilibrage de la capacité pour gérer les interruptions Spot Amazon EC2

Vous pouvez configurer Amazon EC2 Auto Scaling de manière à contrôler et à réagir automatiquement aux modifications qui affectent la disponibilité de vos instances Spot. Le Rééquilibrage de capacité vous permet de maintenir la disponibilité de la charge de travail en augmentant de manière proactive votre flotte avec une nouvelle instance Spot avant qu'une instance en cours d'exécution ne soit interrompue par Amazon EC2.

L'objectif du rééquilibrage de la capacité est de continuer à traiter votre charge de travail sans interruption. Lorsque les instances Spot présentent un risque élevé d'interruption, le service Spot Amazon EC2 signale à Amazon EC2 Auto Scaling avec une recommandation de rééquilibrage d'instance EC2.

Lorsque vous activez le rééquilibrage de la capacité pour votre groupe Auto Scaling, Amazon EC2 Auto Scaling tente de remplacer de manière proactive les instances Spot de votre groupe qui ont reçu une recommandation de rééquilibrage. Cela vous permet de rééquilibrer votre charge de travail en de nouvelles instances Spot qui ne présentent pas un risque élevé d'interruption. Votre charge de travail peut continuer à traiter le travail pendant qu'Amazon EC2 Auto Scaling lance de nouvelles instances Spot et avant que vos instances existantes ne soient interrompues.

Si vous n'utilisez pas le rééquilibrage de la capacité, Amazon EC2 Auto Scaling ne remplace les instances Spot qu'après l'interruption du service Spot d'Amazon EC2 et l'échec de la surveillance d'état correspondante. Avant d'interrompre une instance, Amazon EC2 émet toujours une recommandation de rééquilibrage d'instance EC2 et un préavis d'interruption de deux minutes.

Table des matières

Présentation

Pour utiliser le rééquilibrage de la capacité avec votre groupe Auto Scaling, suivez les étapes de base suivantes :

1. Configurez votre groupe Auto Scaling pour utiliser plusieurs types d'instance et zones de disponibilité. Amazon EC2 Auto Scaling peut ainsi examiner la capacité disponible pour les instances Spot dans chaque zone de disponibilité. Pour plus d'informations, consultez Groupes Auto Scaling combinant plusieurs types d'instances et options d'achat.
2. Ajoutez des hooks de cycle de vie si nécessaire pour arrêter progressivement votre application dans les instances qui reçoivent la notification de rééquilibrage. Pour plus d'informations, consultez Hooks de cycle de vie Amazon EC2 Auto Scaling.

   Les raisons suivantes expliquent pourquoi vous pouvez utiliser un hook de cycle de vie :

   - Pour un arrêt en douceur des employés Amazon SQS

- Pour finaliser le désenregistrement depuis le système de noms de domaine (DNS)
- Pour extraire les journaux système ou les journaux d'application et les charger sur Amazon Simple Storage Service (Amazon S3)

3. Développez une action personnalisée pour le hook de cycle de vie. Pour invoquer votre action personnalisée dès que possible, vous devez savoir quand une instance est prête à être résiliée. Pour cela, détectez l'état du cycle de vie de l'instance.
   - Pour invoquer une action en dehors de l'instance, rédigez une EventBridge règle et automatisez l'action à entreprendre lorsqu'un modèle d'événement correspond à la règle.
   - Pour invoquer une action dans l'instance, configurez l'instance afin d'exécuter un script d'arrêt et récupérer l'état du cycle de vie grâce aux métadonnées de l'instance.

   Il est essentiel de concevoir l'action personnalisée pour terminer en moins de deux minutes. Cela permet de disposer de suffisamment de temps pour terminer les tâches avant la résiliation de l'instance.

Après avoir effectué ces étapes, vous pouvez commencer à utiliser le rééquilibrage de la capacité.

Comportement de rééquilibrage de la capacité

Avec le rééquilibrage de la capacité, Amazon EC2 Auto Scaling se comporte de la manière suivante lorsqu'une instance reçoit une recommandation de rééquilibrage de la capacité :

- Lors du lancement d'une nouvelle instance Spot, Amazon EC2 Auto Scaling attend que celle-ci réussisse sa surveillance de l'état avant de résilier l'ancienne instance. Lors du remplacement de plusieurs instances, la résiliation de chacune des anciennes instances commence une fois que la nouvelle instance a été lancée et qu'elle a réussi sa surveillance d'état.
- Comme Amazon EC2 Auto Scaling tente de lancer de nouvelles instances avant de résilier les anciennes, le fait d'atteindre ou de s'approcher de la capacité maximale spécifiée peut entraver ou stopper les activités de rééquilibrage. Pour contourner ce problème, Amazon EC2 Auto Scaling peut temporairement dépasser la taille maximale du groupe jusqu'à 10 % de la capacité souhaitée.
- Si vous n'avez pas ajouté un hook de cycle de vie à votre groupe Auto Scaling, Amazon EC2 Auto Scaling commence à résilier les anciennes instances dès que les nouvelles réussissent leur surveillance de l'état.
- Si vous avez ajouté un hook de cycle de vie, cela prolonge le temps nécessaire avant que nous ne commencions à résilier les instances précédentes en fonction de la valeur de délai que vous avez indiquée pour le hook de cycle de vie.

- Si vous utilisez des politiques de mise à l’échelle ou une mise à l’échelle planifiée, les activités de mise à l’échelle s’exécutent en parallèle. Si une activité de mise à l'échelle est en cours et que votre groupe Auto Scaling est en dessous de sa nouvelle capacité souhaitée, Amazon EC2 Auto Scaling procède d'abord à la montée en puissance avant de résilier les anciennes instances.

S’il n’y a pas de capacité pour vos types d’instances dans une zone de disponibilité, Amazon EC2 Auto Scaling poursuivra ses tentatives de lancement des instances Spot dans d’autres zones de disponibilité activées jusqu’à ce qu’il y parvienne.

Dans le pire des cas, si le lancement des nouvelles instances échoue, ou si leurs surveillances de l’état échouent, Amazon EC2 Auto Scaling poursuit ses tentatives de lancement. Pendant ce temps, les anciennes finissent par être interrompues et résiliées de force avec un avis d’interruption de deux minutes.

Considérations

Prenez les points suivants en compte lors de l’utilisation du rééquilibrage de capacité :

Concevez votre application de manière à ce qu’elle soit tolérante aux interruptions Spot

Votre application doit être capable de gérer les modifications dynamiques dans le nombre d’instances et la possibilité d’une interruption prématurée d’une instance Spot. Par exemple, si le groupe Auto Scaling se trouve derrière un équilibreur de charge Elastic Load Balancing, Amazon EC2 Auto Scaling attend que l’instance se désenregistre de l’équilibreur de charge avant d’appeler votre hook de cycle de vie. Si le temps nécessaire à l’annulation de l’enregistrement de l’instance et à l’exécution de l’action de cycle de vie complet est trop long, l’instance pourrait être interrompue pendant qu’Amazon EC2 Auto Scaling attend la fin de l’action de cycle de vie pour résilier l’instance.

Amazon EC2 n’est pas toujours capable d’envoyer le signal de recommandation de rééquilibrage avant l’avis d’interruption d’instance Spot de deux minutes. Parfois, le signal de recommandation de rééquilibrage arrive en même temps que l’avis d’interruption de deux minutes. Lorsque cela se produit, Amazon EC2 Auto Scaling appelle le hook de cycle de vie et tente immédiatement de lancer une nouvelle instance Spot.

Éviter un risque élevé d’interruption des instances Spot de remplacement

Vos instances Spot de remplacement peuvent présenter un risque élevé d’interruption si vous utilisez la stratégie d’allocation `lowest-price`. En effet, nous lançons des instances dans le

groupe le moins cher qui dispose de capacités disponibles à ce moment, même si vos instances Spot de remplacement risquent d'être interrompues peu après leur lancement. Pour éviter un risque d'interruption élevé, nous vous recommandons vivement de ne pas utiliser la stratégie d'allocation `lowest-price`. Nous recommandons plutôt la stratégie d'allocation `price-capacity-optimized`. Cette stratégie lance des instances Spot de remplacement dans les groupes Spot les moins susceptibles d'être interrompus et dont le prix est le plus bas possible. Elles sont donc moins susceptibles d'être interrompues dans un futur proche.

Amazon EC2 Auto Scaling ne lancera une nouvelle instance que si la disponibilité est identique ou meilleure

L'un des objectifs du rééquilibrage de capacité est d'améliorer la disponibilité d'une instance Spot. Si une instance Spot existante reçoit une recommandation de rééquilibrage, Amazon EC2 Auto Scaling ne lancera une nouvelle instance que si la nouvelle instance offre une disponibilité supérieure ou égale à celle de l'instance existante. Si le risque d'interruption d'une nouvelle instance est plus important que celui de l'instance existante, Amazon EC2 Auto Scaling ne lancera pas de nouvelle instance. Amazon EC2 Auto Scaling continuera toutefois à évaluer les pools de capacité Spot sur la base des informations fournies par le service Amazon EC2 Spot, et lancera une nouvelle instance si la disponibilité s'améliore.

Il est possible que votre instance existante soit interrompue sans qu'Amazon EC2 Auto Scaling ne lance une nouvelle instance de manière proactive. Lorsque cela se produit, Amazon EC2 Auto Scaling essaye de lancer une nouvelle instance dès qu'il recevra l'avis d'interruption de l'instance Spot. Cela se produit indépendamment du fait que la nouvelle instance présente un risque élevé d'interruption.

Le rééquilibrage de capacité n'augmente pas le taux d'interruption de votre instance Spot

Lorsque vous activez le rééquilibrage de la capacité, cette action n'augmente pas votre Taux d'interruption d'instance Spot.(le nombre d'instances Spot qui sont récupérées lorsqu'Amazon EC2 doit récupérer de l'espace. Toutefois, si le rééquilibrage de la capacité détecte une potentielle interruption d'instance, Amazon EC2 Auto Scaling essayera instantanément de lancer une nouvelle instance. Ainsi, un nombre supérieur d'instances pourraient être remplacées, comparativement au scénario où vous attendez qu'Amazon EC2 Auto Scaling lance une nouvelle instance après l'interruption de l'instance à risque.

Bien que vous puissiez remplacer davantage d'instances lorsque le rééquilibrage de la capacité est activé, vous gagnerez davantage à faire preuve de proactivité que de réactivité. Cela vous donne plus de temps pour agir avant que vos instances ne soient interrompues. En général, après un Avis d'interruption d'instance Spot., vous ne disposez que deux minutes pour arrêter

correctement votre instance. Comme le rééquilibrage de la capacité lance une nouvelle instance à l'avance, vous donnez aux processus existants de plus grandes chances de se terminer sur votre instance à risque. Vous pouvez également démarrer les procédures d'arrêt de votre instance, empêcher la planification de nouveaux travaux sur votre instance à risque et préparer l'instance nouvellement lancée à prendre le contrôle de l'application. Grâce au remplacement proactif dans le rééquilibrage de la capacité, vous bénéficiez d'une continuité.

L'exemple théorique suivant illustre les risques et les avantages liés au rééquilibrage des capacités :

- 14 h 00 – Une recommandation de rééquilibrage est reçue pour l'instance A, et Amazon EC2 Auto Scaling essaye instantanément de lancer une instance B de remplacement, ce qui vous laisse le temps de démarrer vos procédures d'arrêt.
- 14 h 30 – Une recommandation de rééquilibrage est reçue pour l'instance B, remplacée par l'instance C, ce qui vous donne le temps de démarrer vos procédures d'arrêt.
- 14 h 32 – Si le rééquilibrage de la capacité n'était pas activé, et si un avis d'interruption d'instance Spot avait été reçu à 14 h 32 pour l'instance A, vous n'auriez disposé que de deux minutes pour agir. Cependant, l'instance-A aurait été en cours d'exécution jusqu'à ce moment.

Activer le rééquilibrage de la capacité (console)

Vous pouvez activer ou désactiver le Rééquilibrage de la capacité au moment de la création ou de la mise à jour d'un groupe Auto Scaling.

Pour activer le Rééquilibrage de la capacité pour un nouveau groupe Auto Scaling

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Choisissez Créer un groupe Auto Scaling.
3. Pour l'étape 1 : Choisir un modèle de lancement ou une configuration, saisissez un nom pour le groupe Auto Scaling, choisissez un modèle de lancement, puis choisissez Suivant pour passer à l'étape suivante.
4. Pour l'étape 2 : choisissez les options de lancement de l'instance, pour les exigences relatives au type d'instance, choisissez les paramètres pour créer un groupe d'instances mixtes. Cela inclut les types d'instances qu'il peut lancer, les options d'achat d'instances et les stratégies d'allocation pour les instances Spot et à la demande. Par défaut, ces paramètres ne sont pas configurés. Pour les configurer, vous devez sélectionner Override launch template (Remplacer

le modèle de lancement). Pour plus d'informations sur la création de groupes d'instances mixtes, consultez Groupes Auto Scaling combinant plusieurs types d'instances et options d'achat.

5. Sous Réseau, choisissez les options souhaitées. Vérifiez que les sous-réseaux que vous souhaitez utiliser sont dans des zones de disponibilité différentes.
6. Sous la section Stratégies d'allocation, choisissez une stratégie d'allocation des instances Spot. Activer ou désactiver le rééquilibrage de la capacité en cochant ou décochant la case Rééquilibrage de la capacité. Cette option ne s'affiche que si vous demandez à ce qu'un pourcentage du groupe Auto Scaling soit lancé en tant qu'instances Spot dans la section Options d'achat d'Instance.
7. Créez le groupe Auto Scaling.
8. (Facultatif) Ajoutez des hooks de cycle de vie si nécessaire. Pour plus d'informations, consultez Ajouter des hooks de cycle de vie.

Pour activer ou désactiver le Rééquilibrage de la capacité pour un groupe Auto Scaling existant

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling. Un volet fractionné s'ouvre en bas de la page.
3. Dans la page Details (Détails), choisissez Allocation strategies (Stratégies d'allocation), Edit (Modifier).
4. Dans la section Stratégies d'allocation, activez ou désactivez le rééquilibrage de la capacité en cochant ou en décochant la case sous Rééquilibrage de la capacité.
5. Choisissez Mettre à jour.

Activez le rééquilibrage de la capacité (AWS CLI)

Les exemples suivants montrent comment utiliser le AWS CLI pour activer et désactiver le rééquilibrage de capacité.

Utilisez la commande create-auto-scaling-group ou update-auto-scaling-group avec le paramètre suivant :

- `--capacity-rebalance/--no-capacity-rebalance`— Valeur booléenne indiquant si le rééquilibrage des capacités est activé.

Avant d'appeler la commande create-auto-scaling-group, vous devez disposer du nom d'un modèle de lancement configuré pour une utilisation avec un groupe Auto Scaling. Pour plus d'informations, consultez Créer un modèle de lancement pour un groupe Auto Scaling.

> **Note**
>
> Les procédures suivantes expliquent comment utiliser un fichier de configuration au format JSON ou YAML. Si vous utilisez AWS CLI la version 1, vous devez spécifier un fichier de configuration au format JSON. Si vous utilisez AWS CLI la version 2, vous pouvez spécifier un fichier de configuration au format YAML ou JSON.

JSON

Pour créer et configurer un nouveau groupe Auto Scaling

- Utilisez la commande create-auto-scaling-group suivante pour créer un nouveau groupe Auto Scaling et activer le Rééquilibrage de la capacité. Cette commande fait référence à un fichier JSON comme seul paramètre de votre groupe Auto Scaling, au lieu d'un fichier JSON.

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

  Si vous n'avez pas encore de fichier de configuration CLI spécifiant une politique d'instances mixtes, créez-en un.

  Ajoutez la ligne suivante à l'objet JSON de niveau supérieur dans le fichier de configuration.

```
{
    "CapacityRebalance": true
}
```

  Voici un exemple de fichier `config.json`.

```
{
    "AutoScalingGroupName": "my-asg",
    "DesiredCapacity": 12,
    "MinSize": 12,
    "MaxSize": 15,
    "CapacityRebalance": true,
    "MixedInstancesPolicy": {
```

```
        "InstancesDistribution": {
            "OnDemandBaseCapacity": 0,
            "OnDemandPercentageAboveBaseCapacity": 25,
            "SpotAllocationStrategy": "price-capacity-optimized"
        },
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Default"
            },
            "Overrides": [
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                },
                {
                    "InstanceType": "m5.large"
                },
                {
                    "InstanceType": "m5a.large"
                },
                {
                    "InstanceType": "c4.large"
                },
                {
                    "InstanceType": "m4.large"
                },
                {
                    "InstanceType": "c3.large"
                },
                {
                    "InstanceType": "m3.large"
                }
            ]
        }
    },
    "TargetGroupARNs": "arn:aws:elasticloadbalancing:us-
west-2:123456789012:targetgroup/my-alb-target-group/943f017f100becff",
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

YAML

Pour créer et configurer un nouveau groupe Auto Scaling

- Utilisez la commande create-auto-scaling-group suivante pour créer un nouveau groupe Auto Scaling et activer le Rééquilibrage de la capacité. Cette commande fait référence à un fichier YAML comme seul paramètre de votre groupe Auto Scaling, au lieu d'un fichier JSON.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

Ajoutez la ligne suivante à votre fichier de configuration au format YAML.

```
CapacityRebalance: true
```

Voici un exemple de fichier `config.yaml`.

```
---
AutoScalingGroupName: my-asg
DesiredCapacity: 12
MinSize: 12
MaxSize: 15
CapacityRebalance: true
MixedInstancesPolicy:
  InstancesDistribution:
    OnDemandBaseCapacity: 0
    OnDemandPercentageAboveBaseCapacity: 25
    SpotAllocationStrategy: price-capacity-optimized
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
TargetGroupARNs:
```

```
- arn:aws:elasticloadbalancing:us-west-2:123456789012:targetgroup/my-alb-target-
group/943f017f100becff
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Pour activer le Rééquilibrage de la capacité pour un groupe Auto Scaling existant

- Utilisez la commande update-auto-scaling-group suivante pour activer le Rééquilibrage de la capacité.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --capacity-rebalance
```

Pour vérifier que le Rééquilibrage de la capacité est activé pour un groupe Auto Scaling

- Utilisez la commande describe-auto-scaling-groups suivante pour vérifier que le Rééquilibrage de la capacité est activé et pour afficher les détails.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Voici un exemple de réponse.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            ...
            "CapacityRebalance": true
        }
    ]
}
```

Pour désactiver le Rééquilibrage de la capacité.

Utilisez la commande update-auto-scaling-group avec l'option `--no-capacity-rebalance` pour désactiver le Rééquilibrage de la capacité.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
```

```
  --no-capacity-rebalance
```

Ressources connexes

Pour plus d'informations sur le rééquilibrage des capacités, consultez Gérer de manière proactive le cycle de vie des instances Spot à l'aide de la nouvelle fonctionnalité de rééquilibrage des capacités pour EC2 Auto Scaling sur le Compute Blog. AWS

Pour plus d'informations sur les recommandations de rééquilibrage des instances EC2, consultez les recommandations de rééquilibrage des instances EC2 dans le guide de l'utilisateur Amazon EC2.

Pour en savoir plus sur les hooks de cycle de vie, consultez les ressources suivantes.

- Didacticiel : configurer un hook de cycle de vie qui appelle une fonction Lambda(en utilisant EventBridge)
- Tutoriel : configurer les données utilisateur pour récupérer l'état du cycle de vie cible via des métadonnées d'instance

Limites

- Amazon EC2 Auto Scaling ne peut remplacer l’instance qui reçoit la notification de rééquilibrage uniquement que si elle n’est pas protégée contre la mise à l’échelle horizontale. Cependant, la protection de mise à l’échelle horizontale n’empêche pas la résiliation suite à une interruption ponctuelle. Pour plus d’informations, consultez Utiliser la protection de la taille d'instance.
- La prise en charge du rééquilibrage des capacités est disponible dans tous les Régions AWS commerciaux où Amazon EC2 Auto Scaling est disponible, à l’exception de la région du Moyen-Orient (EAU).

Utilisez les réserves de capacité à la demande pour réserver de la capacité dans des zones de disponibilité spécifiques

La réserve de capacité à la demande d’Amazon EC2 vous permet de réserver de la capacité de calcul dans une zone de disponibilité spécifique. Pour commencer à utiliser des réserves de capacité, vous devez créer une réserve de capacité dans une zone de disponibilité spécifique. Vous pouvez ensuite lancer des instances dans la capacité réservée, afficher son utilisation de capacité en temps réel, et augmenter ou diminuer ses capacités en fonction de vos besoins.

Les réserves de capacité sont configurées comme `open` ou `targeted`. Si la réserve de capacité est `open`, toutes les nouvelles instances et les instances existantes dont les attributs correspondent s'exécutent automatiquement dans la capacité de la Réservation de capacité. Si la réserve de capacité est `targeted`, les instances doivent la cibler spécifiquement pour s'exécuter dans la capacité réservée.

Cette rubrique indique comment créer un groupe Auto Scaling qui lance des instances à la demande dans les réserves de capacité `targeted`. Cela vous permet de mieux contrôler le moment où vous devez utiliser des réserves de capacité spécifiques.

Les étapes de base sont les suivantes :

1. Créez des réserves de capacité dans plusieurs zones de disponibilité ayant le même type d'instance, la même plateforme et le même nombre d'instances.
2. Réservations de capacité de groupe à l'aide de AWS Resource Groups.
3. Créez un groupe Auto Scaling avec un modèle de lancement qui cible le groupe de ressources, en utilisant les mêmes zones de disponibilité que les réserves de capacité.

Table des matières

Étape 1 : créer des réserves de capacité

La première étape consiste à créer une réserve de capacité dans chaque zone de disponibilité dans laquelle votre groupe Auto Scaling sera déployé.

> **Note**
>
> Vous ne pouvez créer des réserves `targeted` que lorsque vous créez les réserves de capacité pour la première fois.

Console

Pour créer vos réserves de capacité

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Choisissez Réserves de capacité, puis Créer Réserve de capacité.
3. Sur la page Créer une réserve de capacité, faites attention aux paramètres suivants dans la section Détails de l'instance. Le type d'instance, la plateforme et la zone de disponibilité des instances que vous lancez doivent correspondre au type d'instance, à la plateforme et à la zone de disponibilité que vous spécifiez ici ou la réserve de capacité ne s'applique pas.

   a. Pour le Type d'instance, choisissez le type d'instance à lancer dans la capacité réservée.
   b. Pour la Plateforme, choisissez le système d'exploitation pour vos instances.
   c. Pour la zone de disponibilité, choisissez la première zone de disponibilité dans laquelle vous souhaitez réserver de la capacité.
   d. Pour la capacité totale, choisissez le nombre d'instances dont vous avez besoin. Calculez le nombre total d'instances dont vous avez besoin pour votre groupe Auto Scaling, divisé par le nombre de zones de disponibilité que vous prévoyez d'utiliser.
4. Dans Détails de réserve de capacité, pour l'élément La réservation de capacité se termine, choisissez l'une des options suivantes :

   - À une heure précise — Annulez automatiquement la réservation de capacité à la date et à l'heure spécifiées.
   - Manuellement : réservez la capacité jusqu'à ce que vous l'annuliez explicitement.
5. Pour l'éligibilité des instances, choisissez Ciblée : uniquement les instances qui ciblent la réserve de capacité.
6. (Facultatif) Pour les balises, indiquez les balises à associer à la réserve de capacité.
7. Choisissez Créer.
8. Notez l'identifiant de la réserve de capacité qui vient d'être créée. Vous en aurez besoin pour configurer le groupe de réserves de capacité.

Répétez cette procédure pour chaque zone de disponibilité que vous souhaitez activer pour votre groupe Auto Scaling, en modifiant uniquement la valeur de l'option Zone de disponibilité.

AWS CLI

Pour créer vos réserves de capacité

Utilisez la commande create-capacity-reservation suivante pour créer les réserves de capacité. Remplacez les valeurs d'exemple de `--availability-zone`, `--instance-type`, `--instance-platform`, et `--instance-count`.

```
aws ec2 create-capacity-reservation \
    --availability-zone us-east-1a \
    --instance-type c5.xlarge \
    --instance-platform Linux/UNIX \
    --instance-count 3 \
    --instance-match-criteria targeted
```

Exemple d'ID de réserve de capacité en résultant

```
{
    "CapacityReservation": {
        "CapacityReservationId": "cr-1234567890abcdef1",
        "OwnerId": "123456789012",
        "CapacityReservationArn": "arn:aws:ec2:us-east-1:123456789012:capacity-
reservation/cr-1234567890abcdef1",
        "InstanceType": "c5.xlarge",
        "InstancePlatform": "Linux/UNIX",
        "AvailabilityZone": "us-east-1a",
        "Tenancy": "default",
        "TotalInstanceCount": 3,
        "AvailableInstanceCount": 3,
        "EbsOptimized": false,
        "EphemeralStorage": false,
        "State": "active",
        "StartDate": "2023-07-26T21:36:14+00:00",
        "EndDateType": "unlimited",
        "InstanceMatchCriteria": "targeted",
        "CreateDate": "2023-07-26T21:36:14+00:00"
    }
}
```

Notez l'identifiant de la réserve de capacité qui vient d'être créée. Vous en aurez besoin pour configurer le groupe de réserves de capacité.

Répétez cette commande pour chaque zone de disponibilité que vous souhaitez activer pour votre groupe Auto Scaling, en modifiant uniquement la valeur de l'option `--availability-zone`.

Étape 2 : créer un groupe de réserve de capacité

Lorsque vous avez fini de créer les réservations de capacité, vous pouvez les regrouper à l'aide du service AWS Resource Groups. AWS Resource Groups prend en charge plusieurs types de groupes pour différents usages. Amazon EC2 utilise un groupe spécialisé, connu sous le nom de groupe de ressources lié à un service, pour cibler un groupe de réserves de capacité. Pour interagir avec ce groupe de ressources lié à un service, vous pouvez utiliser le AWS CLI ou un kit SDK, mais pas la console. Pour plus d'informations sur les groupes de ressources liés à un service, consultez la section Configurations de service pour les groupes de ressources dans le Guide de l'utilisateur sur les groupes de ressources AWS .

Pour créer un groupe de réservation de capacité à l'aide du AWS CLI

Utilisez la commande create-group pour créer un groupe de ressources qui ne peut contenir que des réserves de capacité. Dans cet exemple, le groupe de ressources est nommé *my-cr-group*.

```
aws resource-groups create-group \
    --name my-cr-group \
    --configuration '{"Type":"AWS::EC2::CapacityReservationPool"}'
 '{"Type":"AWS::ResourceGroups::Generic", "Parameters": [{"Name": "allowed-resource-
types", "Values": ["AWS::EC2::CapacityReservation"]}]}'
```

Voici un exemple de réponse.

```
{
    "Group": {
        "GroupArn": "arn:aws:resource-groups:us-east-1:123456789012:group/my-cr-group",
        "Name": "my-cr-group"
    },
    "GroupConfiguration": {
        "Configuration": [
            {
                "Type": "AWS::EC2::CapacityReservationPool"
            },
            {
                "Type": "AWS::ResourceGroups::Generic",
                "Parameters": [
                    {
                        "Name": "allowed-resource-types",
                        "Values": [
                            "AWS::EC2::CapacityReservation"
```

```
                        ]
                    }
                ]
            }
        ],
        "Status": "UPDATE_COMPLETE"
    }
}
```

Notez L'ARN du groupe de ressources. Vous en aurez besoin pour configurer le modèle de lancement de votre groupe Auto Scaling.

Pour associer vos réserves de capacité au nouveau groupe à l'aide du AWS CLI

Utilisez la commande group-resources suivante pour associer les réserves de capacité au groupe de réserves de capacité nouvellement créé. Pour l'option `--resource-arns`, indiquez les réserves de capacité à l'aide de leurs ARN. Construisez les ARN à l'aide de la région appropriée, de votre identifiant de compte et des identifiants de réserve que vous avez notés précédemment. Dans cet exemple, les réserves avec les identifiants *cr-1234567890abcdef1* et *cr-54321abcdef567890* sont regroupées dans le groupe intitulé *my-cr-group*.

```
aws resource-groups group-resources \
    --group my-cr-group \
    --resource-arns \
      arn:aws:ec2:region:account-id:capacity-reservation/cr-1234567890abcdef1 \
      arn:aws:ec2:region:account-id:capacity-reservation/cr-54321abcdef567890
```

Voici un exemple de réponse.

```
{
   "Succeeded": [
   "arn:aws:ec2:us-east-1:123456789012:capacity-reservation/cr-1234567890abcdef1",
   "arn:aws:ec2:us-east-1:123456789012:capacity-reservation/cr-54321abcdef567890"
   ],
   "Failed": [],
   "Pending": []
}
```

Pour plus d'informations sur la modification ou la suppression du groupe de ressources, consultez le Guide de référence d'API de groupes de ressources AWS.

Étape 3 : créer un modèle de lancement

Console

Pour créer un modèle de lancement

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/.
2. Dans le volet de navigation, sous Instances, choisissez Modèles de lancement.
3. Choisissez Create launch template (Créer un modèle de lancement). Saisissez un nom et une description pour la version initiale du modèle de lancement.
4. Sous Auto Scaling guidance (Guide Auto Scaling), activez la case à cocher.
5. Créez le modèle de lancement. Choisissez une AMI et un type d'instance qui correspond aux réserves de capacité que vous prévoyez d'utiliser, et éventuellement, une paire de clés, un ou plusieurs groupes de sécurité, et des volumes EBS ou des volumes de stockage d'instances supplémentaires pour vos instances.
6. Développez la section Détails avancés et procédez comme suit :
   a. Pour la réserve de capacité, choisissez Cibler par groupe.
   b. Pour la Réserve de capacité - Cibler par groupe, choisissez le groupe de réserves de capacité que vous avez créé dans la section précédente, puis cliquez sur Enregistrer.
7. Choisissez Create launch template (Créer un modèle de lancement).
8. Sur la page de confirmation, choisissez Create Auto Scaling group (Créer un groupe Auto Scaling).

AWS CLI

Pour créer un modèle de lancement

Utilisez la commande create-launch-template suivante pour créer un modèle de lancement qui indique que la réserve de capacité cible un groupe de ressources particulier. Remplacez la valeur d'exemple de `--launch-template-name`. Remplacez *`c5.xlarge`* par le type d'instance que vous avez utilisé dans la réserve de capacité et *`ami-0123456789EXAMPLE`* par l'identifiant de l'AMI que vous souhaitez utiliser. Remplacez *`arn:aws:resource-groups:region:account-id:group/my-cr-group`* par l'ARN du groupe de ressources que vous avez créé au début de la section précédente.

```
aws ec2 create-launch-template \
```

```
    --launch-template-name my-launch-template \
    --launch-template-data \
        '{"InstanceType": "c5.xlarge",
          "ImageId": "ami-0123456789EXAMPLE",
          "CapacityReservationSpecification":
            {"CapacityReservationTarget":
                { "CapacityReservationResourceGroupArn": "arn:aws:resource-
groups:region:account-id:group/my-cr-group" }
            }
        }'
```

Voici un exemple de réponse.

```
{
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-0dd77bd41dEXAMPLE",
        "LaunchTemplateName": "my-launch-template",
        "CreateTime": "2023-07-26T21:42:48+00:00",
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
        "DefaultVersionNumber": 1,
        "LatestVersionNumber": 1
    }
}
```

Étape 4 : créer un groupe Auto Scaling

Console

Créez votre groupe Auto Scaling comme d'habitude, mais lorsque vous choisissez vos sous-réseaux VPC, sélectionnez un sous-réseau dans chaque zone de disponibilité qui correspond aux réserves de capacité `targeted` que vous avez créées. Ensuite, lorsque votre groupe Auto Scaling lance une instance à la demande dans l'une de ces zones de disponibilité, l'instance est exécutée dans la capacité réservée pour cette zone de disponibilité. Si le groupe de ressources n'a plus de réserves de capacité avant que la capacité souhaitée ne soit atteinte, nous lançons tout ce qui dépasse la capacité réservée en tant que capacité à la demande normale.

Pour créer un simple groupe Auto Scaling

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.

2. Dans la barre de navigation en haut de l'écran, choisissez le même Région AWS que celui que vous avez utilisé lors de la création du modèle de lancement.
3. Choisissez Créer un groupe Auto Scaling.
4. Dans la page Choisir un modèle de lancement ou une configuration, dans Nom du groupe Auto Scaling, entrez un nom pour le groupe Auto Scaling.
5. Dans Launch template (Modèle de lancement), choisissez un modèle de lancement existant.
6. Pour Version du modèle de lancement, indiquez si le groupe Auto Scaling utilise la version par défaut, la version la plus récente ou une version spécifique du modèle de lancement lors de l'évolutivité horizontale.
7. Sur la page Choisir les options de lancement, ignorez la section Exigences relatives au type d'instance pour utiliser le type d'instance EC2 indiqué dans le modèle de lancement.
8. Sous Network (Réseau), pour VPC, choisissez un VPC. Le groupe Auto Scaling doit être créé dans le même VPC que le groupe de sécurité que vous avez spécifié dans votre modèle de lancement. Si vous n'avez pas indiqué de groupe de sécurité dans votre modèle de lancement, vous pouvez choisir un VPC qui dispose d'un sous-réseau dans les mêmes zones de disponibilité que vos réserves de capacité.
9. Pour les des sous-réseaux et les sous-réseaux, choisissez les sous-réseaux de chaque zone de disponibilité que vous souhaitez inclure, en fonction des zones de disponibilité dans lesquelles se trouvent vos réserves de capacité.
10. Choisissez Next (Suivant) deux fois.
11. Sur la page Configurer la taille du groupe et les politiques de mise à l'échelle, pour la capacité souhaitée, saisissez le nombre initial d'instances à lancer. Lorsque vous modifiez ce nombre pour une valeur en dehors des limites de capacité minimale ou maximale, vous devez mettre à jour les valeurs de Minimum capacity (Capacité minimale) ou Maximum capacity (Capacité maximale). Pour plus d'informations, consultez Définissez des limites de mise à l'échelle pour votre groupe Auto Scaling.
12. Choisissez Skip to review (Passer à la révision).
13. Sur la page Vérifier, sélectionnez Créer un groupe Auto Scaling.

AWS CLI

Pour créer un simple groupe Auto Scaling

Utilisez la commande create-auto-scaling-group suivante et indiquez le nom et la version de votre modèle de lancement comme valeur de l'option `--launch-template`. Remplacez les valeurs

d'exemple de `--auto-scaling-group-name`, `--min-size`, `--max-size`, et `--vpc-zone-identifier`.

Pour l'option `--availability-zones`, indiquez les zones de disponibilité pour lesquelles vous avez créé des réserves de capacité. Par exemple, si vos réserves de capacité indiquent les zones de disponibilité `us-east-1a` et `us-east-1b`, vous devez créer votre groupe Auto Scaling dans les mêmes zones. Ensuite, lorsque votre groupe Auto Scaling lance une instance à la demande dans l'une de ces zones de disponibilité, l'instance est exécutée dans la capacité réservée pour cette zone de disponibilité. Si le groupe de ressources n'a plus de réserves de capacité avant que la capacité souhaitée ne soit atteinte, nous lançons tout ce qui dépasse la capacité réservée en tant que capacité à la demande normale.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateName=my-launch-template,Version='1' \
    --min-size 6 \
    --max-size 6 \
    --vpc-zone-identifier "subnet-5f46ec3b,subnet-0ecac448" \
    --availability-zones us-east-1a us-east-1b
```

Ressources connexes

Pour un exemple d'implémentation, consultez le AWS CloudFormation modèle dans le GitHub référentiel AWS d'exemples suivant : https://github.com/aws-samples/aws-auto-scaling-backed-by-on-demand-capacity-reservations/.

Les rubriques connexes suivantes peuvent vous être utiles pour en savoir plus sur les réserves de capacité.

- On-Demand Capacity Reservations
  - Créez une réservation de capacité dans le guide de l'utilisateur Amazon EC2
  - Réservations de capacité à la demande dans le guide de l'utilisateur Amazon EC2
  - Ciblez un groupe de réservations de capacité à la demande sur Amazon EC2 sur le blog AWS Cloud Operations & Migrations
- Blocs de capacité (réserves de capacité d'une durée définie)
  - Blocs de capacité pour le ML dans le guide de l'utilisateur Amazon EC2
  - Utilisation Capacity Blocks pour les charges de travail liées à l'apprentissage automatique

Créez des groupes Auto Scaling depuis la ligne de commande en utilisant AWS CloudShell

Si cela est pris en charge Régions AWS, vous pouvez exécuter AWS CLI des commandes à l'aide AWS CloudShell d'un shell pré-authentifié basé sur un navigateur qui se lance directement depuis le. AWS Management Console Vous pouvez exécuter AWS CLI des commandes sur des services à l'aide de votre shell préféré (shell Bash ou Z). PowerShell

Vous pouvez lancer AWS CloudShell depuis le AWS Management Console en utilisant l'une des deux méthodes suivantes :

- Choisissez l' AWS CloudShell icône dans la barre de navigation de la console. Il est situé à droite de la zone de recherche.
- Utilisez le champ de recherche de la barre de navigation de la console pour rechercher, CloudShellpuis choisissez l'CloudShelloption.

Lors du premier AWS CloudShell lancement dans une nouvelle fenêtre de navigateur, un panneau de bienvenue s'affiche et répertorie les principales fonctionnalités. Après avoir fermé ce panneau, des mises à jour de l'état sont fournies pendant que le shell configure et transmet les informations d'identification de votre console. Lorsque l'invite de commandes s'affiche, le shell est prêt pour l'interaction.

Pour en savoir plus sur ce service, consultez le guide de l'utilisateur AWS CloudShell.

Créer un groupe Auto Scaling avec AWS CloudFormation

Amazon EC2 Auto Scaling est intégré à AWS CloudFormation un service qui vous aide à modéliser et à configurer vos AWS ressources afin que vous puissiez passer moins de temps à créer et à gérer vos ressources et votre infrastructure. Vous créez un modèle qui décrit toutes les AWS ressources que vous souhaitez (telles que les groupes Auto Scaling), puis vous AWS CloudFormation approvisionnez et configurez ces ressources pour vous.

Lorsque vous l'utilisez AWS CloudFormation, vous pouvez réutiliser votre modèle pour configurer vos ressources Amazon EC2 Auto Scaling de manière cohérente et répétée. Décrivez vos ressources une seule fois, puis fournissez les mêmes ressources encore et encore dans plusieurs Comptes AWS régions.

Amazon EC2 Auto Scaling et modèles AWS CloudFormation

Pour allouer et configurer les ressources pour Amazon EC2 Auto Scaling et les services associés, vous devez maîtriser les modèles AWS CloudFormation. Les modèles sont des fichiers texte formatés en JSON ou YAML. Ces modèles décrivent les ressources que vous souhaitez mettre à disposition dans vos AWS CloudFormation piles. Si vous n'êtes pas familiarisé avec JSON ou YAML, vous pouvez utiliser AWS CloudFormation Designer pour vous aider à démarrer avec les AWS CloudFormation modèles. Pour plus d'informations, voir Qu'est-ce que AWS CloudFormation Designer ? dans le guide de AWS CloudFormation l'utilisateur.

Pour commencer à créer vos propres modèles de pile pour Amazon EC2 Auto Scaling, exécutez les tâches suivantes :

- Créez un modèle de lancement à l'aide de AWS::EC2::LaunchTemplate.
- Créez un groupe Auto Scaling à l'aide de AWS::AutoScaling::AutoScalingAWS::AutoScaling::AutoScalingGroup .

Pour une démonstration vous expliquant comment déployer un groupe Auto Scaling derrière un Application Load Balancer, consultez Procédure de création d'un serveur web scalable à équilibrage de charge dans le Guide de l'utilisateur AWS CloudFormation .

Vous trouverez d'autres exemples utiles d'extraits de modèles permettant de créer des groupes Auto Scaling et des ressources associées dans les sections suivantes du guide de l'AWS CloudFormation utilisateur :

- Référence du type de ressource Amazon EC2 Auto Scaling Référence du type de ressource
- Configurez les ressources Amazon EC2 Auto Scaling avec AWS CloudFormation

En savoir plus sur AWS CloudFormation

Pour en savoir plus AWS CloudFormation, consultez les ressources suivantes :

- AWS CloudFormation
- AWS CloudFormation Guide de l'utilisateur
- AWS CloudFormation API Reference
- AWS CloudFormation Guide de l'utilisateur de l'interface de ligne de commande

AWS Compute Optimizer À utiliser pour obtenir des recommandations pour le type d'instance d'un groupe Auto Scaling

AWS fournit des recommandations relatives aux instances Amazon EC2 pour vous aider à améliorer les performances, à économiser de l'argent, ou les deux, en utilisant des fonctionnalités optimisées par. AWS Compute Optimizer Vous pouvez utiliser ces recommandations pour décider de passer à un nouveau type d'instance.

Pour formuler des recommandations, Compute Optimizer analyse vos spécifications d'instances existantes ainsi que l'historique récent des métriques. Les données compilées sont ensuite utilisées pour recommander les types d'instances Amazon EC2 qui sont les plus à même de gérer la charge de travail existante. Les recommandations sont renvoyées avec la tarification horaire des instances.

> **Note**
>
> Pour obtenir des recommandations de Compute Optimizer, vous devez d'abord vous inscrire à Compute Optimizer. Pour plus d'informations, consultez Démarrer avec AWS Compute Optimizer dans le Guide de l'utilisateur AWS Compute Optimizer .

Table des matières

Limites

Compute Optimizer génère des recommandations pour les instances des groupes Auto Scaling configurés pour lancer et exécuter les types d'instances M, C, R, T et X. Cependant, il ne génère pas de recommandations pour les types d'instances -g alimentés par les processeurs AWS Graviton2 (par exemple, C6g), ni pour les types d'instances -n qui offrent des performances de bande passante réseau supérieures (par exemple, M5n).

Les groupes Auto Scaling doivent également être configurés pour exécuter un type d'instance unique (c'est-à-dire aucun type d'instance mixte), ne doivent pas avoir de politique de mise à

l'échelle associée et avoir les mêmes valeurs pour la capacité désirée, minimale et maximale (c'est-à-dire un groupe Auto Scaling avec un nombre fixe d'instances). Compute Optimizer génère des recommandations pour les instances de groupes Auto Scaling qui répondent à toutes ces exigences de configuration.

Conclusions

Compute Optimizer classe ses résultats pour les groupes Auto Scaling comme suit :

- Non optimisé – Un groupe Auto Scaling est considéré comme non optimisé lorsque Compute Optimizer a identifié une recommandation pouvant fournir de meilleures performances pour votre application.
- Optimisé – Un groupe Auto Scaling est considéré comme optimisé lorsque Compute Optimizer détermine que ce groupe est correctement provisionné pour exécuter votre application, en fonction du type d'instance choisi. Pour les ressources optimisées, Compute Optimizer peut parfois recommander un type d'instance de nouvelle génération.
- Aucune – Il n'y a aucune recommandation pour ce groupe Auto Scaling. Cela peut se produire si vous êtes inscrit à Computer Optimizer depuis moins de 12 heures, ou lorsque le groupe Auto Scaling s'exécute depuis moins de 30 heures, ou lorsque le type d'instance ou groupe Auto Scaling n'est pas pris en charge par Compute Optimizer. Pour plus d’informations, consultez la section Limites.

Afficher les recommandations

Après votre inscription à Compute Optimizer, vous pouvez afficher les résultats et les recommandations qu'il génère pour vos groupes Auto Scaling. Si vous venez de vous inscrire, les recommandations peuvent ne pas être disponibles avant un délai de 12 heures.

Pour afficher les recommandations générées pour un groupe Auto Scaling

1. Ouvrez la console Compute Optimizer à l’adresse https://console.aws.amazon.com/compute-optimizer/.

   La page Tableau de bord s'ouvre.
2. Choisissez View recommendations for all Auto Scaling groups (Afficher les recommandations pour tous les groupes Auto Scaling).
3. Sélectionnez votre groupe Auto Scaling.

4. Choisissez View detail (Afficher les détails).

   La vue change pour afficher jusqu'à trois recommandations d'instances différentes dans une vue préconfigurée, en fonction des paramètres de table par défaut. Il fournit également des données CloudWatch métriques récentes (utilisation moyenne du processeur, entrée réseau moyenne et sortie réseau moyenne) pour le groupe Auto Scaling.

Déterminez si vous souhaitez utiliser l'une des recommandations. Décidez s'il convient d'optimiser les performances et/ou de réduire les coûts.

Pour modifier le type d'instance dans votre groupe Auto Scaling, mettez à jour le modèle de lancement ou mettez à jour le groupe Auto Scaling pour utiliser une nouvelle configuration du lancement. Les instances existantes continuent d'utiliser la configuration précédente. Pour mettre à jour les instances existantes, résiliez-les afin qu'elles soient remplacées par votre groupe Auto Scaling ou permettez à la scalabilité automatique de remplacer progressivement les anciennes instances par des instances plus récentes en fonction de vos politiques de résiliation.

> **Note**
>
> Grâce aux fonctionnalités de durée de vie maximale et d'actualisation de l'instance, vous pouvez également remplacer les instances existantes du groupe Auto Scaling pour lancer de nouvelles instances qui utilisent le nouveau modèle de lancement ou la nouvelle configuration du lancement. Pour plus d'informations, consultez Remplacer des instances Auto Scaling en fonction de la durée de vie maximale de l'instance et Utiliser une actualisation d'instance pour mettre à jour les instances d'un groupe Auto Scaling.

Considérations relatives à l'évaluation des recommandations

Avant de passer à un nouveau type d'instance, tenez compte des éléments suivants :

- Les recommandations ne prédisent pas votre utilisation. Les recommandations sont basées sur votre historique d'utilisation au cours de la période de 14 jours la plus récente. Veillez à choisir un type d'instance censé répondre à vos futurs besoins en termes d'utilisation.
- Concentrez-vous sur le graphique des métriques pour déterminer si l'utilisation réelle est inférieure à la capacité d'instance. Vous pouvez également consulter les données métriques (moyenne, pic, percentile) afin d'évaluer plus en détail les recommandations relatives CloudWatch à votre instance EC2. Par exemple, notez l'évolution des métriques de pourcentage d'UC pendant la journée et s'il y

a des pics qui doivent être pris en compte. Pour plus d'informations, consultez la section Affichage des statistiques disponibles dans le guide de CloudWatch l'utilisateur Amazon.

- Compute Optimizer peut fournir des recommandations pour les instances de performance à capacité extensible, à savoir les instances T3, T3a et T2. Si vous dépassez régulièrement votre niveau de base, assurez-vous que vous pouvez continuer à le faire en fonction des vCPU du nouveau type d'instance. Pour plus d'informations, consultez la section Crédits du processeur et performances de base pour les instances de performance en rafale dans le guide de l'utilisateur Amazon EC2.
- Si vous avez acheté une Instance réservée, votre instance à la demande peut être facturée au prix d'une Instance réservée. Avant de modifier votre type d'instance actuel, commencez par évaluer l'impact sur l'utilisation et la couverture de l'Instance réservée.
- Dans la mesure du possible, envisagez des conversions vers des instances de nouvelle génération.
- Lors de la migration vers une autre famille d'instances, assurez-vous que le type d'instance actuel et le nouveau type d'instance sont compatibles, en termes de virtualisation, d'architecture ou de type de réseau par exemple. Pour plus d'informations, consultez la section Compatibilité pour le redimensionnement des instances dans le guide de l'utilisateur Amazon EC2.
- Enfin, tenez compte de la note de risque de performances fournie pour chaque recommandation. Le risque de performances correspond à l'effort que vous pourriez avoir à consacrer pour valider si le type d'instance recommandé répond aux exigences de performances de votre charge de travail. Nous recommandons également des tests rigoureux de charge et de performance avant et après toute modification.

Ressources supplémentaires

Outre les rubriques de cette page, consultez les ressources suivantes :

- Types d'instances Amazon EC2
- AWS Compute Optimizer Guide de l'utilisateur

Utiliser Elastic Load Balancing pour répartir le trafic sur les instances dans votre groupe Auto Scaling.

Elastic Load Balancing répartit automatiquement le trafic d'application entrant sur toutes les instances EC2 que vous exécutez. Elastic Load Balancing contribue à gérer les demandes entrantes en acheminant le trafic de manière optimale afin qu'aucune instance ne soit submergée.

Pour utiliser Elastic Load Balancing avec votre groupe Auto Scaling, attachez l'équilibreur de charge à votre groupe Auto Scaling. Cela permet d'enregistrer le groupe auprès de l'équilibreur de charge et de bénéficier d'un point de contact unique pour tout le trafic web entrant dans votre groupe Auto Scaling.

Lorsque vous utilisez Elastic Load Balancing avec votre groupe Auto Scaling, il n'est pas nécessaire d'enregistrer les instances EC2 individuelles auprès de l'équilibreur de charge. Les instances qui sont lancées par votre groupe Auto Scaling sont automatiquement enregistrées auprès de l'équilibreur de charge. De même, l'enregistrement des instances qui sont résiliées par votre groupe Auto Scaling est automatiquement annulé auprès de l'équilibreur de charge.

Après avoir attaché un équilibreur de charge à votre groupe Auto Scaling, vous pouvez configurer ce dernier afin qu'il utilise des métriques Elastic Load Balancing (comme le nombre de demandes Application Load Balancer par cible) pour mettre à l'échelle le nombre d'instances du groupe en fonction de l'évolution de la demande.

Si vous le souhaitez, vous pouvez ajouter des surveillances d'état Elastic Load Balancing à votre groupe Auto Scaling afin qu'Amazon EC2 Auto Scaling puisse identifier et remplacer les instances défectueuses en fonction de ces surveillances d'état supplémentaires. Sinon, vous pouvez créer une CloudWatch alarme qui vous avertira si le nombre d'hôtes sains du groupe cible est inférieur au nombre autorisé.

Table des matières

- Vérifier l’état d’attachement de votre équilibreur de charge
- Ajouter et supprimer des zones de disponibilité
- Exemples d'utilisation d'Elastic Load Balancing avec le AWS Command Line Interface

Types d'équilibreurs de charge Elastic Load Balancing

Elastic Load Balancing fournit quatre types d'équilibreurs de charge qui peuvent être utilisés avec votre groupe Auto Scaling : Application Load Balancer, Network Load Balancer, Gateway Load Balancer et Classic Load Balancers.

Il existe une différence clé dans la configuration des équilibreurs de charge. Avec les types Application Load Balancer, Network Load Balancer et Gateway Load Balancer, les instances sont enregistrées en tant que cibles auprès d'un groupe cible et vous acheminez le trafic vers le groupe cible. Avec le type Classic Load Balancer, les instances sont enregistrées directement auprès de l'équilibreur de charge.

Application Load Balancer

Achemine et équilibre les charges au niveau de la couche d'application (HTTP/HTTPS) et prend en charge le routage basé sur le chemin d'accès. Un équilibreur de charge Application Load Balancer peut acheminer les demandes vers les ports d'une ou de plusieurs cibles enregistrées, telles que des instances EC2, au sein de votre cloud privé virtuel (VPC).

Network Load Balancer

Achemine et équilibre les charges au niveau de la couche de transport (TCP/UDP couche 4) en fonction des informations d'adresse extraites de l'en-tête de paquet TCP. Un équilibreur de charge Network Load Balancer peut traiter les pics de trafic, conserver l'adresse IP source du client et utiliser une adresse IP fixe pendant la durée de vie de l'équilibreur de charge.

Gateway Load Balancer

Distribue le trafic à une flotte d'instances d'appliances. Offre évolutivité, disponibilité et simplicité pour les appliances virtuelles tierces, telles que les pare-feu, les systèmes de détection et de prévention des intrusions et d'autres appliances. Les équilibreurs de charge Gateway Load Balancer fonctionnent avec des appliances virtuelles qui prennent en charge le protocole GENEVE. Une intégration technique supplémentaire est nécessaire. Veuillez consulter le guide de l'utilisateur avant de choisir un équilibreur de charge Gateway Load Balancer.

Classic Load Balancer

Achemine et équilibre les charges au niveau de la couche de transport (TCP/SSL) ou de la couche d'application (HTTP/HTTPS).

Pour mieux comprendre les différents types d'équilibreurs de charge disponibles, consultez les ressources suivantes :

- Qu'est-ce qu'Elastic Load Balancing ?
- Qu'est-ce qu'un équilibreur de charge Application Load Balancer ?
- Qu'est-ce qu'un équilibreur de charge Network Load Balancer ?
- Qu'est-ce qu'un équilibreur de charge Gateway Load Balancer ?
- Qu'est-ce qu'un équilibreur de charge Classic Load Balancer ?

Préparez-vous à associer un équilibreur de charge Elastic Load Balancing à votre groupe Auto Scaling

Avant d'associer un équilibreur de charge Elastic Load Balancing à votre groupe Auto Scaling, vous devez remplir les conditions préalables suivantes :

- Vous devez déjà avoir créé l'équilibreur de charge et le groupe cible utilisés pour acheminer le trafic vers votre groupe Auto Scaling.

  Il existe deux méthodes pour créer l'équilibreur de charge et le groupe cible :

  - Utilisation d'Elastic Load Balancing : suivez les procédures décrites dans la documentation d'Elastic Load Balancing pour créer et configurer l'équilibreur de charge et le groupe cible avant de créer le groupe Auto Scaling. Ignorer l'étape d'enregistrement de vos instances Amazon EC2. Amazon EC2 Auto Scaling prend automatiquement en charge l'enregistrement (et le désenregistrement) des instances lorsque vous associez un groupe cible à votre groupe Auto Scaling. Pour plus d'informations, consultez Prise en main d'Elastic Load Balancing dans le Guide de l'utilisateur Elastic Load Balancing.
  - Utilisation d'Amazon EC2 Auto Scaling : créez, configurez et associez l'équilibreur de charge et le groupe cible à l'aide d'une configuration de base depuis la console Amazon EC2 Auto Scaling. Pour plus d'informations, consultez Configurer un équilibreur de charge Application Load Balancer ou Network Load Balancer depuis la console Amazon EC2 Auto Scaling.

- Avant de créer un équilibreur de charge, déterminez le type d'équilibreur de charge dont vous avez besoin. Pour plus d'informations, consultez Types d'équilibreurs de charge Elastic Load Balancing.
- L'équilibreur de charge et son groupe cible doivent se trouver dans le même Compte AWS VPC et dans la même région que votre groupe Auto Scaling.
- Le groupe cible doit préciser un type de `instance` cible. Vous ne pouvez pas préciser un type de `ip` cible lorsque vous utilisez un groupe Auto Scaling.
- Si le modèle de lancement de votre groupe Auto Scaling ne contient pas le groupe de sécurité approprié pour autoriser le trafic entrant nécessaire depuis l'équilibreur de charge, vous devez mettre à jour le modèle de lancement. Les règles recommandées dépendent du type d'équilibreur de charge et des types de backends que l'équilibreur de charge utilise. Par exemple, pour acheminer le trafic vers des serveurs web, autorisez l'accès HTTP entrant sur le port 80 à partir de l'équilibreur de charge. Les instances existantes ne sont pas mises à jour avec les nouveaux paramètres lorsque le modèle de lancement est modifié. Pour mettre à jour les instances existantes, vous pouvez lancer une actualisation des instances pour les remplacer. Pour plus d'informations, consultez Utiliser une actualisation d'instance pour mettre à jour les instances d'un groupe Auto Scaling.
- Les groupes de sécurité du modèle de lancement doivent également autoriser l'accès depuis l'équilibreur de charge sur le port approprié pour qu'Elastic Load Balancing puisse effectuer ses contrôles de santé.
- Lors du déploiement d'appareils virtuels derrière un Gateway Load Balancer, l'Amazon Machine Image (AMI) figurant dans le modèle de lancement doit spécifier l'ID d'une AMI compatible avec le protocole GENEVE afin de permettre au groupe Auto Scaling d'échanger du trafic avec un Gateway Load Balancer. En outre, les groupes de sécurité du modèle de lancement doivent autoriser le trafic UDP sur le port 6081.

> **Tip**
>
> Si vous avez des scripts d'amorçage qui prennent un certain temps à se terminer, vous pouvez éventuellement ajouter un hook de cycle de vie de lancement à votre groupe Auto Scaling pour retarder l'enregistrement des instances derrière l'équilibreur de charge, avant que vos scripts d'amorçage se soient déroulés avec succès et que les applications présentes sur les instances soient prêtes à accepter le trafic. Vous ne pouvez pas ajouter un hook de cycle de vie lorsque vous créez initialement un groupe Auto Scaling dans la console Amazon EC2 Auto Scaling. Cependant, vous pouvez ajouter un hook de cycle de vie une fois le

groupe créé. Pour plus d'informations, consultez Hooks de cycle de vie Amazon EC2 Auto Scaling.

Configuration des contrôles de santé pour les cibles

Vous pouvez configurer des contrôles de santé pour vos cibles enregistrées auprès d'un équilibreur de charge Elastic Load Balancing afin de vous assurer qu'elles sont en mesure de gérer correctement le trafic. Les étapes spécifiques varient en fonction du type d'équilibreur de charge que vous utilisez. Pour plus d'informations, consultez les ressources suivantes :

- Application Load Balancer : consultez la section Contrôles de santé de vos groupes cibles dans le guide de l'utilisateur pour les équilibreurs de charge d'application.
- Network Load Balancer : consultez la section Contrôles de santé de vos groupes cibles dans le Guide de l'utilisateur pour les Network Load Balancers.
- Gateway Load Balancer — Consultez les bilans de santé de vos groupes cibles dans le guide de l'utilisateur des Gateway Load Balancers.
- Classic Load Balancer : voir Configurer les contrôles de santé de votre Classic Load Balancer dans le guide de l'utilisateur des Classic Load Balancer.

Par défaut, Amazon EC2 Auto Scaling ne considère pas qu'une instance est défectueuse et la remplace si elle échoue aux tests de santé d'Elastic Load Balancing. Les surveillances d'état par défaut d'un groupe Auto Scaling correspondent uniquement aux surveillances de l'état EC2. Pour plus d'informations, consultez Surveillance de l'état des instances dans un groupe Auto Scaling.

Pour permettre à Amazon EC2 Auto Scaling de remplacer les instances signalées comme défectueuses par Elastic Load Balancing, vous pouvez configurer votre groupe Auto Scaling pour qu'il utilise les bilans de santé d'Elastic Load Balancing. Ce faisant, Amazon EC2 Auto Scaling considère que l'instance n'est pas saine si elle échoue aux tests de santé EC2 ou à ceux d'Elastic Load Balancing. Si vous attachez plusieurs groupes cibles d'équilibreurs de charge ou un équilibreur de charge Classic Load Balancer au groupe, ils doivent tous indiquer que l'instance est saine pour que le groupe la considère comme saine. Si l'un d'eux signale qu'une instance est défectueuse, le groupe Auto Scaling la remplace, même si d'autres la signalent comme saine.

Pour plus d'informations sur la façon d'activer ces contrôles de santé pour votre groupe Auto Scaling, consultezAssociez un équilibreur de charge Elastic Load Balancing à votre groupe Auto Scaling.

> **Note**
>
> Pour que ces bilans de santé débutent dès que possible, assurez-vous que le délai de grâce du bilan de santé de votre groupe n'est pas trop élevé, mais suffisamment élevé pour que vos bilans de santé Elastic Load Balancing puissent déterminer si une cible est disponible pour traiter les demandes. Pour plus d'informations, consultez Définir la période de grâce de la surveillance de l'état pour un groupe Auto Scaling.

Associez un équilibreur de charge Elastic Load Balancing à votre groupe Auto Scaling

Cette rubrique décrit comment associer un équilibreur de charge Elastic Load Balancing à un groupe Auto Scaling. Il décrit également comment activer les contrôles de santé d'Elastic Load Balancing pour permettre à Amazon EC2 Auto Scaling de remplacer les instances signalées par Elastic Load Balancing comme étant défectueuses.

Par défaut, Amazon EC2 Auto Scaling remplace uniquement les instances non saines ou inaccessibles sur la base des surveillances de l'état Amazon EC2. Si vous activez les contrôles de santé d'Elastic Load Balancing, Amazon EC2 Auto Scaling peut remplacer une instance en cours d'exécution si l'un des équilibreurs de charge Elastic Load Balancing que vous associez au groupe Auto Scaling indique qu'il ne fonctionne pas correctement.

Pour un didacticiel sur l'attachement d'un Application Load Balancer à votre groupe Auto Scaling, consultez. Didacticiel : configurer une application redimensionnée et à charge équilibrée

> **Important**
>
> Avant de continuer, remplissez l'ensemble des conditions préalables de la section précédente.

Table des matières

Associer un groupe cible ou un Classic Load Balancer

Lorsque vous créez ou mettez à jour un groupe Auto Scaling, vous pouvez y associer un ou plusieurs groupes cibles ou Classic Load Balancers. Lorsque vous attachez un Application Load Balancer, un Network Load Balancer ou un Gateway Load Balancer, vous attachez un groupe cible plutôt que l'équilibreur de charge lui-même.

Suivez les étapes de cette section pour utiliser la console afin de :

- Associer un groupe cible ou un Classic Load Balancer à un groupe Auto Scaling
- Activez les tests de santé pour Elastic Load Balancing

Pour attacher un équilibreur de charge existant lors de la création d'un groupe Auto Scaling

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Dans la barre de navigation en haut de l'écran, choisissez Région AWS celui dans lequel vous avez créé votre équilibreur de charge.
3. Choisissez Créer un groupe Auto Scaling.
4. Aux étapes 1 et 2, choisissez les options souhaitées et passez à Étape 3 : configurer des options avancées.
5. Dans le champ Équilibrage de charge, choisissez Attacher à un équilibreur de charge existant.
6. Sous Attacher à un équilibreur de charge existant, effectuez l'une des opérations suivantes :

   a. Pour les équilibreurs de charge Application Load Balancer, Network Load Balancer et Gateway Load Balancer :

      Sélectionnez Choisir parmi vos groupes cibles d'équilibreurs de charge, puis choisissez un groupe cible dans le champ Groupes cibles d'équilibreurs de charge existants.

   b. Pour les équilibreurs de charge Classic Load Balancer :

      Sélectionnez Choisir parmi les équilibreurs de charge Classic Load Balancer, puis choisissez votre équilibreur de charge dans le champ Équilibreurs de charge Classic Load Balancer.

7. (Facultatif) Pour les surveillances de l'état et les types de surveillance de l'état supplémentaires, sélectionnez Activer les surveillances de l'état Elastic Load Balancing.

8. (Facultatif) Dans le champ Période de grâce de la surveillance de l'état, saisissez le délai en secondes. Il s'agit de la durée pendant laquelle Amazon EC2 Auto Scaling doit attendre avant de procéder à la surveillance de l'état d'une instance une fois qu'elle est passée à l'état `InService`. Pour plus d'informations, consultez Définir la période de grâce de la surveillance de l'état pour un groupe Auto Scaling.
9. Procédez à la création du groupe Auto Scaling. Une fois votre groupe Auto Scaling créé, vos instances seront automatiquement enregistrées dans l'équilibreur de charge.

Pour attacher un équilibreur de charge existant à votre groupe Auto Scaling une fois celui-ci créé

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page Auto Scaling groups (Groupes Auto Scaling).
3. Sous l'onglet Details (Détails), choisissez Load balancing (Équilibrage de charge), Edit (Modifier).
4. Sous Load balancing (Équilibrage de charge), effectuez l'une des opérations suivantes :

   a. Pour les Groupes cibles d'équilibreurs de charge Application Load Balancer, Network Load Balancer ou Gateway Load Balancer, cochez la case et choisissez un groupe cible.

   b. Pour les Équilibreurs de charge Classic Load Balancer, cochez la case et choisissez votre équilibreur de charge.
5. Choisissez Mettre à jour.

Lorsque vous avez terminé de fixer l'équilibreur de charge, vous pouvez éventuellement activer les contrôles de santé qui l'utilisent.

Pour activer les contrôles de santé d'Elastic Load Balancing

1. Sous l'onglet Détails choisissez Vérifications de l'états, Modifier.
2. Pour les surveillances de l'état et les types de surveillances de l'état supplémentaires, sélectionnez Activer les surveillances de l'état Elastic Load Balancing.
3. Dans le champ Période de grâce de la surveillance de l'état, saisissez le délai en secondes. Il s'agit de la durée pendant laquelle Amazon EC2 Auto Scaling doit attendre avant de procéder à la surveillance de l'état d'une instance une fois qu'elle est passée à l'état `InService`. Pour plus

d'informations, consultez Définir la période de grâce de la surveillance de l'état pour un groupe Auto Scaling.

4. Choisissez Mettre à jour.

> **Note**
>
> Vous pouvez surveiller l'état de l'équilibreur de charge lorsqu'il est attaché à l'aide de l' AWS CLI. Lorsqu'Amazon EC2 Auto Scaling a enregistré avec succès les instances et qu'au moins une instance enregistrée réussit les surveillances de l'état, vous obtenez un état `InService`. Pour plus d'informations, consultez Vérifier l'état d'attachement de votre équilibreur de charge.

Détacher un groupe cible ou un Classic Load Balancer

Lorsque vous n'avez plus besoin de l'équilibreur de charge, suivez la procédure ci-dessous pour le détacher de votre groupe Auto Scaling.

Pour détacher un équilibreur de charge d'un groupe

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Activez la case à cocher en regard d'un groupe existant.

   Un volet fractionné s'ouvre en bas de la page Auto Scaling groups (Groupes Auto Scaling).
3. Sous l'onglet Details (Détails), choisissez Load balancing (Équilibrage de charge), Edit (Modifier).
4. Sous Load balancing (Équilibrage de charge), effectuez l'une des opérations suivantes :

   a. Pour les Groupes cibles d'équilibreurs de charge Application Load Balancer, Network Load Balancer ou Gateway Load Balancer, choisissez l'icône de suppression (X) en regard du groupe cible.

   b. Pour les Classic Load Balancers (Équilibreurs de charge Classic Load Balancer), choisissez l'icône de suppression (X) en regard de l'équilibreur de charge.
5. Choisissez Mettre à jour.

Lorsque vous avez terminé de détacher le groupe cible, vous pouvez désactiver les tests de santé d'Elastic Load Balancing.

Pour désactiver les contrôles de santé d'Elastic Load Balancing

1. Sous l'onglet Détails choisissez Vérifications de l'états, Modifier.
2. Pour les bilans de santé et les types de bilans de santé supplémentaires, désélectionnez Turn on Elastic Load Balancing health checks.
3. Choisissez Mettre à jour.

Configurer un équilibreur de charge Application Load Balancer ou Network Load Balancer depuis la console Amazon EC2 Auto Scaling

Suivez la procédure ci-dessous pour créer et attacher un équilibreur de charge Application Load Balancer ou Network Load Balancer lors de la création de votre groupe Auto Scaling.

Pour créer et attacher un nouvel équilibreur de charge lors de la création d'un groupe Auto Scaling

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Choisissez Créer un groupe Auto Scaling.
3. Aux étapes 1 et 2, choisissez les options souhaitées et passez à Étape 3 : configurer des options avancées.
4. Dans le champ Équilibrage de charge, choisissez Attacher à un nouvel équilibreur de charge.

    a. Sous Attacher à un nouvel équilibreur de charge, accédez à Type d'équilibreur de charge et choisissez de créer un équilibreur de charge de type Application Load Balancer ou Network Load Balancer.

    b. Dans le champ Nom de l'équilibreur de charge, attribuez un nom à l'équilibreur de charge ou conservez le nom par défaut.

    c. Dans le champ Schéma de l'équilibreur de charge, choisissez de créer un équilibreur de charge accessible sur Internet ou de conserver la configuration par défaut, à savoir un équilibreur de charge interne.

    d. Dans le champ Zones de disponibilité et sous-réseaux, sélectionnez le sous-réseau public de chacune des zones de disponibilité dans lesquelles vous avez choisi de lancer vos instances EC2. (Ces champs sont préremplis à partir de l'étape 2).

e. Dans le champ Écouteurs et routage, mettez à jour le numéro de port de votre écouteur (si nécessaire), et sous Routage par défaut, choisissezCréer un groupe cible. Vous pouvez également choisir un groupe cible existant dans la liste déroulante.

f. Si vous avez choisi Create a target group (Créer un groupe cible) à l'étape précédente, dans le champ New target group name (Nom du nouveau groupe cible), attribuez un nom au groupe cible ou conservez le nom par défaut.

g. Pour ajouter des étiquettes à votre équilibreur de charge, choisissez Add tag (Ajouter une étiquette), et fournissez une clé et une valeur d'étiquette pour chaque étiquette.

5. (Facultatif) Pour les surveillances de l'état et les types de surveillance de l'état supplémentaires, sélectionnez Activer les surveillances de l'état Elastic Load Balancing.

6. (Facultatif) Dans le champ Période de grâce de la surveillance de l'état, saisissez le délai en secondes. Il s'agit de la durée pendant laquelle Amazon EC2 Auto Scaling doit attendre avant de procéder à la surveillance de l'état d'une instance une fois qu'elle est passée à l'état `InService`. Pour plus d'informations, consultez Définir la période de grâce de la surveillance de l'état pour un groupe Auto Scaling.

7. Procédez à la création du groupe Auto Scaling. Une fois votre groupe Auto Scaling créé, vos instances seront automatiquement enregistrées dans l'équilibreur de charge.

> **Note**
>
> Une fois votre groupe Auto Scaling créé, vous pouvez utiliser la console Elastic Load Balancing pour créer des écouteurs supplémentaires. Ceci est utile si vous devez créer un écouteur avec un protocole sécurisé, tel que HTTPS, ou un écouteur UDP. Vous pouvez ajouter d'autres écouteurs aux équilibreurs de charge existants, à condition d'utiliser des ports distincts.

Vérifier l'état d'attachement de votre équilibreur de charge

Après avoir attaché un équilibreur de charge, il passe en statut `Adding` pendant l'enregistrement des instances dans le groupe. Lorsque toutes les instances du groupe sont enregistrées, il passe à l'état `Added`. Lorsqu'au moins une instance enregistrée réussit les surveillances de l'état, il passe en statut `InService`. Une fois que l'équilibreur de charge est passé à l'état `InService`, Amazon EC2 Auto Scaling peut résilier les instances signalées comme défectueuses et les remplacer. Si aucune instance enregistrée ne réussit les surveillances d'état (par exemple, en raison d'une mauvaise

configuration de celles-ci), l'équilibreur de charge ne passe pas à l'état `InService`. Amazon EC2 Auto Scaling ne résilie donc pas les instances et ne les remplace pas.

Lorsque vous détachez un équilibreur de charge, il passe en statut `Removing` pendant le désenregistrement des instances dans le groupe. Les instances restent en cours d'exécution après leur désenregistrement. Par défaut, Connection Draining (délai d'annulation d'enregistrement) est activé pour les équilibreurs de charge Application Load Balancer, Network Load Balancer et Gateway Load Balancer. Si Connection Draining est activé, Elastic Load Balancing attend que les demandes à la volée soient terminées ou que le délai maximal expire (selon la première éventualité) avant d'annuler l'enregistrement des instances.

Vous pouvez vérifier l'état de la pièce jointe à l'aide du AWS Command Line Interface (AWS CLI) ou AWS des SDK. Vous ne pouvez pas vérifier l'état d'attachement sur la console.

Pour utiliser le AWS CLI pour vérifier l'état de la pièce jointe

La commande describe-traffic-sources suivante renvoie l'état d'attachement de toutes les sources de trafic pour le groupe Auto Scaling indiqué.

```
aws autoscaling describe-traffic-sources --auto-scaling-group-name my-asg
```

L'exemple renvoie l'ARN du groupe cible Elastic Load Balancing attaché au groupe Auto Scaling, ainsi que l'état d'attachement du groupe cible dans l'élément `State`.

```
{
    "TrafficSources": [
        {
            "Identifier": "arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/1234567890123456",
            "State": "InService",
            "Type": "elbv2"
        }
    ]
}
```

Ajouter et supprimer des zones de disponibilité

Pour profiter de la sécurité et de la fiabilité de la redondance géographique, étendez votre groupe Auto Scaling sur plusieurs zones de disponibilité au sein de la région où vous travaillez, puis attachez un équilibreur de charge pour répartir le trafic entrant entre ces zones de disponibilité.

Lorsqu'une zone de disponibilité devient défaillante ou inaccessible, Amazon EC2 Auto Scaling lance de nouvelles instances dans une zone de disponibilité non affectée. Lorsque la zone de disponibilité défaillante redevient saine, Amazon EC2 Auto Scaling répartit de nouveau automatiquement les instances d'application de manière uniforme dans toutes les zones de disponibilité de votre groupe Auto Scaling. Pour ce faire, Amazon EC2 Auto Scaling tente de lancer de nouvelles instances dans la zone de disponibilité qui contient le moins d'instances. Si la tentative échoue, Amazon EC2 Auto Scaling tente de lancer des instances dans d'autres zones de disponibilité jusqu'à ce qu'il y parvienne.

Elastic Load Balancer crée un nœud d'équilibreur de charge dans chaque zone de disponibilité que vous activez pour l'équilibreur de charge. Si vous activez l'équilibrage de charge entre zones pour votre équilibreur de charge, chaque nœud d'équilibreur de charge répartit le trafic de manière uniforme entre les instances enregistrées dans toutes les zones de disponibilité activées. Si l'équilibrage de charge entre zones est désactivé, chaque nœud de l'équilibreur de charge répartit les demandes uniformément entre les instances enregistrées dans sa zone de disponibilité uniquement.

Vous devez spécifier au moins une zone de disponibilité lors de la création de votre groupe Auto Scaling. Par la suite, vous pourrez étendre la disponibilité de votre application en ajoutant une zone de disponibilité à votre groupe Auto Scaling et en activant cette zone de disponibilité pour votre équilibreur de charge (si celui-ci la prend en charge).

Table des matières

Ajouter une zone de disponibilité

Suivez la procédure ci-dessous pour étendre votre groupe Auto Scaling et votre équilibreur de charge à un sous-réseau dans une zone de disponibilité supplémentaire.

Pour ajouter une zone de disponibilité

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Activez la case à cocher en regard d'un groupe existant.

Un volet fractionné s'ouvre en bas de la page Auto Scaling groups (Groupes Auto Scaling).

3. Sous l'onglet Details (Détails) choisissez Network (Réseau), Edit (Modifier).
4. Dans le champ Sous-réseaux, choisissez le sous-réseau correspondant à la zone de disponibilité que vous souhaitez ajouter au groupe Auto Scaling.
5. Choisissez Mettre à jour.
6. Pour mettre à jour les zones de disponibilité de votre équilibreur de charge afin qu'il partage les mêmes zones de disponibilité que votre groupe Auto Scaling, procédez comme suit :

   a. Dans le panneau de navigation, sous Load Balancing (Équilibrage de charge), choisissez Load Balancers (Équilibreurs de charge).

   b. Choisissez votre équilibreur de charge .

   c. Effectuez l’une des actions suivantes :

      - Pour les équilibreurs de charge Application Load Balancer et Network Load Balancer :

        1. Sous l'onglet Description, accédez au champ Zones de disponibilité et choisissez Modifier les sous-réseaux.
        2. Sur la page Modifier les sous-réseaux, accédez à Zones de disponibilité et cochez la case correspondant à la zone de disponibilité à ajouter. Si cette zone ne comporte qu'un seul sous-réseau, celui-ci est sélectionné. En présence de plusieurs sous-réseaux dans cette zone, sélectionnez l'un d'eux.

      - Pour les équilibreurs de charge Classic Load Balancer situés dans un VPC :

        1. Sous l'onglet Instances, choisissez Edit Availability Zones (Modifier des zones de disponibilité).
        2. Sur la page Ajouter et supprimer des sous-réseaux, accédez à Sous-réseaux disponibles et sélectionnez le sous-réseau à l'aide de l'icône d'ajout (+) correspondante. Le sous-réseau est déplacé sous Selected subnets (Sous-réseaux sélectionnés).

   d. Choisissez Enregistrer.

Enlever une zone de disponibilité

Pour supprimer une zone de disponibilité de votre groupe Auto Scaling et de votre équilibreur de charge, suivez la procédure ci-dessous.

Pour supprimer une zone de disponibilité

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Activez la case à cocher en regard d'un groupe existant.

   Un volet fractionné s'ouvre en bas de la page Auto Scaling groups (Groupes Auto Scaling).
3. Sous l'onglet Details (Détails) choisissez Network (Réseau), Edit (Modifier).
4. Dans le champ Sous-réseaux, choisissez l'icône de suppression (X) du sous-réseau correspondant à la zone de disponibilité que vous souhaitez supprimer du groupe Auto Scaling. Si plusieurs sous-réseaux sont associés à cette zone, choisissez l'icône de suppression (X) pour chacun d'eux.
5. Choisissez Mettre à jour.
6. Pour mettre à jour les zones de disponibilité de votre équilibreur de charge afin qu'il partage les mêmes zones de disponibilité que votre groupe Auto Scaling, procédez comme suit :

   a. Dans le panneau de navigation, sous Load Balancing (Équilibrage de charge), choisissez Load Balancers (Équilibreurs de charge).

   b. Choisissez votre équilibreur de charge .

   c. Effectuez l’une des actions suivantes :

      - Pour les équilibreurs de charge Application Load Balancer et Network Load Balancer :

        1. Sous l'onglet Description, accédez au champ Zones de disponibilité et choisissez Modifier les sous-réseaux.
        2. Sur la page Modifier les sous-réseaux, accédez à Zones de disponibilité et décochez la case pour supprimer le sous-réseau associé à cette zone de disponibilité.

      - Pour les équilibreurs de charge Classic Load Balancer situés dans un VPC :

        1. Sous l'onglet Instances, choisissez Edit Availability Zones (Modifier des zones de disponibilité).
        2. Sur la page Ajouter et supprimer des sous-réseaux, accédez à Sous-réseaux disponibles et supprimez le sous-réseau à l'aide de l'icône de suppression (-) correspondante. Le sous-réseau est déplacé sous Sous-réseaux disponibles.

   d. Choisissez Enregistrer.

Ressources connexes

Amazon EC2 Auto Scaling rééquilibre votre groupe lorsque vous modifiez les zones de disponibilité. Cela implique le remplacement et la redistribution de certaines instances. Pour plus d'informations, consultez Exemple : répartir les instances dans les zones de disponibilité.

Si vous avez enregistré des cibles dans des zones de disponibilité qui ne sont pas activées pour l'équilibreur de charge, celui-ci n'achemine pas le trafic vers ces cibles. Pour de plus amples informations, consultez la section Fonctionnement d'Elastic Load Balancing, dans le Guide de l'utilisateur Elastic Load Balancing.

Limites

Pour mettre à jour les zones de disponibilité activées pour votre équilibreur de charge, vous devez tenir compte des limitations suivantes :

- Lorsque vous activez une zone de disponibilité pour votre équilibreur de charge, vous spécifiez un sous-réseau depuis cette zone de disponibilité. Notez que vous ne pouvez activer qu'un seul sous-réseau par zone de disponibilité pour votre équilibreur de charge.
- Pour les équilibreurs de charge accessibles sur Internet, les sous-réseaux que vous spécifiez doivent avoir au moins huit adresses IP disponibles.
- Pour les équilibreurs de charge Application Load Balancer, vous devez activer au moins deux zones de disponibilité.
- Pour les équilibreurs de charge Network Load Balancer, vous ne pouvez pas désactiver les zones de disponibilité activées, mais vous pouvez en activer d'autres.
- Pour les équilibreurs de charge Gateway, vous ne pouvez pas désactiver les zones de disponibilité activées, mais vous pouvez en activer d'autres.

Exemples d'utilisation d'Elastic Load Balancing avec le AWS Command Line Interface

Utilisez le AWS Command Line Interface (AWS CLI) pour attacher, détacher et décrire les équilibreurs de charge et les groupes cibles, ajouter et supprimer les tests de santé d'Elastic Load Balancing et modifier les zones de disponibilité activées.

Cette rubrique présente des exemples de AWS CLI commandes qui exécutent des tâches courantes pour Amazon EC2 Auto Scaling.

> **Important**
>
> Pour d'autres exemples de commandes, consultez `aws elbv2` et `aws elb` dans le guide de référence des commandes AWS CLI .

Table des matières

Attachez votre groupe cible ou votre équilibreur Classic Load Balancer

Utilisez la commande create-auto-scaling-group suivante pour créer un groupe Auto Scaling et attacher simultanément un groupe cible en indiquant son Amazon Resource Name (ARN). Le groupe cible peut être associé à un équilibreur Application Load Balancer, un équilibreur Network Load Balancer ou un équilibreur Gateway Load Balancer.

Remplacez les valeurs d'exemple de `--auto-scaling-group-name`, `--vpc-zone-identifier`, `--min-size`, et `--max-size`. Pour l'option `--launch-template`, remplacez *`my-launch-template`* et *`1`* par le nom et la version d'un modèle de lancement de votre groupe Auto Scaling. Pour l'option `--traffic-sources`, remplacez l'exemple d'ARN par l'ARN d'un groupe cible pour un équilibreur Application Load Balancer, Network Load Balancer ou Gateway Load Balancer.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-template LaunchTemplateName=my-launch-template,Version='1' \
  --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
  --min-size 1 --max-size 5 \
  --traffic-sources "Identifier=arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/12345678EXAMPLE1"
```

Utilisez la commande attach-traffic-sources pour attacher des groupes cibles supplémentaires au groupe Auto Scaling après sa création.

La commande suivante ajoute un autre groupe cible au même groupe.

```
aws autoscaling attach-traffic-sources --auto-scaling-group-name my-asg \
  --traffic-sources "Identifier=arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/12345678EXAMPLE2"
```

Vous pouvez également attacher un équilibreur Classic Load Balancer à votre groupe. Renseignez les options `--traffic-sources` et `--type` lorsque vous utilisez create-auto-scaling-group ou attach-traffic-sources, comme dans l'exemple suivant. Remplacez *my-classic-load-balancer* par le nom d'un équilibreur Classic Load Balancer. Pour l'option `--type`, indiquez une valeur de **elb**.

```
--traffic-sources "Identifier=my-classic-load-balancer" --type elb
```

Décrivez vos groupes cibles ou vos équilibreurs Classic Load Balancer

Pour décrire les équilibreurs de charge ou les groupes cibles attachés à votre groupe Auto Scaling, utilisez la commande describe-traffic-sources suivante. Remplacez *my-asg* par le nom de votre groupe.

```
aws autoscaling describe-traffic-sources --auto-scaling-group-name my-asg
```

L'exemple renvoie l'ARN des groupes cibles Elastic Load Balancing que vous avez attachés au groupe Auto Scaling.

```
{
    "TrafficSources": [
        {
            "Identifier": "arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/12345678EXAMPLE1",
            "State": "InService",
            "Type": "elbv2"
        },
        {
            "Identifier": "arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/12345678EXAMPLE2",
            "State": "InService",
            "Type": "elbv2"
        }
    ]
```

```
}
```

Pour en savoir plus sur le champ `State` de la sortie, consultez Vérifier l'état d'attachement de votre équilibreur de charge.

Ajouter des surveillances d'état Elastic Load Balancing

Pour ajouter des surveillances de l'état Elastic Load Balancing aux surveillances de l'état des instances de votre groupe Auto Scaling, utilisez la commande update-auto-scaling-group suivante et définissez l'option `--health-check-type` sur **ELB**. Remplacez *my-asg* par le nom de votre groupe.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-type "ELB"
```

Les nouvelles instances ont souvent besoin de temps pour un bref échauffement avant de pouvoir passer un bilan de santé. Si le délai de grâce ne fournit pas un temps de préchauffage suffisant, les instances peuvent ne pas sembler prêtes à traiter le trafic. Amazon EC2 Auto Scaling peut considérer ces instances comme défectueuses et les remplacer.

Pour mettre à jour la période de grâce de la surveillance de l'état, utilisez l'option `--health-check-grace-period` lorsque vous utilisez update-auto-scaling-group, comme dans l'exemple suivant. Remplacez *300* par le nombre de secondes nécessaires pour maintenir les nouvelles instances en service avant de les résilier si elles s'avèrent défectueuses.

```
--health-check-grace-period 300
```

Pour plus d'informations, consultez Surveillance de l'état des instances dans un groupe Auto Scaling.

Modifier vos zones de disponibilité

La modification de vos zones de disponibilité présente des limites que vous devez connaître. Pour plus d'informations, consultez Limites.

Pour modifier les zones de disponibilité d'un équilibreur Application Load Balancer ou Network Load Balancer

1. Avant de modifier les zones de disponibilité de l'équilibreur de charge, il est conseillé de mettre à jour les zones de disponibilité du groupe Auto Scaling afin de vérifier que vos types d'instances sont disponibles dans les zones indiquées.

Pour mettre à jour les zones de disponibilité de votre groupe Auto Scaling, utilisez la commande update-auto-scaling-group suivante. Remplacez les identifiants des sous-réseaux en exemple par les identifiants des sous-réseaux des zones de disponibilité à activer. Les sous-réseaux indiqués remplacent les sous-réseaux précédemment activés. Remplacez *my-asg* par le nom de votre groupe.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --vpc-zone-identifier "subnet-41767929,subnet-cb663da2,subnet-8360a9e7"
```

2. Utilisez la commande describe-auto-scaling-groups suivante pour vérifier que les instances dans les nouveaux sous-réseaux ont été lancées. Si les instances ont été lancées, la liste de ces instances apparaît, avec leur statut. Remplacez *my-asg* par le nom de votre groupe.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

3. Utilisez la commande set-subnets suivante pour indiquer les sous-réseaux de votre équilibreur de charge. Remplacez les identifiants des sous-réseaux en exemple par les identifiants des sous-réseaux des zones de disponibilité à activer. Vous pouvez spécifier un seul sous-réseau par zone de disponibilité. Les sous-réseaux indiqués remplacent les sous-réseaux précédemment activés. Remplacez *my-lb-arn* par l'ARN de votre équilibreur de charge.

```
aws elbv2 set-subnets --load-balancer-arn my-lb-arn \
  --subnets subnet-41767929 subnet-cb663da2 subnet-8360a9e7
```

Pour modifier les zones de disponibilité d'un équilibreur Classic Load Balancer

1. Avant de modifier les zones de disponibilité de l'équilibreur de charge, il est conseillé de mettre à jour les zones de disponibilité du groupe Auto Scaling afin de vérifier que vos types d'instances sont disponibles dans les zones indiquées.

   Pour mettre à jour les zones de disponibilité de votre groupe Auto Scaling, utilisez la commande update-auto-scaling-group suivante. Remplacez les identifiants des sous-réseaux en exemple par les identifiants des sous-réseaux des zones de disponibilité à activer. Les sous-réseaux indiqués remplacent les sous-réseaux précédemment activés. Remplacez *my-asg* par le nom de votre groupe.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
```

```
  --vpc-zone-identifier "subnet-41767929,subnet-cb663da2"
```

2. Utilisez la commande describe-auto-scaling-groups suivante pour vérifier que les instances dans les nouveaux sous-réseaux ont été lancées. Si les instances ont été lancées, la liste de ces instances apparaît, avec leur statut. Remplacez *my-asg* par le nom de votre groupe.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

3. Utilisez la commande attach-load-balancer-to-subnets suivante pour activer une nouvelle zone de disponibilité pour votre Classic Load Balancer. Remplacez l'identifiant du sous-réseau en exemple par l'identifiant du sous-réseau de la zone de disponibilité à activer. Remplacez *my-lb* par le nom de votre équilibreur de charge.

```
aws elb attach-load-balancer-to-subnets --load-balancer-name my-lb \
  --subnets subnet-cb663da2
```

   Pour désactiver une zone de disponibilité, exécutez la commande detach-load-balancer-from-subnets suivante. Remplacez l'identifiant du sous-réseau en exemple par l'identifiant du sous-réseau de la zone de disponibilité à désactiver. Remplacez *my-lb* par le nom de votre équilibreur de charge.

```
aws elb detach-load-balancer-from-subnets --load-balancer-name my-lb \
  --subnets subnet-8360a9e7
```

Détachez votre groupe cible ou votre équilibreur Classic Load Balancer

La commande detach-traffic-sources suivante vous permet de détacher un groupe cible d'un groupe Auto Scaling lorsque vous n'en avez plus besoin.

Pour l'option `--auto-scaling-group-name`, remplacez *my-asg* par le nom de votre groupe. Pour l'option `--traffic-sources`, remplacez l'exemple d'ARN par l'ARN d'un groupe cible pour un équilibreur Application Load Balancer, Network Load Balancer ou Gateway Load Balancer.

```
aws autoscaling detach-traffic-sources --auto-scaling-group-name my-asg \
  --traffic-sources "Identifier=arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/1234567890123456"
```

Pour détacher un équilibreur Classic Load Balancer de votre groupe, renseignez les options `--traffic-sources` et `--type`, comme dans l'exemple suivant. Remplacez *my-classic-load-*

balancer par le nom d'un équilibreur Classic Load Balancer. Pour l'option `--type`, indiquez une valeur de **elb**.

```
--traffic-sources "Identifier=my-classic-load-balancer" --type elb
```

Supprimer les surveillances de l'état Elastic Load Balancing

Pour enlever des surveillances de l'état Elastic Load Balancing de votre groupe Auto Scaling, utilisez la commande update-auto-scaling-group suivante et définissez l'option `--health-check-type` sur **EC2**. Remplacez *my-asg* par le nom de votre groupe.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-type "EC2"
```

Pour plus d'informations, consultez Surveillance de l'état des instances dans un groupe Auto Scaling.

Anciennes commandes

Les exemples suivants montrent comment utiliser les anciennes commandes CLI pour attacher, détacher et décrire les équilibreurs de charge et les groupes cibles. Ils servent de référence pour les clients qui souhaitent les utiliser. Nous continuons à prendre en charge les anciennes commandes CLI, mais nous vous recommandons d'utiliser les nouvelles commandes CLI « sources de trafic », qui peuvent attacher et détacher plusieurs types de sources de trafic. Vous pouvez utiliser les anciennes commandes CLI et les commandes CLI « sources de trafic » sur le même groupe Auto Scaling.

Attachez votre groupe cible ou votre équilibreur Classic Load Balancer (ancien)

Pour attacher votre groupe cible

La commande create-auto-scaling-group suivante permet de créer un groupe Auto Scaling auquel est attaché un groupe cible : Spécifiez l'Amazon Resource Name (ARN) d'un groupe cible pour un équilibreur de charge Application Load Balancer, Network Load Balancer ou Gateway Load Balancer.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-template LaunchTemplateName=my-launch-template,Version='1' \
  --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
  --target-group-arns "arn:aws:elasticloadbalancing:region:account-id:targetgroup/my-
targets/1234567890123456" \
  --min-size 1 --max-size 5
```

La commande attach-load-balancer-target-groups suivante permet d'attacher un groupe cible à un groupe Auto Scaling existant.

```
aws autoscaling attach-load-balancer-target-groups --auto-scaling-group-name my-asg \
  --target-group-arns "arn:aws:elasticloadbalancing:region:account-id:targetgroup/my-targets/1234567890123456"
```

Pour attacher votre Classic Load Balancer

La commande create-auto-scaling-group suivante permet de créer un groupe Auto Scaling auquel est attaché un équilibreur de charge Classic Load Balancer.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-configuration-name my-launch-config \
  --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
  --load-balancer-names "my-load-balancer" \
  --min-size 1 --max-size 5
```

La commande attach-load-balancers suivante permet d'attacher l'équilibreur de charge Classic Load Balancer spécifié à un groupe Auto Scaling existant.

```
aws autoscaling attach-load-balancers --auto-scaling-group-name my-asg \
  --load-balancer-names my-lb
```

Décrivez votre groupe cible ou votre équilibreur Classic Load Balancer (ancien)

Pour décrire les groupes cibles

Pour décrire les groupes cibles associés à un groupe Auto Scaling, utilisez la commande describe-load-balancer-target-groups. L'exemple suivant répertorie les groupes cibles de *my-asg*.

```
aws autoscaling describe-load-balancer-target-groups --auto-scaling-group-name my-asg
```

Pour décrire les Classic Load Balancer.

Pour décrire les équilibreurs de charge Classic Load Balancer associés à un groupe Auto Scaling, utilisez la commande describe-load-balancers. L'exemple suivant répertorie les équilibreurs de charge Classic Load Balancer de *my-asg*.

```
aws autoscaling describe-load-balancers --auto-scaling-group-name my-asg
```

Détachez votre groupe cible ou votre équilibreur Classic Load Balancer (ancien)

Pour détacher un groupe cible

La commande detach-load-balancer-target-groups suivante vous permet de détacher un groupe cible d'un groupe Auto Scaling lorsque vous n'en avez plus besoin.

```
aws autoscaling detach-load-balancer-target-groups --auto-scaling-group-name my-asg \
  --target-group-arns "arn:aws:elasticloadbalancing:region:account-id:targetgroup/my-targets/1234567890123456"
```

Pour détacher un équilibreur de charge Classic Load Balancer

La commande detach-load-balancers suivante vous permet de détacher un équilibreur de charge Classic Load Balancer du groupe Auto Scaling lorsque vous n'en avez plus besoin.

```
aws autoscaling detach-load-balancers --auto-scaling-group-name my-asg \
  --load-balancer-names my-lb
```

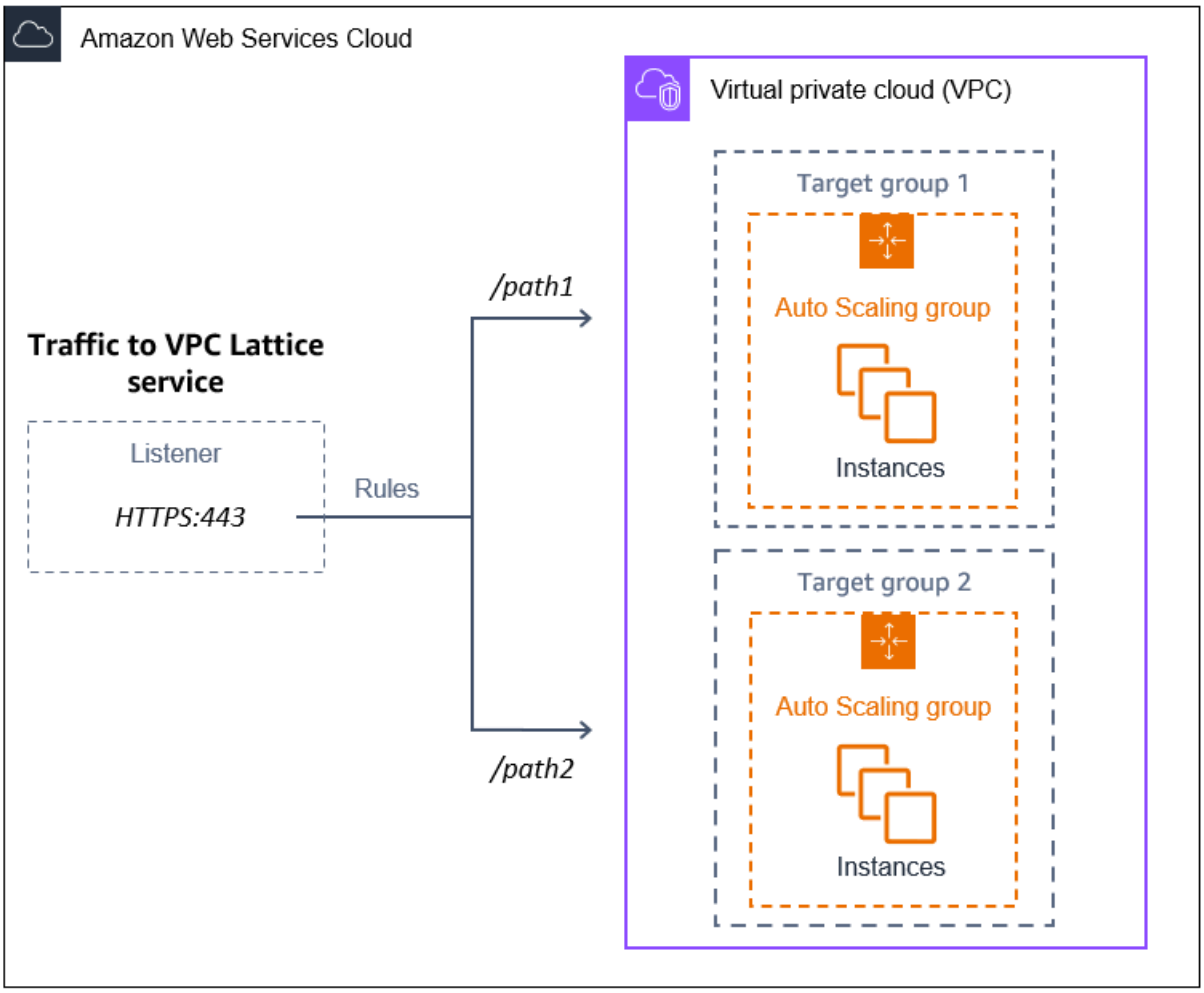
Acheminer le trafic vers votre groupe Auto Scaling avec un groupe cible VPC Lattice

Vous pouvez utiliser Amazon VPC Lattice pour gérer le flux de trafic et les appels d'API entre vos applications et services qui s'exécutent sur des ressources distinctes, comme les groupes Auto Scaling ou les fonctions Lambda. VPC Lattice est un service de mise en réseau d'applications que vous pouvez utiliser pour connecter, sécuriser et surveiller vos services sur plusieurs comptes et clouds privés virtuels (VPC). Pour en savoir plus sur VPC Lattice, voir Qu'est-ce que VPC Lattice ?

Pour commencer à utiliser VPC Lattice, créez d'abord les ressources VPC Lattice nécessaires pour permettre aux ressources d'un VPC associé à un réseau de services de se connecter entre elles. Ces ressources incluent les services, les écouteurs, les règles des écouteurs et les groupes cibles.

Pour associer un groupe Auto Scaling à un service VPC Lattice, créez un groupe cible pour le service qui achemine les demandes vers les instances enregistrées par ID d'instance, et ajoutez au service un écouteur qui envoie les demandes au groupe cible. Puis, attachez le groupe cible au groupe Auto Scaling. Amazon EC2 Auto Scaling enregistre automatiquement les instances EC2 en tant que cibles auprès du groupe cible. Plus tard, lorsqu'Amazon EC2 Auto Scaling doit mettre fin à une instance, il désenregistre automatiquement l'instance du groupe cible avant la résiliation.

Une fois que vous avez attaché le groupe cible, c'est le point d'entrée de toutes les demandes entrantes vers votre groupe Auto Scaling. Comme le montre l'exemple du schéma suivant, les demandes entrantes peuvent alors être acheminées vers le groupe cible approprié à l'aide des règles d'écoute spécifiées pour un service VPC Lattice.



Lorsque le trafic est acheminé via VPC Lattice vers votre groupe Auto Scaling, VPC Lattice équilibre les demandes entre les instances du groupe en utilisant la répartition de charge avec un routage en tourniquet. VPC Lattice peut également surveiller l'état de ses instances enregistrées et n'acheminer le trafic qu'aux instances saines.

Pour que vos instances restent disponibles pour les demandes entrantes, vous pouvez éventuellement ajouter des surveillance de l'état VPC Lattice à votre groupe Auto Scaling. Ainsi, si l'une des instances EC2 a une défaillance, votre groupe Auto Scaling lance automatiquement une nouvelle instance pour la remplacer. Le comportement des surveillances de l'état VPC Lattice est similaire à celui des surveillances de l'état Elastic Load Balancing. Les surveillances d'état par défaut d'un groupe Auto Scaling correspondent uniquement aux surveillances de l'état EC2.

Pour en savoir plus sur VPC Lattice, consultez Simplifier la connectivité, la sécurité et la surveillance entre services avec Amazon VPC Lattice, désormais disponible sur le blog. AWS

Table des matières

Se préparer à attacher un groupe cible VPC Lattice à votre groupe Auto Scaling

Avant d'attacher un groupe cible VPC Lattice à votre groupe Auto Scaling, vous devez remplir les conditions préalables suivantes :

- Vous devez déjà avoir créé un réseau de services VPC Lattice, un service, un écouteur et un groupe cible. Pour plus d'informations, consultez les rubriques suivantes dans le Guide de l'utilisateur VPC Lattice :
  - Réseaux de services
  - Services
  - Écouteurs
  - Groupes cibles
- Le groupe cible doit appartenir au même Compte AWS VPC et à la même région que votre groupe Auto Scaling.
- Le groupe cible doit préciser un type de `instance` cible. Vous ne pouvez pas préciser un type de `ip` cible lorsque vous utilisez un groupe Auto Scaling.
- Vous devez disposer d'autorisations IAM suffisantes pour attacher le groupe cible au groupe Auto Scaling. L'exemple de politique suivant illustre les autorisations minimales requises pour attacher et détacher des groupes cibles.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "autoscaling:AttachTrafficSources",
```

```
                "autoscaling:DetachTrafficSources",
                "autoscaling:DescribeTrafficSources",
                "vpc-lattice:RegisterTargets",
                "vpc-lattice:DeregisterTargets"
            ],
            "Resource": "*"
        }
    ]
}
```

- Si le modèle de lancement de votre groupe Auto Scaling ne contient pas les paramètres corrects pour VPC Lattice, tels qu'un groupe de sécurité compatible, vous devez mettre à jour le modèle de lancement. Les instances existantes ne sont pas mises à jour avec les nouveaux paramètres lorsque le modèle de lancement est modifié. Pour mettre à jour les instances existantes, vous pouvez lancer une actualisation des instances pour les remplacer. Pour plus d'informations, consultez Utiliser une actualisation d'instance pour mettre à jour les instances d'un groupe Auto Scaling.
- Avant d'activer les surveillances de l'état VPC Lattice sur votre groupe Auto Scaling, vous pouvez configurer une surveillance de l'état basée sur l'application pour vérifier que votre application répond comme prévu. Pour plus d'informations, consultez la section Surveillances de l'état de vos groupes cibles dans le Guide de l'utilisateur VPC Lattice.

Groupes de sécurité : règles entrantes et sortantes

Les groupes de sécurité font office de pare-feu pour les instances EC2 associées, en contrôlant le trafic entrant et le trafic sortant au niveau de l'instance.

> **Note**
>
> La configuration réseau est suffisamment complexe pour que nous vous recommandions fortement de créer un nouveau groupe de sécurité à utiliser avec VPC Lattice. Cela vous permet également de vous AWS Support aider plus facilement si vous devez les contacter. Les sections suivantes partent du principe que vous suivez cette recommandation.
> Pour en savoir plus sur la création de groupes de sécurité pour VPC Lattice que vous pouvez utiliser avec votre groupe Auto Scaling, consultez la section Contrôler le trafic à l'aide de groupes de sécurité dans le Guide de l'utilisateur VPC Lattice. Pour résoudre les problèmes liés au flux de trafic, consultez le Guide de l'utilisateur VPC Lattice pour plus d'informations.

Pour plus d'informations sur la création d'un groupe de sécurité, consultez la section Créer un groupe de sécurité dans le guide de l'utilisateur Amazon EC2 et utilisez le tableau suivant pour déterminer les options à sélectionner.

Option	Valeur
Nom	Un nom facile à retenir.
Description	Description de la règle du groupe de sécurité vous permettant de l'identifier.
VPC	Le même VPC que le groupe Auto Scaling.

Règles entrantes

Lorsque vous créez un groupe de sécurité, il n'existe pas de règles entrantes. Aucun trafic entrant issu de clients d'un réseau de service VPC Lattice vers votre instance n'est autorisé tant que vous n'avez pas ajouté des règles entrantes au groupe de sécurité.

Pour permettre aux clients d'un réseau de service VPC Lattice de se connecter aux instances de votre groupe Auto Scaling, le groupe de sécurité de votre groupe Auto Scaling doit être correctement configuré. Dans ce cas, attribuez-lui une règle entrante pour autoriser le trafic provenant du nom de la liste de AWS préfixes gérée pour VPC Lattice, au lieu d'une adresse IP spécifique. La liste des préfixes VPC Lattice est une plage d'adresses IP utilisées par VPC Lattice en notation CIDR. Pour plus d'informations, consultez la section Travailler avec des listes de préfixes AWS gérées dans le guide de l'utilisateur Amazon VPC.

Pour plus d'informations sur l'ajout de règles à un groupe de sécurité, consultez Ajouter des règles à votre groupe de sécurité dans le Guide de l'utilisateur Amazon VPC et utilisez le tableau suivant pour déterminer quelles options sélectionner.

Option	Valeur
Règle HTTP	Type : HTTP

Option	Valeur	
	Source : com.amazonaws.*region*.vpc-lattice	
Règle HTTP	Type : HTTPS Source : com.amazonaws.*region*.vpc-lattice	

Le groupe de sécurité est avec état : il autorise le trafic issu des clients du réseau de service VPC Lattice vers les instances de votre groupe Auto Scaling, puis renvoie la réponse au client qu'il a quitté.

Règles sortantes

Par défaut, un groupe de sécurité inclut une règle sortante qui autorise tout le trafic sortant. Vous pouvez éventuellement supprimer cette règle par défaut et ajouter une règle sortante pour répondre à des besoins de sécurité spécifiques.

Limites

- Les groupes d'instances mixtes ne sont pas pris en charge. Si vous essayez d'associer un groupe cible VPC Lattice à un groupe Auto Scaling doté d'une politique d'instances mixtes, vous recevez le message d'erreur Actuellement, les groupes Auto Scaling comportant des instances mixtes ne peuvent pas être intégrés à un service VPC Lattice. Cela est dû au fait que l'algorithme d'équilibrage de charge répartit uniformément la charge sur toutes les ressources disponibles et suppose que les instances sont suffisamment similaires pour gérer des charges égales.

Attacher un groupe cible VPC Lattice à votre groupe Auto Scaling

Cette rubrique explique comment attacher un groupe cible VPC Lattice à un groupe Auto Scaling. Elle décrit également comment activer les surveillances de l'état VPC Lattice pour permettre à Amazon EC2 Auto Scaling de remplacer les instances signalées par VPC Lattice comme non saines.

Par défaut, Amazon EC2 Auto Scaling remplace uniquement les instances non saines ou inaccessibles sur la base des surveillances de l'état Amazon EC2. Si vous activez les surveillances de l'état VPC Lattice, Amazon EC2 Auto Scaling peut remplacer une instance en cours d'exécution

si l'un des groupes cibles VPC Lattice que vous attachez au groupe Auto Scaling le signale comme étant non sain. Pour plus d'informations, consultez Surveillance de l'état des instances dans un groupe Auto Scaling.

> **Important**
>
> Avant de continuer, remplissez l'ensemble des conditions préalables de la section précédente.

Associer un groupe cible VPC Lattice

Vous pouvez associer un ou plusieurs groupes cibles à un groupe Auto Scaling lorsque vous créez ou mettez à jour le groupe.

Console

Suivez les étapes de cette section pour utiliser la console afin de :

- Attacher un groupe cible VPC Lattice à votre groupe Auto Scaling
- Activer les surveillances de l'état pour VPC Lattice

Pour attacher un groupe cible VPC Lattice à un nouveau groupe Auto Scaling

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Dans la barre de navigation située en haut de l'écran, choisissez la Région AWS dans laquelle vous avez créé votre groupe Auto Scaling.
3. Choisissez Créer un groupe Auto Scaling.
4. Aux étapes 1 et 2, choisissez vos options souhaitées et passez à Étape 3 : configurer des options avancées.
5. Pour Options d'intégration VPC Lattice, choisissez Attacher au service VPC Lattice.
6. Sous Choisir un groupe cible VPC Lattice, choisissez votre groupe cible.
7. (Facultatif) Pour les surveillances de l'état et les types de surveillance de l'état supplémentaires, sélectionnez Activer les surveillances de l'état du réseau VPC.
8. (Facultatif) Dans le champ Période de grâce de la surveillance de l'état, saisissez le délai en secondes. Il s'agit de la durée pendant laquelle Amazon EC2 Auto Scaling doit attendre

avant de procéder à la surveillance de l'état d'une instance une fois qu'elle est passée à l'état `InService`. Pour plus d'informations, consultez Définir la période de grâce de la surveillance de l'état pour un groupe Auto Scaling.

9. Procédez à la création du groupe Auto Scaling. Vos instances seront automatiquement enregistrées dans le groupe cible VPC Lattice une fois le groupe Auto Scaling créé.

Pour attacher un groupe cible VPC Lattice à un groupe Auto Scaling existant

Suivez la procédure ci-dessous pour attacher un groupe cible pour un service à un groupe Auto Scaling existant.

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Cochez la case située en regard de votre groupe Auto Scaling.

   Un volet fractionné s'ouvre en bas de la page.
3. Dans l'onglet Détails, choisissez Options d'intégration VPC Lattice, puis Modifier.
4. Sous Options d'intégration VPC Lattice, choisissez Attacher au service VPC Lattice.
5. Sous Choisir un groupe cible VPC Lattice, choisissez votre groupe cible.
6. Choisissez Mettre à jour.

Lorsque vous avez terminé d'associer le groupe cible, vous pouvez éventuellement activer les surveillances de l'état qui l'utilisent.

Activer les surveillances de l'état VPC Lattice

1. Sous l'onglet Détails choisissez Vérifications de l'états, Modifier.
2. Pour les surveillances de l'état et les types de surveillance de l'état supplémentaires, sélectionnez Activer les surveillances de l'état du réseau VPC.
3. Dans le champ Période de grâce de la surveillance de l'état, saisissez le délai en secondes. Il s'agit de la durée pendant laquelle Amazon EC2 Auto Scaling doit attendre avant de procéder à la surveillance de l'état d'une instance une fois qu'elle est passée à l'état `InService`. Pour plus d'informations, consultez Définir la période de grâce de la surveillance de l'état pour un groupe Auto Scaling.
4. Choisissez Mettre à jour.

AWS CLI

Suivez les étapes décrites dans cette section pour utiliser le AWS CLI pour :

- Attacher un groupe cible VPC Lattice à votre groupe Auto Scaling
- Activer les surveillances de l'état pour VPC Lattice

Pour attacher un groupe cible VPC Lattice à votre groupe Auto Scaling

Utilisez la commande create-auto-scaling-group suivante pour créer un groupe Auto Scaling et attacher simultanément un groupe cible VPC Lattice en indiquant son Amazon Resource Name (ARN).

Remplacez les valeurs d'exemple de `--auto-scaling-group-name`, `--vpc-zone-identifier`, `--min-size`, et `--max-size`. Pour l'option `--launch-template`, remplacez *`my-launch-template`* et *`1`* par le nom et la version du modèle de lancement que vous avez créé pour les instances enregistrées auprès d'un groupe cible VPC Lattice. Pour l'option `--traffic-sources`, remplacez l'exemple d'ARN par l'ARN de votre groupe cible VPC Lattice.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-template LaunchTemplateName=my-launch-template,Version='1' \
  --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
  --min-size 1 --max-size 5 \
  --traffic-sources "Identifier=arn:aws:vpc-lattice:region:account-id:targetgroup/
tg-0e2f2665eEXAMPLE"
```

Utilisez la commande attach-traffic-sources pour attacher un groupe cible VPC Lattice au groupe Auto Scaling après sa création.

```
aws autoscaling attach-traffic-sources --auto-scaling-group-name my-asg \
  --traffic-sources "Identifier=arn:aws:vpc-lattice:region:account-id:targetgroup/
tg-0e2f2665eEXAMPLE"
```

Pour activer les surveillances de l'état pour VPC Lattice

Si vous avez configuré un contrôle de santé basé sur les applications pour votre groupe cible VPC Lattice, vous pouvez activer ces surveillances de l'état. Utilisez les commandes create-auto-scaling-group ou update-auto-scaling-group avec l'option `--health-check-type` et une valeur de **`VPC_LATTICE`**. Pour spécifier le délai de grâce pour les surveillances de l'état effectuées par

votre groupe Auto Scaling, incluez l'option `--health-check-grace-period` et indiquez sa valeur en secondes.

```
--health-check-type "VPC_LATTICE" --health-check-grace-period 60
```

Détacher un groupe cible VPC Lattice

Lorsque vous n'avez plus besoin d'utiliser VPC Lattice, suivez la procédure ci-dessous pour détacher le groupe cible de votre groupe Auto Scaling.

Console

Suivez les étapes de cette section pour utiliser la console afin de :

- Détacher un groupe cible VPC Lattice d'un groupe Auto Scaling
- Désactiver les surveillances de l'état pour VPC Lattice

Pour détacher un groupe cible VPC Lattice d'un groupe Auto Scaling

1. Ouvrez la console Amazon EC2 à l'adresse https://console.aws.amazon.com/ec2/ et choisissez Groupes Auto Scaling dans le panneau de navigation.
2. Activez la case à cocher en regard d'un groupe existant.

   Un volet fractionné s'ouvre en bas de la page.
3. Dans l'onglet Détails, choisissez Options d'intégration VPC Lattice, puis Modifier.
4. Sous Options d'intégration VPC Lattice, choisissez l'icône de suppression (X) en regard du groupe cible.
5. Choisissez Mettre à jour.

Lorsque vous avez terminé de détacher le groupe cible, vous pouvez désactiver les surveillances de l'état du VPC Lattice.

Pour désactiver les surveillances de l'état VPC Lattice

1. Sous l'onglet Détails choisissez Vérifications de l'états, Modifier.
2. Pour les surveillances de l'état et les types de surveillance de l'état supplémentaires, désélectionnez Activer les surveillances de l'état du réseau VPC.

3. Choisissez Mettre à jour.

AWS CLI

Suivez les étapes décrites dans cette section pour utiliser le AWS CLI pour :

- Détacher un groupe cible VPC Lattice d'un groupe Auto Scaling
- Désactiver les surveillances de l'état pour VPC Lattice

Utilisez la commande detach-traffic-sources pour détacher un groupe cible de votre groupe Auto Scaling lorsque vous n'en avez plus besoin.

```
aws autoscaling detach-traffic-sources --auto-scaling-group-name my-asg \
  --traffic-sources "Identifier=arn:aws:vpc-lattice:region:account-id:targetgroup/
tg-0e2f2665eEXAMPLE"
```

Pour mettre à jour les surveillances de l'état d'un groupe Auto Scaling afin qu'il n'utilise plus les surveillances de l'état VPC Lattice, utilisez la commande update-auto-scaling-group. Incluez l'option `--health-check-type` et une valeur de**EC2**.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-type "EC2"
```

Vérifier l'état d'attachement de votre groupe cible VPC Lattice

Une fois que vous avez attaché un groupe cible VPC Lattice à un groupe Auto Scaling, celui-ci entre dans l'état `Adding` lors de l'enregistrement des instances dans le groupe. Lorsque toutes les instances du groupe sont enregistrées, il passe à l'état `Added`. Lorsqu'au moins une instance enregistrée réussit les surveillances de l'état, il passe en statut `InService`. Une fois que le groupe cible est passé à l'état `InService`, Amazon EC2 Auto Scaling peut résilier les instances signalées comme non saines et les remplacer. Si aucune instance enregistrée ne réussit les surveillances d'état (par exemple, en raison d'une mauvaise configuration de la surveillance de l'état), le groupe cible ne passe pas à l'état `InService`. Amazon EC2 Auto Scaling ne résilie donc pas les instances et ne les remplace pas.

Lorsque vous détachez un groupe cible, il passe en statut `Removing` pendant le désenregistrement des instances dans le groupe. Les instances restent en cours d'exécution après leur

désenregistrement. Par défaut, le drainage de la connexion (délai d'annulation d'enregistrement) est activé. Si le drainage de la connexion est activé, VPC Lattice attend que les demandes à la volée soient terminées ou que le délai maximal expire (selon la première éventualité) avant d'annuler l'enregistrement des instances.

Vous pouvez vérifier l'état de la pièce jointe à l'aide du AWS Command Line Interface (AWS CLI) ou AWS des SDK. Vous ne pouvez pas vérifier l'état d'attachement sur la console.

Pour utiliser le AWS CLI pour vérifier l'état de la pièce jointe

La commande describe-traffic-sources suivante renvoie l'état d'attachement de toutes les sources de trafic pour le groupe Auto Scaling indiqué.

```
aws autoscaling describe-traffic-sources --auto-scaling-group-name my-asg
```

L'exemple renvoie l'ARN du groupe cible VPC Lattice attaché au groupe Auto Scaling, ainsi que l'état d'attachement du groupe cible dans l'élément `State`.

```
{
    "TrafficSources": [
        {
            "Identifier": "arn:aws:vpc-lattice:region:account-
id:targetgroup/tg-0e2f2665eEXAMPLE",
            "State": "InService",
            "Type": "vpc-lattice"
        }
    ]
}
```

EventBridge À utiliser pour gérer les événements Auto Scaling

Amazon EventBridge, anciennement appelé CloudWatch Events, vous aide à configurer des règles basées sur les événements qui surveillent les ressources et initient des actions ciblées utilisant d'autres AWS services.

Les événements d'Amazon EC2 Auto Scaling sont transmis EventBridge en temps quasi réel. Vous pouvez établir des EventBridge règles qui invoquent des actions programmatiques et des notifications en réponse à divers événements de ce type. Par exemple, lorsque les instances sont en cours de lancement ou d'arrêt, vous pouvez appeler une AWS Lambda fonction pour exécuter une tâche préconfigurée.

Les cibles des EventBridge règles peuvent inclure AWS Lambda des fonctions, des rubriques Amazon SNS, des destinations d'API, des bus d'événements Comptes AWS, etc. Pour plus d'informations sur les cibles prises en charge, consultez EventBridge les cibles Amazon dans le guide de EventBridge l'utilisateur Amazon.

Commencez par créer des EventBridge règles avec un exemple à l'aide d'une rubrique Amazon SNS et d'une EventBridge règle. Ensuite, lorsqu'un utilisateur démarre une actualisation d'instance, Amazon SNS vous avertit par e-mail chaque fois qu'un point de contrôle est atteint. Pour plus d'informations, consultez Créez des EventBridge règles pour les événements d'actualisation, par exemple.

Table des matières

Référence de l'événement Amazon EC2 Auto Scaling

À l'aide d'Amazon EventBridge, vous pouvez créer des règles qui correspondent aux événements entrants et les acheminer vers des cibles à des fins de traitement.

Table des matières

Événements d'action de cycle de vie

Lorsque vous ajoutez des hooks de cycle de vie à votre groupe Auto Scaling, Amazon EC2 Auto Scaling envoie des événements EventBridge lorsqu'une instance passe en état d'attente. Les événements sont générés sur la base du meilleur effort.

Types d'événements

Action du cycle de vie de montée en puissance

L'événement suivant en exemple montre que Amazon EC2 Auto Scaling a déplacé une instance vers un état `Pending:Wait` en raison d'un hook de cycle de vie de lancement.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-launch Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken": "87654321-4321-4321-4321-210987654321",
    "AutoScalingGroupName": "my-asg",
    "LifecycleHookName": "my-lifecycle-hook",
    "EC2InstanceId": "i-1234567890abcdef0",
    "LifecycleTransition": "autoscaling:EC2_INSTANCE_LAUNCHING",
    "NotificationMetadata": "additional-info",
    "Origin": "EC2",
    "Destination": "AutoScalingGroup"
  }
}
```

Action du cycle de vie de mise à l'échelle horizontale

L'événement suivant en exemple montre que Amazon EC2 Auto Scaling a déplacé une instance vers un état `Terminating:Wait` en raison d'un hook de cycle de vie de résiliation.

> **Important**
>
> Lorsqu'un groupe Auto Scaling renvoie des instances vers un groupe chaud lors de la mise à l'échelle horizontale, ce renvoi peut également générer des événements `EC2 Instance-terminate Lifecycle Action`. Les événements générés lorsqu'une instance passe à l'état d'attente lors de mise à l'échelle horizontale indiquent `WarmPool` comme valeur de `Destination`. Pour plus d'informations, consultez Instance reuse policy.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-terminate Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken":"87654321-4321-4321-4321-210987654321",
    "AutoScalingGroupName":"my-asg",
    "LifecycleHookName":"my-lifecycle-hook",
    "EC2InstanceId":"i-1234567890abcdef0",
    "LifecycleTransition":"autoscaling:EC2_INSTANCE_TERMINATING",
    "NotificationMetadata":"additional-info",
    "Origin": "AutoScalingGroup",
    "Destination": "EC2"
  }
}
```

Événements réussis de mise à l'échelle

Les exemples suivants montrent les types d'événements en cas d'événements réussis de mise à l'échelle. Les événements sont générés sur la base du meilleur effort.

Types d'événements

- Événement réussi de montée en puissance
- Événement réussi de mise à l'échelle horizontale

Événement réussi de montée en puissance

L'événement suivant en exemple montre que Amazon EC2 Auto Scaling a lancé une instance avec succès.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
```

```
  "detail-type": "EC2 Instance Launch Successful",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn",
    "instance-arn"
  ],
  "detail": {
    "StatusCode": "InProgress",
    "Description": "Launching a new EC2 instance: i-12345678",
    "AutoScalingGroupName": "my-asg",
    "ActivityId": "87654321-4321-4321-4321-210987654321",
    "Details": {
      "Availability Zone": "us-west-2b",
      "Subnet ID": "subnet-12345678"
    },
    "RequestId": "12345678-1234-1234-1234-123456789012",
    "StatusMessage": "",
    "EndTime": "yyyy-mm-ddThh:mm:ssZ",
    "EC2InstanceId": "i-1234567890abcdef0",
    "StartTime": "yyyy-mm-ddThh:mm:ssZ",
    "Cause": "description-text",
    "Origin": "EC2",
    "Destination": "AutoScalingGroup"
  }
}
```

Événement réussi de mise à l'échelle horizontale

L'événement suivant en exemple montre que Amazon EC2 Auto Scaling a résilié une instance avec succès.

> **Important**
>
> Lorsqu'un groupe Auto Scaling renvoie des instances vers un groupe chaud lors de la mise à l'échelle horizontale, ce renvoi peut également générer des événements `EC2 Instance Terminate Successful`. Les événements générés lorsqu'une instance retourne avec succès dans le groupe chaud indiquent `WarmPool` comme valeur de `Destination`. Pour plus d'informations, consultez Instance reuse policy.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance Terminate Successful",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn",
    "instance-arn"
  ],
  "detail": {
    "StatusCode": "InProgress",
    "Description": "Terminating EC2 instance: i-12345678",
    "AutoScalingGroupName": "my-asg",
    "ActivityId": "87654321-4321-4321-4321-210987654321",
    "Details": {
      "Availability Zone": "us-west-2b",
      "Subnet ID": "subnet-12345678"
    },
    "RequestId": "12345678-1234-1234-1234-123456789012",
    "StatusMessage": "",
    "EndTime": "yyyy-mm-ddThh:mm:ssZ",
    "EC2InstanceId": "i-1234567890abcdef0",
    "StartTime": "yyyy-mm-ddThh:mm:ssZ",
    "Cause": "description-text",
    "Origin": "AutoScalingGroup",
    "Destination": "EC2"
  }
}
```

Échec des événements de mise à l'échelle

Les exemples suivants montrent les types d'événements en cas d'échec des événements de mise à l'échelle. Les événements sont générés sur la base du meilleur effort.

Types d'événements

- Échec de l'événement de montée en puissance
- Échec de l'événement de mise à l'échelle horizontale

Échec de l’événement de montée en puissance

L’événement suivant en exemple montre que Amazon EC2 Auto Scaling n’a pas pu lancer une instance.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance Launch Unsuccessful",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn",
    "instance-arn"
  ],
  "detail": {
    "StatusCode": "Failed",
    "AutoScalingGroupName": "my-asg",
    "ActivityId": "87654321-4321-4321-4321-210987654321",
    "Details": {
      "Availability Zone": "us-west-2b",
      "Subnet ID": "subnet-12345678"
    },
    "RequestId": "12345678-1234-1234-1234-123456789012",
    "StatusMessage": "message-text",
    "EndTime": "yyyy-mm-ddThh:mm:ssZ",
    "EC2InstanceId": "i-1234567890abcdef0",
    "StartTime": "yyyy-mm-ddThh:mm:ssZ",
    "Cause": "description-text",
    "Origin": "EC2",
    "Destination": "AutoScalingGroup"
  }
}
```

Échec de l’événement de mise à l’échelle horizontale

L’événement suivant en exemple montre que Amazon EC2 Auto Scaling n’a pas pu résilier une instance.

> **Important**
>
> Lorsqu'un groupe Auto Scaling renvoie des instances vers un groupe chaud lors de la mise à l'échelle horizontale, l'échec de ce renvoi peut également générer des événements EC2 `Instance Terminate Unsuccessful`. Les événements générés lorsqu'une instance ne parvient pas à retourner dans le groupe chaud indiquent `WarmPool` comme valeur de `Destination`. Pour plus d'informations, consultez Instance reuse policy.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance Terminate Unsuccessful",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn",
    "instance-arn"
  ],
  "detail": {
    "StatusCode": "Failed",
    "AutoScalingGroupName": "my-asg",
    "ActivityId": "87654321-4321-4321-4321-210987654321",
    "Details": {
      "Availability Zone": "us-west-2b",
      "Subnet ID": "subnet-12345678"
    },
    "RequestId": "12345678-1234-1234-1234-123456789012",
    "StatusMessage": "message-text",
    "EndTime": "yyyy-mm-ddThh:mm:ssZ",
    "EC2InstanceId": "i-1234567890abcdef0",
    "StartTime": "yyyy-mm-ddThh:mm:ssZ",
    "Cause": "description-text",
    "Origin": "AutoScalingGroup",
    "Destination": "EC2"
  }
}
```

Événements d’actualisation d’instance

Les exemples suivants illustrent des événements de la fonction d’actualisation des instances. Les événements sont générés sur la base du meilleur effort.

Types d’événements

- Point de contrôle atteint
- Début de l’actualisation de l’instance
- Actualisation de l’instance réussie
- Échec de l’actualisation de l’instance
- Annulation de l’actualisation de l’instance
- L'annulation de l'actualisation de l'instance a commencé
- Annulation de l'actualisation de l'instance réussie
- L'annulation de l'actualisation de l'instance a échoué

Point de contrôle atteint

Lorsque le nombre d’instances qui ont été remplacées atteint le seuil (en pourcentage) défini pour le point de contrôle, Amazon EC2 Auto Scaling envoie l’événement suivant.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Checkpoint Reached",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "ab00cf8f-9126-4f3c-8010-dbb8cad6fb86",
    "AutoScalingGroupName": "my-asg",
    "CheckpointPercentage": "50",
    "CheckpointDelay": "300"
  }
}
```

Début de l'actualisation de l'instance

Lorsque le statut d'une actualisation d'instance passe à `InProgress`, Amazon EC2 Auto Scaling envoie les événements suivants.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Started",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
}
```

Actualisation de l'instance réussie

Lorsque le statut d'une actualisation d'instance passe à `Successful`, Amazon EC2 Auto Scaling envoie les événements suivants.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Succeeded",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
```

```
}
```

Échec de l’actualisation de l’instance

Lorsque le statut d’une actualisation d’instance passe à `Failed`, Amazon EC2 Auto Scaling envoie les événements suivants.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Failed",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
}
```

Annulation de l’actualisation de l’instance

Lorsque le statut d’une actualisation d’instance passe à `Cancelled`, Amazon EC2 Auto Scaling envoie les événements suivants.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Cancelled",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
```

```
    "AutoScalingGroupName": "my-asg"
  }
}
```

L'annulation de l'actualisation de l'instance a commencé

Lorsque le statut d'une actualisation d'instance passe à `RollbackInProgress`, Amazon EC2 Auto Scaling envoie les événements suivants.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Rollback Started",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
}
```

Annulation de l'actualisation de l'instance réussie

Lorsque le statut d'une actualisation d'instance passe à `RollbackSuccessful`, Amazon EC2 Auto Scaling envoie les événements suivants.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Rollback Succeeded",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
```

```
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
}
```

L'annulation de l'actualisation de l'instance a échoué

Lorsque le statut d'une actualisation d'instance passe à `Failed`, Amazon EC2 Auto Scaling envoie les événements suivants.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Rollback Failed",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
}
```

Exemples d'événements et de modèles de groupe chaud

Amazon EC2 Auto Scaling prend en charge plusieurs modèles prédéfinis dans Amazon. EventBridge Cela simplifie la création d'un modèle d'événement. Vous sélectionnez les valeurs des champs dans un formulaire et vous EventBridge générez le modèle pour vous. Pour le moment, Amazon EC2 Auto Scaling ne prend pas en charge les modèles prédéfinis pour les événements émis par un groupe Auto Scaling doté d'un groupe d'instances pré-initialisées. Vous devez entrer le modèle en tant qu'objet JSON. Cette section et la rubrique Créez des EventBridge règles pour les événements en piscine chaude vous montre comment utiliser un modèle d'événements pour sélectionner des événements et les envoyer à des cibles.

Pour créer des EventBridge règles filtrant les événements liés au warm pool auxquels Amazon EC2 Auto Scaling envoie EventBridge, incluez `Origin` les champs `Destination` et de la section de `detail` l'événement.

Les valeurs de `Origin` et `Destination` peuvent être les suivantes :

`EC2 | AutoScalingGroup | WarmPool`

Table des matières

Exemples d’événements

Lorsque vous ajoutez des hooks de cycle de vie à votre groupe Auto Scaling, Amazon EC2 Auto Scaling envoie des événements EventBridge lorsqu'une instance passe en état d'attente. Pour plus d’informations, consultez Utiliser des hooks de cycle de vie avec un groupe d'instances pré-initialisées.

Cette section comprend des exemples de ces événements lorsque votre groupe Auto Scaling dispose d’un groupe chaud. Les événements sont générés sur la base du meilleur effort.

> **Note**
>
> Pour les événements auxquels Amazon EC2 Auto Scaling envoie une EventBridge fois le dimensionnement réussi, consultez. Événements réussis de mise à l’échelle Pour les événements où la mise à l’échelle échoue, consultez Échec des événements de mise à l’échelle.

Exemples d'événements

Action du cycle de vie de montée en puissance

Les événements générés lorsqu’une instance passe à un état d’attente pour des événements de montée en puissance indiquent `EC2 Instance-launch Lifecycle Action` comme valeur de `detail-type`. Dans l’objet `detail`, les valeurs des attributs `Origin` et `Destination` indiquent d’où vient l’instance et où elle va.

Dans cet exemple d'événement de montée en puissance, une nouvelle instance est lancée et son état passe à `Warmed:Pending:Wait` parce qu'elle est ajoutée au groupe chaud. Pour plus d'informations, consultez Transitions de l'état du cycle de vie pour les instances dans un groupe d'instances pré-initialisées.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-launch Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "2021-01-13T00:12:37.214Z",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken": "71514b9d-6a40-4b26-8523-05e7eEXAMPLE",
    "AutoScalingGroupName": "my-asg",
    "LifecycleHookName": "my-launch-lifecycle-hook",
    "EC2InstanceId": "i-1234567890abcdef0",
    "LifecycleTransition": "autoscaling:EC2_INSTANCE_LAUNCHING",
    "NotificationMetadata": "additional-info",
    "Origin": "EC2",
    "Destination": "WarmPool"
  }
}
```

Dans cet exemple d'événement de montée en puissance, l'état de l'instance passe à `Pending:Wait` parce qu'elle est ajoutée au groupe Auto Scaling depuis le groupe chaud. Pour plus d'informations, consultez Transitions de l'état du cycle de vie pour les instances dans un groupe d'instances pré-initialisées.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-launch Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "2021-01-19T00:35:52.359Z",
  "region": "us-west-2",
  "resources": [
```

```
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken": "19cc4d4a-e450-4d1c-b448-0de67EXAMPLE",
    "AutoScalingGroupName": "my-asg",
    "LifecycleHookName": "my-launch-lifecycle-hook",
    "EC2InstanceId": "i-1234567890abcdef0",
    "LifecycleTransition": "autoscaling:EC2_INSTANCE_LAUNCHING",
    "NotificationMetadata": "additional-info",
    "Origin": "WarmPool",
    "Destination": "AutoScalingGroup"
  }
}
```

Action du cycle de vie de mise à l'échelle horizontale

Les événements générés lorsqu'une instance passe à un état d'attente pour des événements de mise à l'échelle horizontale indiquent `EC2 Instance-terminate Lifecycle Action` comme valeur de `detail-type`. Dans l'objet `detail`, les valeurs des attributs `Origin` et `Destination` indiquent d'où vient l'instance et où elle va.

Dans cet exemple d'événement de type mise à l'échelle horizontale, l'état d'une instance passe à `Warmed:Pending:Wait` car elle est retournée au groupe chaud. Pour plus d'informations, consultez Transitions de l'état du cycle de vie pour les instances dans un groupe d'instances pré-initialisées.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-terminate Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "2022-03-28T00:12:37.214Z",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken": "42694b3d-4b70-6a62-8523-09a1eEXAMPLE",
    "AutoScalingGroupName": "my-asg",
    "LifecycleHookName": "my-termination-lifecycle-hook",
    "EC2InstanceId": "i-1234567890abcdef0",
    "LifecycleTransition": "autoscaling:EC2_INSTANCE_TERMINATING",
    "NotificationMetadata": "additional-info",
```

```
    "Origin": "AutoScalingGroup",
    "Destination": "WarmPool"
  }
}
```

Exemples de modèles d'événement

La section précédente fournit des exemples d'événements émis par Amazon EC2 Auto Scaling.

EventBridge les modèles d'événements ont la même structure que les événements auxquels ils correspondent. Le modèle place entre guillemets les champs que vous voulez faire correspondre et fournit les valeurs que vous recherchez.

Les champs suivants de l'événement constituent le modèle d'événement défini dans la règle permettant d'appeler une action :

`"source": "aws.autoscaling"`

Identifie que l'événement provient de Amazon EC2 Auto Scaling.

`"detail-type": "EC2 Instance-launch Lifecycle Action"`

Identifie le type d'événement.

`"Origin": "EC2"`

Identifie l'origine de l'instance.

`"Destination": "WarmPool"`

Identifie la destination de l'instance.

Utilisez l'exemple de modèle d'événement suivant pour capturer tous les événements `EC2 Instance-launch Lifecycle Action` associés aux instances qui intègrent le groupe chaud.

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-launch Lifecycle Action" ],
  "detail": {
      "Origin": [ "EC2" ],
      "Destination": [ "WarmPool" ]
   }
}
```

Utilisez l’exemple de modèle d’événement suivant pour capturer tous les événements EC2 `Instance-launch Lifecycle Action` associés aux instances qui sortent du groupe chaud en réponse à un événement de montée en puissance.

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-launch Lifecycle Action" ],
  "detail": {
      "Origin": [ "WarmPool" ],
      "Destination": [ "AutoScalingGroup" ]
   }
}
```

Utilisez l’exemple de modèle d’événement suivant pour capturer tous les événements EC2 `Instance-launch Lifecycle Action` associés aux instances lancées directement dans le groupe Auto Scaling.

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-launch Lifecycle Action" ],
  "detail": {
      "Origin": [ "EC2" ],
      "Destination": [ "AutoScalingGroup" ]
   }
}
```

Utilisez l’exemple de modèle d’événement suivant pour capturer tous les événements EC2 `Instance-terminate Lifecycle Action` associés aux instances renvoyées vers le groupe chaud mis à l’échelle horizontale.

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-terminate Lifecycle Action" ],
  "detail": {
      "Origin": [ "AutoScalingGroup" ],
      "Destination": [ "WarmPool" ]
   }
}
```

Utilisez l'exemple de modèle d'événement suivant pour capturer tous les événements associés à EC2 `Instance-launch Lifecycle Action`, quelle que soit l'origine ou la destination.

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-launch Lifecycle Action" ]
}
```

Créez des EventBridge règles

Lorsqu'un événement est émis par Amazon EC2 Auto Scaling, une notification d'événement est envoyée à Amazon EventBridge sous forme de fichier JSON. Vous pouvez écrire une EventBridge règle pour automatiser les actions à effectuer lorsqu'un modèle d'événement correspond à la règle. S'il EventBridge détecte un modèle d'événement correspondant à un modèle défini dans une règle, EventBridge invoque la cible (ou les cibles) spécifiée dans la règle.

Vous pouvez utiliser les procédures d'exemple de cette section comme point de départ.

Vous pourriez également trouver la documentation suivante utile.

- Pour exécuter des actions personnalisées sur les instances au moment de leur lancement ou avant qu'elles ne soient terminées à l'aide d'une fonction Lambda, consultez la rubrique Didacticiel : configurer un hook de cycle de vie qui appelle une fonction Lambda.
- Pour appeler une fonction Lambda sur les appels d'API enregistrés avec CloudTrail, consultez Tutoriel : journalisation des appels d' AWS API EventBridge à l'aide du guide de l' EventBridge utilisateur Amazon.
- Pour plus d'informations sur la création de règles relatives aux événements, consultez la section Création de EventBridge règles Amazon qui réagissent aux événements dans le guide de EventBridge l'utilisateur Amazon.

Rubriques

- Créez des EventBridge règles pour les événements d'actualisation, par exemple
- Créez des EventBridge règles pour les événements en piscine chaude

Créez des EventBridge règles pour les événements d'actualisation, par exemple

L'exemple suivant crée une EventBridge règle pour envoyer une notification par e-mail. Elle effectue cette opération chaque fois que votre groupe Auto Scaling émet un événement lorsqu'un point de contrôle est atteint pendant une actualisation d'instance. La procédure de configuration des notifications par e-mail à l'aide d'Amazon SNS est incluse. Pour utiliser Amazon SNS afin d'envoyer

des notifications par e-mail, vous devez d'abord créer une rubrique, puis abonner les adresses e-mail requises à cette rubrique.

Pour plus d'informations sur la fonction d'actualisation d'instance, consultez la rubrique Utiliser une actualisation d'instance pour mettre à jour les instances d'un groupe Auto Scaling.

Créer une rubrique Amazon SNS

Une rubrique SNS est un point d'accès logique, un canal de communication utilisé par le groupe Auto Scaling pour envoyer les notifications. Pour créer une rubrique, donnez-lui un nom.

Les noms de rubrique doivent respecter les critères suivants :

- Avoir 1 à 256 caractères
- Contenir des lettres majuscules et minuscules ASCII, des chiffres, des traits de soulignement ou de traits d'union

Pour plus d'informations, consultez Création d'une rubrique Amazon SNS dans le Guide du développeur Amazon Simple Notification Service.

Abonner à la rubrique Amazon SNS

Pour recevoir les notifications que votre groupe Auto Scaling envoie à la rubrique, vous devez abonner un point de terminaison à cette dernière. Dans cette procédure, sous Point de terminaison, spécifiez l'adresse e-mail à laquelle vous souhaitez recevoir les notifications envoyées par Amazon EC2 Auto Scaling.

Pour plus d'informations, consultez Abonnement à une rubrique Amazon SNS dans le Guide du développeur Amazon Simple Notification Service.

Confirmer votre abonnement Amazon SNS

Amazon SNS envoie un e-mail de confirmation à l'adresse que vous avez spécifiée à l'étape précédente.

Assurez-vous d'ouvrir l'e-mail depuis AWS Notifications et de choisir le lien pour confirmer l'abonnement avant de passer à l'étape suivante.

Vous recevrez un message d'accusé de réception de. AWS Amazon SNS est maintenant configuré pour recevoir des notifications et envoyer la notification par e-mail à l'adresse spécifiée.

Acheminer les événements vers votre rubrique Amazon SNS

Créez une règle qui correspond aux événements sélectionnés et les achemine vers votre rubrique Amazon SNS pour envoyer des notifications aux adresses e-mail abonnées.

Pour créer une règle qui envoie des notifications à votre rubrique Amazon SNS

1. Ouvrez la EventBridge console Amazon à l'adresse https://console.aws.amazon.com/events/.
2. Dans le volet de navigation, choisissez Règles.
3. Choisissez Créer une règle.
4. Pour Define rule detail (Définir les détails de la règle), procédez comme suit :
   a. Entrez un nom et éventuellement une description pour la règle.

      Une règle ne peut pas avoir le même nom qu'une autre règle de la même région et sur le même bus d'événement.
   b. Pour Event bus (Bus d'événement), choisissez default (défaut). Lorsqu'un AWS service de votre compte génère un événement, celui-ci est toujours redirigé vers le bus d'événements par défaut de votre compte.
   c. Pour Type de règle, choisissez Règle avec un modèle d'événement.
   d. Choisissez Suivant.
5. Pour Build event pattern (Créer un modèle d'événement), procédez comme suit :
   a. Dans Source de l'événement, choisissez AWS des événements ou des événements EventBridge partenaires.
   b. Pour Event pattern (Modèle d'événement), procédez comme suit :
      i. Pour Event source (Source d'événement), choisissez Services AWS.
      ii. Pour Service AWS, choisissez Auto Scaling.
      iii. Dans Type d'événement, choisissez Actualisation d'instance.
      iv. Par défaut, la règle correspond à tout événement d'actualisation d'instance. Pour créer une règle permettant d'envoyer une notification chaque fois qu'un point de contrôle est atteint pendant une actualisation d'instance, choisissez Specific instance event(s) (Événement[s] d'instance spécifique[s]) et sélectionnez EC2 Auto Scaling Instance Refresh Checkpoint Reached (Point de contrôle lié à une actualisation d'instance EC2 Auto Scaling atteint).

v. Par défaut, la règle correspond à tout groupe Auto Scaling de la région. Pour que la règle corresponde à un groupe Auto Scaling spécifique, choisissez Nom(s) de groupe spécifique(s), puis sélectionnez un ou plusieurs groupes Auto Scaling.

vi. Choisissez Next (Suivant).

6. Pour Select target(s) (Sélectionner la ou les cibles), procédez comme suit :

   a. Pour Target types (Types de cibles), choisissez Service AWS.

   b. Pour Select a target (Sélectionnez une cible), choisissez SNS Topic (Rubrique SNS).

   c. Pour Topic (Rubrique), choisissez votre rubrique Amazon SNS.

   d. (Facultatif) Sous Additional settings (Paramètres supplémentaires), vous pouvez configurer des paramètres supplémentaires. Pour plus d'informations, consultez la section Création de EventBridge règles Amazon qui réagissent aux événements dans le guide de EventBridge l'utilisateur Amazon.

   e. Choisissez Suivant.

7. (Facultatif) Pour Tags (Identifications), vous pouvez également attribuer une ou plusieurs identifications à votre règle, puis choisir Next (Suivant).

8. Pour Review and create (Vérifier et créer), examinez les détails de la règle et modifiez-les si nécessaire. Puis choisissez Create rule (Créer une règle).

Créez des EventBridge règles pour les événements en piscine chaude

L'exemple suivant crée une EventBridge règle pour invoquer des actions programmatiques. Elle effectue cette opération chaque fois que votre groupe Auto Scaling émet un événement lorsqu'une nouvelle instance est ajoutée au groupe d'instances pré-initialisées.

Avant de créer la règle, créez la AWS Lambda fonction que vous souhaitez que la règle utilise comme cible. Vous devez spécifier cette fonction comme étant la cible de la règle. La procédure suivante décrit uniquement les étapes permettant de créer la EventBridge règle qui agit lorsque de nouvelles instances entrent dans le pool de chaleur. Pour un tutoriel d'introduction qui vous montre comment créer une simple fonction Lambda à invoquer lorsqu'un événement entrant correspond à une règle, consultez la rubrique Didacticiel : configurer un hook de cycle de vie qui appelle une fonction Lambda.

Pour plus d'informations sur la création et l'utilisation de groupes d'instances pré-initialisées, consultez la rubrique Groupes d'instances pré-initialisées pour Amazon EC2 Auto Scaling.

Pour créer une règle d'événement qui invoque une fonction Lambda

1. Ouvrez la EventBridge console Amazon à l'adresse https://console.aws.amazon.com/events/.
2. Dans le volet de navigation, choisissez Règles.
3. Choisissez Créer une règle.
4. Pour Define rule detail (Définir les détails de la règle), procédez comme suit :

   a. Entrez un nom et éventuellement une description pour la règle.

      Une règle ne peut pas avoir le même nom qu'une autre règle de la même région et sur le même bus d'événement.

   b. Pour Event bus (Bus d'événement), choisissez default (défaut). Lorsqu'un événement est généré Service AWS dans votre compte, il est toujours redirigé vers le bus d'événements par défaut de votre compte.

   c. Pour Type de règle, choisissez Règle avec un modèle d'événement.

   d. Choisissez Suivant.
5. Pour Build event pattern (Créer un modèle d'événement), procédez comme suit :

   a. Dans Source de l'événement, choisissez AWS des événements ou des événements EventBridge partenaires.

   b. Pour Event pattern (Modèle d'événement), choisissez Custom pattern (JSON editor) (Modèle personnalisé [éditeur JSON]), puis collez le motif suivant dans le Event pattern (Modèle d'événement), en remplaçant le texte en *italique* par le nom de votre groupe Auto Scaling.

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-launch Lifecycle Action" ],
  "detail": {
      "AutoScalingGroupName": [ "my-asg" ],
      "Origin": [ "EC2" ],
      "Destination": [ "WarmPool" ]
   }
}
```

      Pour créer une règle qui correspond à d'autres événements, modifiez le modèle d'événement. Pour plus d'informations, consultez la rubrique Exemples de modèles d'événement.

c. Choisissez Next (Suivant).

6. Pour Select target(s) (Sélectionner la ou les cibles), procédez comme suit :

   a. Pour Target types (Types de cibles), choisissez Service AWS.

   b. Pour Select a target (Sélectionner une cible), choisissez Lambda Function (Fonction Lambda).

   c. Pour Function (Fonction), choisissez la fonction à laquelle vous souhaitez envoyer les événements.

   d. (Facultatif) Pour Configure version/alias (Configurer la version/l'alias), saisissez les paramètres de version et d'alias pour la fonction Lambda cible.

   e. (Facultatif) Pour Additional settings (Paramètres supplémentaires), saisissez tout paramètre supplémentaire approprié à votre application. Pour plus d'informations, consultez la section Création de EventBridge règles Amazon qui réagissent aux événements dans le guide de EventBridge l'utilisateur Amazon.

   f. Choisissez Suivant.

7. (Facultatif) Pour Tags (Identifications), vous pouvez également attribuer une ou plusieurs identifications à votre règle, puis choisir Next (Suivant).

8. Pour Review and create (Vérifier et créer), examinez les détails de la règle et modifiez-les si nécessaire. Puis choisissez Create rule (Créer une règle).

Fournir une connectivité réseau pour vos instances Auto Scaling à l'aide d'Amazon VPC

Amazon Virtual Private Cloud (Amazon VPC) est un service qui vous permet de lancer des AWS ressources telles que des groupes Auto Scaling dans un réseau virtuel isolé de manière logique que vous définissez.

Un sous-réseau dans Amazon VPC est une sous-division dans une zone de disponibilité définie par un segment de la plage d'adresses IP du VPC. Avec les sous-réseaux, vous pouvez regrouper les instances selon vos besoins sécuritaires et opérationnels. Un sous-réseau réside entièrement dans la zone de disponibilité dans laquelle il a été créé. Vous lancez des instances Auto Scaling dans les sous-réseaux.

Pour activer la communication entre Internet et les instances des sous-réseaux, vous devez créer une passerelle Internet et l'attacher à votre VPC. Une passerelle Internet permet aux ressources des

sous-réseaux de se connecter à Internet via le périmètre du réseau Amazon EC2. Si le trafic de votre sous-réseau est acheminé vers une passerelle Internet, le sous-réseau est reconnu comme un sous-réseau public. Si le trafic du sous-réseau n'est pas acheminé vers une passerelle Internet, le sous-réseau est reconnu comme un sous-réseau privé. Utilisez un sous-réseau public pour les ressources qui doivent être connectées à Internet et un sous-réseau privé pour les ressources qui ne doivent pas être connectées à Internet. Pour plus d'informations sur l'octroi d'un accès Internet aux instances dans un VPC, consultez Accès à Internet dans le Guide de l'utilisateur Amazon VPC.

Table des matières

VPC par défaut

Si vous avez créé votre groupe Compte AWS après le 4 décembre 2013 ou si vous créez votre groupe Auto Scaling dans un nouveau groupe Région AWS, nous créons un VPC par défaut pour vous. Le VPC par défaut s'accompagne d'un sous-réseau par défaut dans chaque zone de disponibilité. Si vous avez un VPC par défaut, le groupe Auto Scaling est créé dans le VPC par défaut.

Vous pouvez afficher vos VPC sur la page de vos VPC de la console Amazon VPC.

Pour plus d'informations sur le VPC par défaut, consultez la rubrique VPC par défaut dans le Guide de l'utilisateur Amazon VPC.

VPC personnalisé

Vous pouvez choisir de créer des VPC supplémentaires en accédant à la page du tableau de bord VPC dans la AWS Management Console et en sélectionnant Create VPC (Créer un VPC).

Pour de plus amples informations, consultez le Guide de l'utilisateur Amazon VPC.

> **Note**
>
> Un VPC couvre toutes les zones de disponibilité de l' Région AWS. Lorsque vous ajoutez des sous-réseaux à votre VPC, choisissez plusieurs zones de disponibilité pour vous assurer que les applications hébergées dans ces sous-réseaux sont hautement disponibles. Une zone de disponibilité est un ou plusieurs centres de données discrets dotés d'une alimentation, d'un réseau et d'une connectivité redondants dans une Région AWS. Les zones de disponibilité vous aident à rendre les applications de production hautement disponibles, tolérantes aux pannes et évolutives.

Considérations à prendre en compte lors du choix des sous-réseaux VPC

Notez les considérations à prendre en compte lors du choix des sous-réseaux VPC pour votre groupe Auto Scaling :

- Si vous associez un équilibreur de charge Elastic Load Balancing à votre groupe Auto Scaling, les instances peuvent être lancées dans des sous-réseaux publics ou privés. Cependant, l'équilibreur de charge doit être créé dans des sous-réseaux publics afin de prendre en charge la résolution DNS.
- Si vous accédez à vos instances Auto Scaling directement via SSH, les instances peuvent être lancées dans des sous-réseaux publics uniquement.
- Si vous accédez à des instances Auto Scaling sans entrée à l'aide de AWS Systems Manager Session Manager, les instances peuvent être lancées dans des sous-réseaux publics ou privés.
- Si vous utilisez des sous-réseaux privés, vous pouvez autoriser les instances Auto Scaling à accéder à Internet à l'aide d'une passerelle NAT publique.
- Par défaut, les sous-réseaux par défaut d'un VPC par défaut sont des sous-réseaux publics.

Adressage IP dans un VPC

Lorsque vous lancez des instances Auto Scaling dans un VPC, une adresse IP privée est automatiquement attribuée à vos instances à partir de la plage CIDR du sous-réseau dans lequel l'instance est lancée. Cela permet aux instances de communiquer avec d'autres instances dans le VPC.

Vous pouvez configurer un modèle de lancement ou une configuration du lancement pour affecter des adresses IPv4 publiques à vos instances. L'attribution d'adresses IP publiques à vos instances leur permet de communiquer avec Internet ou d'autres AWS services.

Lorsque vous lancez des instances dans un sous-réseau configuré pour attribuer automatiquement des adresses IPv6 aux instances, les instances reçoivent des adresses IPv4 et IPv6. Sinon, elles reçoivent uniquement des adresses IPv4. Pour plus d'informations, consultez la section Adresses IPv6 dans le Guide de l'utilisateur Amazon EC2.

Pour plus d'informations sur la spécification de plages CIDR pour votre VPC ou sous-réseau, consultez le Guide de l'utilisateur Amazon VPC.

Amazon EC2 Auto Scaling peut automatiquement attribuer des adresses IP privées supplémentaires lors du lancement de l'instance lorsque vous utilisez un modèle de lancement spécifiant des interfaces réseau supplémentaires. Une adresse IP privée unique est attribuée à chaque interface réseau à partir de la plage d'adresses CIDR du sous-réseau dans lequel l'instance est lancée. Dans ce cas, le système ne peut plus attribuer automatiquement d'adresse IPv4 publique à l'interface réseau principale. Vous ne pourrez pas vous connecter vos instances par une adresse IPv4 publique à moins d'associer des adresses IP élastiques disponibles aux instances à scalabilité automatique.

Interfaces réseau dans un VPC

Chaque instance d'un VPC a une interface réseau par défaut (appelée interface réseau principale). Vous ne pouvez pas détacher une interface réseau principale d’une instance. Vous pouvez créer et attacher une Network Interface supplémentaire dans n'importe quelle instance de votre VPC. Le nombre d'interfaces réseau que vous pouvez attacher varie en fonction du type d'instance.

Lors du lancement d'une instance à l'aide d'un modèle de lancement, vous pouvez spécifier des interfaces réseau supplémentaires. Toutefois, le lancement d'une instance Auto Scaling avec plusieurs interfaces réseau crée automatiquement chaque interface dans le même sous-réseau que l'instance. C'est parce qu'Amazon EC2 Auto Scaling ignore les sous-réseaux définis dans le modèle de lancement au profit de ce qui est spécifié dans le groupe Auto Scaling. Pour plus d'informations, consultez Création d'un modèle de lancement pour un groupe Auto Scaling.

Si vous créez ou attachez au moins deux interfaces réseau du même sous-réseau à une instance, vous risquez de rencontrer des problèmes de réseaux tels que le routage asymétrique, notamment sur les instances utilisant une variante de Linux non Amazon. Si vous avez besoin de ce type de configuration, vous devez configurer l'interface réseau secondaire dans le système d'exploitation. Par

exemple, consultez Comment puis-je faire fonctionner mon interface réseau secondaire dans mon instance Ubuntu EC2 ? dans le AWS Knowledge Center.

Location de placement de l'instance

Par défaut, toutes les instances du VPC s'exécutent comme des instances à location partagée. Amazon EC2 Auto Scaling prend également en charge les instances dédiées et les hôtes dédiés. Pour plus d'informations, consultez Créer un modèle de lancement à l'aide de paramètres avancés.

AWS Outposts

AWS Outposts étend un Amazon VPC d'une AWS région à un avant-poste avec les composants VPC accessibles dans la région, notamment les passerelles Internet, les passerelles privées virtuelles, les passerelles de transit Amazon VPC et les points de terminaison VPC. Un Outpost est hébergé dans une zone de disponibilité dans la Région et est une extension de cette zone de disponibilité que vous pouvez utiliser pour assurer la résilience.

Pour plus d'informations, consultez le Guide de l'utilisateur AWS Outposts.

Pour obtenir un exemple de déploiement d'un groupe Auto Scaling qui dessert le trafic d'un Application Load Balancer dans un Outpost, consultez l'article de blog suivant Configuration d'un Application Load Balancer sur AWS Outposts.

Ressources supplémentaires pour en savoir plus sur les VPC

Utilisez les rubriques suivantes pour en savoir plus sur les VPC et les sous-réseaux.

- Sous-réseaux privés dans un VPC
  - Exemple : VPC avec des serveurs dans des sous-réseaux privés et NAT
  - Passerelles NAT
- Sous-réseaux publics dans un VPC
  - Exemple : VPC pour un environnement de test
  - Exemple : VPC pour serveurs web et de base de données
- Sous-réseaux pour votre Application Load Balancer
  - Sous-réseaux pour votre équilibreur de charge
- Informations générales sur les VPC
  - Amazon VPC User Guide

- Connexion de VPC avec l’appairage de VPC
- Interfaces réseau Elastic
- Utiliser des points de terminaison d'un VPC pour la connectivité privée

Sécurité dans Amazon EC2 Auto Scaling

La sécurité du cloud AWS est la priorité absolue. En tant que AWS client, vous bénéficiez d'un centre de données et d'une architecture réseau conçus pour répondre aux exigences des entreprises les plus sensibles en matière de sécurité.

La sécurité est une responsabilité partagée entre vous AWS et vous. Le modèle de responsabilité partagée décrit cela comme la sécurité du cloud et la sécurité dans le cloud :

- Sécurité du cloud : AWS est chargée de protéger l'infrastructure qui exécute les AWS services dans le AWS cloud. AWS vous fournit également des services que vous pouvez utiliser en toute sécurité. Des auditeurs tiers testent et vérifient régulièrement l'efficacité de notre sécurité dans le AWS cadre des programmes de de). Pour en savoir plus sur les programmes de conformité qui s'appliquent à Amazon EC2 Auto Scaling, consultez la section AWS services concernés par programme de conformité et .
- Sécurité dans le cloud — Votre responsabilité est déterminée par le AWS service que vous utilisez. Vous êtes également responsable d'autres facteurs, y compris la sensibilité de vos données, les exigences de votre entreprise, ainsi que la législation et la réglementation applicables.

Cette documentation vous aide à comprendre comment appliquer le modèle de responsabilité partagée lors de l'utilisation de Amazon EC2 Auto Scaling. Les rubriques suivantes vous montrent comment configurer Amazon EC2 Auto Scaling pour répondre à vos objectifs de sécurité et de conformité. Vous apprendrez également à utiliser d'autres AWS services qui vous aident à surveiller et à sécuriser vos ressources Amazon EC2 Auto Scaling.

Rubriques

Sécurité de l’infrastructure dans Amazon EC2 Auto Scaling

En tant que service géré, Amazon EC2 Auto Scaling est protégé par la sécurité du réseau AWS mondial. Pour plus d'informations sur les services AWS de sécurité et sur la manière dont AWS l'infrastructure est protégée, consultez la section Sécurité du AWS cloud. Pour concevoir votre AWS environnement en utilisant les meilleures pratiques en matière de sécurité de l'infrastructure, consultez la section Protection de l'infrastructure dans le cadre AWS bien architecturé du pilier de sécurité.

Vous utilisez des appels d'API AWS publiés pour accéder à Amazon EC2 Auto Scaling via le réseau. Les clients doivent prendre en charge les éléments suivants :

- Protocole TLS (Transport Layer Security). Nous exigeons TLS 1.2 et recommandons TLS 1.3.
- Ses suites de chiffrement PFS (Perfect Forward Secrecy) comme DHE (Ephemeral Diffie-Hellman) ou ECDHE (Elliptic Curve Ephemeral Diffie-Hellman). La plupart des systèmes modernes tels que Java 7 et les versions ultérieures prennent en charge ces modes.

En outre, les demandes doivent être signées à l’aide d’un ID de clé d’accès et d’une clé d’accès secrète associée à un principal IAM. Vous pouvez également utiliser AWS Security Token Service (AWS STS) pour générer des informations d’identification de sécurité temporaires et signer les demandes.

Vous pouvez également utiliser un point de terminaison de cloud privé virtuel (VPC) pour Amazon EC2 Auto Scaling. Les points de terminaison d'un VPC d'interface permettent à vos ressources Amazon VPC d'utiliser leurs adresses IP privées pour accéder à Amazon EC2 Auto Scaling sans exposition à l'Internet public. Pour plus d’informations, consultez Amazon EC2 Auto Scaling et points de terminaison d'un VPC d'interface.

Ressources connexes

Pour plus d'informations sur les fonctionnalités permettant d'isoler le trafic de service fourni par Amazon EC2, consultez la section Sécurité de l'infrastructure dans Amazon EC2 dans le guide de l'utilisateur Amazon EC2.

Résilience dans Amazon EC2 Auto Scaling

L'infrastructure AWS mondiale est construite autour Régions AWS de zones de disponibilité. Régions AWS fournissent plusieurs zones de disponibilité physiquement séparées et isolées,

connectées par un réseau à faible latence, à haut débit et hautement redondant. Avec les zones de disponibilité, vous pouvez concevoir et exploiter des applications et des bases de données qui basculent automatiquement d'une zone à l'autre sans interruption. Les zones de disponibilité sont davantage disponibles, tolérantes aux pannes et ont une plus grande capacité de mise à l'échelle que les infrastructures traditionnelles à un ou plusieurs centres de données.

Pour plus d'informations sur les zones de disponibilité Régions AWS et les zones de disponibilité, consultez la section Infrastructure AWS globale.

Pour bénéficier de la redondance géographique du concept de zone de disponibilité, procédez comme suit :

- Répartissez votre groupe Auto Scaling sur plusieurs zones de disponibilité.
- Maintenez au moins une instance dans chaque zone de disponibilité.
- Attachez un équilibreur de charge pour distribuer le trafic entrant à travers les mêmes zones de disponibilité. Si vous utilisez un Application Load Balancer (équilibreur de charge), assurez-vous que chaque instance EC2 reçoit un volume de trafic similaire en maintenant activé l'équilibrage de charge entre zones. Cela permet de limiter l'impact d'une charge accrue sur les instances existantes lors d'un incident de basculement et améliore la résilience par rapport à l'absence d'équilibrage de charge entre zones.
- Assurez-vous que les contrôles d'état d'Elastic Load Balancing sont correctement configurés et qu'ils sont activés sur le groupe Auto Scaling. Ensuite, si une instance échoue à sa surveillance de l'état, Elastic Load Balancing arrête de lui envoyer du trafic et redirige le trafic vers des instances saines, tandis qu'Amazon EC2 Auto Scaling remplace l'instance défectueuse.

Amazon EC2 Auto Scaling vous aide à répondre aux besoins de résilience de vos applications Amazon EC2 Auto Scaling de la manière suivante :

- Il vérifie les instances pour détecter les problèmes d'état et d'accessibilité. Lorsque l'état d'une instance se dégrade, il la résilie et en lance une nouvelle.
- Si des politiques de mise à l'échelle dynamiques sont en vigueur, il adapte automatiquement la capacité en fonction du trafic entrant.
- Détecte les problèmes de fiabilité des CloudWatch métriques Amazon qui prennent en charge les politiques de dimensionnement et suspend les activités de scale-in lorsque des métriques fiables ne sont pas disponibles, par exemple lorsque des points de données sont manquants.
- Il tente de maintenir automatiquement un nombre équivalent d'instances dans chaque zone de disponibilité activée.

- Il utilise les zones de disponibilité pour maintenir une haute disponibilité. Lorsque l'état d'une zone de disponibilité se dégrade, Amazon EC2 Auto Scaling procède comme suit :
  - Il lance de nouvelles instances dans une zone de disponibilité différente activée pour votre groupe Auto Scaling.
  - Il redistribue les instances entre toutes les zones de disponibilité activées lorsque la zone de disponibilité défectueuse revient à un état sain.
- Il continue d'essayer de lancer des instances dans d'autres zones de disponibilité activées si une instance ne parvient pas à être lancée dans une zone de disponibilité donnée.
- Il enregistre et désenregistre automatiquement les instances auprès des équilibreurs de charge associés à votre groupe Auto Scaling. De cette façon, vous n'avez pas besoin d'enregistrer et de désenregistrer séparément les instances.

Ressources connexes

Pour plus d'informations sur les fonctionnalités destinées à répondre à vos besoins en matière de résilience des données fournies par Amazon EBS, consultez Resilience in Amazon Elastic Block Store dans le guide de l'utilisateur d'Amazon EBS.

Protection des données dans Amazon EC2 Auto Scaling

Le modèle de responsabilité AWS partagée Le modèle s'applique à la protection des données dans Amazon EC2 Auto Scaling. Comme décrit dans ce modèle, AWS est chargé de protéger l'infrastructure mondiale qui gère tous les AWS Cloud. La gestion du contrôle de votre contenu hébergé sur cette infrastructure relève de votre responsabilité. Vous êtes également responsable des tâches de configuration et de gestion de la sécurité des Services AWS que vous utilisez. Pour plus d'informations sur la confidentialité des données, consultez Questions fréquentes (FAQ) sur la confidentialité des données. Pour en savoir plus sur la protection des données en Europe, consultez le billet de blog Modèle de responsabilité partagée AWS et RGPD (Règlement général sur la protection des données) sur le Blog de sécuritéAWS .

À des fins de protection des données, nous vous recommandons de protéger les Compte AWS informations d'identification et de configurer les utilisateurs individuels avec AWS IAM Identity Center ou AWS Identity and Access Management (IAM). Ainsi, chaque utilisateur se voit attribuer uniquement les autorisations nécessaires pour exécuter ses tâches. Nous vous recommandons également de sécuriser vos données comme indiqué ci-dessous :

- Utilisez l’authentification multifactorielle (MFA) avec chaque compte.
- Utilisez le protocole SSL/TLS pour communiquer avec les ressources. AWS Nous exigeons TLS 1.2 et recommandons TLS 1.3.
- Configurez l'API et la journalisation de l'activité des utilisateurs avec AWS CloudTrail.
- Utilisez des solutions de AWS chiffrement, ainsi que tous les contrôles de sécurité par défaut qu'ils contiennent Services AWS.
- Utilisez des services de sécurité gérés avancés tels qu’Amazon Macie, qui contribuent à la découverte et à la sécurisation des données sensibles stockées dans Amazon S3.
- Si vous avez besoin de modules cryptographiques validés par la norme FIPS 140-2 pour accéder AWS via une interface de ligne de commande ou une API, utilisez un point de terminaison FIPS. Pour plus d’informations sur les points de terminaison FIPS (Federal Information Processing Standard) disponibles, consultez Federal Information Processing Standard (FIPS) 140-2 (Normes de traitement de l’information fédérale).

Nous vous recommandons fortement de ne jamais placer d’informations confidentielles ou sensibles, telles que les adresses e-mail de vos clients, dans des balises ou des champs de texte libre tels que le champ Name (Nom). Cela inclut lorsque vous travaillez avec Amazon EC2 Auto Scaling ou un autre outil à Services AWS l'aide de la console, de l'API ou AWS des AWS CLI SDK. Toutes les données que vous entrez dans des balises ou des champs de texte de forme libre utilisés pour les noms peuvent être utilisées à des fins de facturation ou dans les journaux de diagnostic. Si vous fournissez une adresse URL à un serveur externe, nous vous recommandons fortement de ne pas inclure d’informations d’identification dans l’adresse URL permettant de valider votre demande adressée à ce serveur.

Lorsque vous lancez une instance Amazon EC2, vous avez la possibilité de transmettre des données utilisateur à l'instance pour effectuer une configuration supplémentaire au démarrage de l'instance. Nous vous recommandons également de ne jamais inclure d’informations confidentielles ou sensibles dans les données utilisateur qui seront transmises à une instance.

AWS KMS keys À utiliser pour chiffrer les volumes Amazon EBS

Vous pouvez configurer votre groupe Auto Scaling pour chiffrer les données de volume Amazon EBS stockées dans le cloud avec AWS KMS keys. Amazon EC2 Auto Scaling prend en charge les clés AWS gérées et gérées par le client pour chiffrer vos données. Notez que l'option `KmsKeyId` permettant de spécifier une clé gérée par le client n'est pas disponible lorsque vous utilisez une configuration du lancement. Pour spécifier votre clé gérée par le client, utilisez plutôt un modèle de

lancement. Pour plus d'informations, consultez Créer un modèle de lancement pour un groupe Auto Scaling. Pour plus d'informations sur la création, le stockage et la gestion de vos clés de AWS KMS chiffrement, consultez le guide du AWS Key Management Service développeur.

Vous pouvez également configurer une clé gérée par le client dans votre AMI basée sur EBS avant de configurer le modèle ou la configuration du lancement, ou utiliser le chiffrement par défaut pour chiffrer les nouveaux volumes EBS et les copies d'instantané que vous créez. Pour plus d'informations, consultez Utiliser le chiffrement avec les AMI basées sur EBS dans le guide de l'utilisateur Amazon EC2 et Chiffrement par défaut dans le guide de l'utilisateur Amazon EBS.

> **Note**
>
> Pour plus d'informations sur la configuration de la stratégie de clé dont vous avez besoin pour lancer des instances Auto Scaling lorsque vous utilisez une clé gérée par le client pour le chiffrement, consultez Politique de AWS KMS clé requise pour une utilisation avec des volumes chiffrés.

Ressources connexes

Pour connaître les directives de protection des données fournies par Amazon EBS, consultez la section Protection des données dans Amazon Elastic Block Store dans le guide de l'utilisateur d'Amazon EBS.

Politique de AWS KMS clé requise pour une utilisation avec des volumes chiffrés

Amazon EC2 Auto Scaling utilise des rôles liés à un service pour déléguer des autorisations à d'autres personnes. Services AWS Les rôles liés au service Amazon EC2 Auto Scaling sont prédéfinis et incluent les autorisations dont Amazon EC2 Auto Scaling a besoin pour appeler d'autres personnes en votre nom. Services AWS Les autorisations prédéfinies incluent également l'accès à votre Clés gérées par AWS. Elles n'incluent pas toutefois l'accès à vos clés gérées par le client, ce qui vous permet de garder un contrôle total sur ces clés.

Cette rubrique explique comment configurer la politique de clé dont vous avez besoin pour lancer des instances Auto Scaling lorsque vous spécifiez une clé gérée par le client pour le chiffrement Amazon EBS.

> **Note**
>
> Amazon EC2 Auto Scaling n'a pas besoin d'une autorisation supplémentaire pour utiliser l' Clé gérée par AWS par défaut et protéger les volumes chiffrés de votre compte.

Table des matières

Présentation

Les éléments suivants AWS KMS keys peuvent être utilisés pour le chiffrement Amazon EBS lorsqu'Amazon EC2 Auto Scaling lance des instances :

- Clé gérée par AWS— Une clé de chiffrement dans votre compte créée, détenue et gérée par Amazon EBS. Il s'agit de la clé de chiffrement par défaut d'un nouveau compte. Le Clé gérée par AWS est utilisé pour le chiffrement, sauf si vous spécifiez une clé gérée par le client.
- Clé gérée par le client : clé de chiffrement personnalisée que vous créez, détenez et gérez. Pour plus d'informations, consultez Création des clés dans le Guide du développeur AWS Key Management Service .

  Remarque : la clé doit être symétrique. Amazon EBS ne prend pas en charge les clés asymétriques gérées par le client.

Vous pouvez configurer les clés gérées par le client lors de la création d'instantanés chiffrés ou d'un modèle de lancement qui spécifie les volumes chiffrés, ou de l'activation du chiffrement par défaut.

Configurer des politiques de clé

Vos clés KMS doivent être associées à une politique de clé qui permet à Amazon EC2 Auto Scaling de lancer des instances avec des volumes Amazon EBS chiffrés à l'aide d'une clé gérée par le client.

Utilisez les exemples de cette page pour configurer une politique de clé pour donner à Amazon EC2 Auto Scaling l'accès à votre clé gérée par le client. Vous pouvez modifier la politique de clé de la clé gérée par le client lors de la création de la clé ou ultérieurement.

Vous devez ajouter au moins deux déclarations de politique à la politique de clé de votre clé pour qu'elle fonctionne avec Amazon EC2 Auto Scaling.

- La première instruction permet à l'identité IAM spécifiée dans l'élément `Principal` d'utiliser directement la clé gérée par le client. Il inclut les autorisations permettant d'effectuer les `DescribeKey` opérations AWS KMS `Encrypt` `DecryptReEncrypt*`,`GenerateDataKey*`,, et sur la clé.
- La deuxième instruction permet à l'identité IAM spécifiée dans l'`Principal`élément d'utiliser l'`CreateGrant`opération pour générer des autorisations déléguant un sous-ensemble de ses propres autorisations à des entités intégrées à Services AWS AWS KMS ou à un autre principal. Il est ainsi possible d'utiliser la clé pour créer des ressources chiffrées en votre nom.

Lorsque vous ajoutez les nouvelles instructions de politique à votre politique de clé, ne changez pas les déclarations existantes dans la politique.

Pour chacun des exemples suivants, les arguments qui doivent être remplacés, tels qu'un identifiant de clé ou le nom d'un rôle lié à un service, sont affichés sous forme de texte *`d'espace réservé à l'utilisateur`*. Dans la plupart des cas, vous pouvez remplacer le nom du rôle lié à un service par le nom d'un rôle lié à un service Amazon EC2 Auto Scaling.

Pour plus d’informations, consultez les ressources suivantes :

- Pour créer une clé avec le AWS CLI, voir create-key.
- Pour mettre à jour une politique clé avec le AWS CLI, consultez put-key-policy.
- Pour trouver l’ID et l’Amazon Resource Name (ARN) d’une clé, consultez Recherche de l’ID et de l’ARN d’une clé dans le Guide du développeur AWS Key Management Service .
- Pour plus d'informations sur les rôles liés à un service Amazon EC2 Auto Scaling, consultez Rôles liés à un service pour Amazon EC2 Auto Scaling.
- Pour plus d'informations sur le chiffrement Amazon EBS et KMS en général, consultez le guide de l'utilisateur Amazon EBS et le guide du développeur sur le AWS Key Management Service chiffrement Amazon EBS.

Exemple 1 : sections de la politique de clé qui autorisent l'accès à la clé gérée par le client

Ajoutez les deux instructions de politique suivantes à la politique de clé de la clé gérée par le client, en remplaçant l'ARN par l'ARN du rôle lié à un service approprié qui est autorisé à accéder à la clé. Dans cet exemple, les sections de politique donnent au rôle lié au service nommé AWSServiceRoleForAutoScalingles autorisations d'utiliser la clé gérée par le client.

```
{
   "Sid": "Allow service-linked role use of the customer managed key",
   "Effect": "Allow",
   "Principal": {
       "AWS": [
           "arn:aws:iam::account-id:role/aws-service-role/
autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling"
       ]
   },
   "Action": [
       "kms:Encrypt",
       "kms:Decrypt",
       "kms:ReEncrypt*",
       "kms:GenerateDataKey*",
       "kms:DescribeKey"
   ],
   "Resource": "*"
}
```

```
{
   "Sid": "Allow attachment of persistent resources",
   "Effect": "Allow",
   "Principal": {
       "AWS": [
           "arn:aws:iam::account-id:role/aws-service-role/
autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling"
       ]
   },
   "Action": [
       "kms:CreateGrant"
   ],
   "Resource": "*",
   "Condition": {
       "Bool": {
```

```
            "kms:GrantIsForAWSResource": true
        }
    }
}
```

Exemple 2 : sections de la politique de clé autorisant l'accès entre comptes à la clé gérée par le client

Si votre clé gérée par le client se trouve dans un compte différent du groupe Auto Scaling, vous devez utiliser un octroi en combinaison avec la politique de clé pour autoriser l'accès intercompte à la clé.

Deux étapes doivent être effectuées dans l'ordre suivant :

1. Tout d'abord, ajoutez les deux énoncés de politique suivants à la politique clé de la clé gérée par le client. Remplacez l'exemple d'ARN par l'ARN de l'autre compte, en veillant à remplacer *111122223333* par l'ID de compte réel du compte dans Compte AWS lequel vous souhaitez créer le groupe Auto Scaling. Cela vous permet de donner à un utilisateur ou à un rôle IAM du compte spécifié l'autorisation de créer un octroi pour la clé à l'aide de la commande CLI suivante. Cependant, cela ne donne en soi aucun accès à la clé aux utilisateurs.

```
{
   "Sid": "Allow external account 111122223333 use of the customer managed key",
   "Effect": "Allow",
   "Principal": {
       "AWS": [
           "arn:aws:iam::111122223333:root"
       ]
   },
   "Action": [
       "kms:Encrypt",
       "kms:Decrypt",
       "kms:ReEncrypt*",
       "kms:GenerateDataKey*",
       "kms:DescribeKey"
   ],
   "Resource": "*"
}
```

```
{
```

```
    "Sid": "Allow attachment of persistent resources in external
 account 111122223333",
    "Effect": "Allow",
    "Principal": {
        "AWS": [
            "arn:aws:iam::111122223333:root"
        ]
    },
    "Action": [
        "kms:CreateGrant"
    ],
    "Resource": "*"
}
```

2. Ensuite, à partir du compte dans lequel vous voulez créer le groupe Auto Scaling, créez un octroi qui délègue les autorisations pertinentes au rôle approprié lié au service. L'élément `Grantee Principal` du bénéficiaire est l'ARN du rôle lié à un service approprié. L'`key-id` est l'ARN de la clé.

   *Voici un exemple de commande CLI create-grant qui autorise le rôle lié au service nommé **AWSServiceRoleForAutoScaling**dans le compte 111122223333 à utiliser la clé gérée par le client dans le compte 444455556666.*

```
aws kms create-grant \
  --region us-west-2 \
  --key-id arn:aws:kms:us-
west-2:444455556666:key/1a2b3c4d-5e6f-1a2b-3c4d-5e6f1a2b3c4d \
  --grantee-principal arn:aws:iam::111122223333:role/aws-service-role/
autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling \
  --operations "Encrypt" "Decrypt" "ReEncryptFrom" "ReEncryptTo" "GenerateDataKey"
 "GenerateDataKeyWithoutPlaintext" "DescribeKey" "CreateGrant"
```

   Pour que cette commande réussisse, l'utilisateur qui fait la demande doit avoir les autorisations pour l'action `CreateGrant`.

   L'exemple de politique IAM suivant permet à une identité IAM (utilisateur ou rôle) du compte *111122223333* de créer un octroi pour la clé gérée par le client du compte *444455556666*.

```
{
  "Version": "2012-10-17",
  "Statement": [
```

```
    {
      "Sid": "AllowCreationOfGrantForTheKMSKeyinExternalAccount444455556666",
      "Effect": "Allow",
      "Action": "kms:CreateGrant",
      "Resource": "arn:aws:kms:us-
west-2:444455556666:key/1a2b3c4d-5e6f-1a2b-3c4d-5e6f1a2b3c4d"
    }
  ]
}
```

Pour plus d'informations sur la création d'un octroi pour une clé KMS dans un autre Compte AWS, consultez la rubrique Octrois dans AWS KMS dans le Guide du développeur AWS Key Management Service .

> **Important**
>
> Le nom du rôle lié au service spécifié en tant que principal bénéficiaire doit être le nom d'un rôle existant. Après avoir créé l'autorisation, pour vous assurer qu'elle permet à Amazon EC2 Auto Scaling d'utiliser la clé KMS spécifiée, ne supprimez pas et ne recréez pas le rôle lié au service.

Modifier des politiques de clé dans la console AWS KMS

Les exemples des sections précédentes montrent comment ajouter des déclarations à une politique de clé, qui est simplement un moyen de modifier une politique de clé. Le moyen le plus simple de modifier une politique clé consiste à utiliser la vue par défaut de la AWS KMS console pour les politiques clés et à faire d'une identité IAM (utilisateur ou rôle) l'un des principaux utilisateurs de la stratégie clé appropriée. Pour plus d'informations, consultez la section Utilisation de l'affichage AWS Management Console par défaut dans le Guide du AWS Key Management Service développeur.

> **Important**
>
> Soyez prudent. Les déclarations de politique d'affichage par défaut de la console incluent les autorisations permettant d'effectuer AWS KMS `Revoke` des opérations sur la clé gérée par le client. Si vous accordez Compte AWS l'accès à une clé gérée par le client dans votre compte et que vous révoquez accidentellement l'autorisation qui leur a accordé cette autorisation, les

utilisateurs externes ne peuvent plus accéder à leurs données chiffrées ou à la clé utilisée pour chiffrer leurs données.

Identity and Access Management pour Amazon EC2 Auto Scaling

AWS Identity and Access Management (IAM) est un outil Service AWS qui permet à un administrateur de contrôler en toute sécurité l'accès aux AWS ressources. Des administrateurs IAM contrôlent les personnes qui peuvent être authentifiées(connectées) et autorisées (disposant d'autorisations) pour utiliser des ressources Amazon EC2 Auto Scaling. IAM est un Service AWS outil que vous pouvez utiliser sans frais supplémentaires.

Pour utiliser Amazon EC2 Auto Scaling, vous avez besoin d'un identifiant Compte AWS et de vos identifiants de sécurité pour vous connecter à votre compte. Pour plus d'informations, consultez les informations d'identification AWS de sécurité dans le guide de l'utilisateur IAM.

Pour une documentation IAM complète, consultez le Guide de l'utilisateur IAM.

Contrôle d’accès

Vous pouvez disposer d'informations d'identification valides pour authentifier vos demandes, mais à moins d'avoir des autorisations, vous ne pouvez pas créer de ressources Amazon EC2 Auto Scaling n'y accéder. Par exemple, vous devez disposer d’autorisations pour créer des groupes Auto Scaling, lancer des instances avec des modèles de lancement, etc.

Les sections suivantes fournissent des détails sur la façon dont un administrateur IAM peut utiliser IAM pour contribuer à sécuriser vos ressources Amazon EC2 Auto Scaling en contrôlant qui peut effectuer des actions Amazon EC2 Auto Scaling.

Nous vous recommandons de lire d'abord les rubriques Amazon EC2. Consultez la section Gestion des identités et des accès pour Amazon EC2 dans le guide de l'utilisateur Amazon EC2. Après avoir lu les rubriques de cette section, vous devriez avoir une bonne idée de ce que représentent les autorisations de contrôle d'accès à Amazon EC2 et de la façon dont elles peuvent s'intégrer à vos autorisations de ressources Amazon EC2 Auto Scaling.

Rubriques

- Fonctionnement d'Amazon EC2 Auto Scaling avec IAM

- Autorisations API Amazon EC2 Auto Scaling
- AWS politiques gérées pour Amazon EC2 Auto Scaling
- Rôles liés à un service pour Amazon EC2 Auto Scaling
- Exemples de politiques Amazon EC2 Auto Scaling basées sur l'identité
- Prévention du problème de l’adjoint confus entre services
- Support de modèle de lancement
- Rôle IAM pour les applications qui s'exécutent sur des instances Amazon EC2

Fonctionnement d'Amazon EC2 Auto Scaling avec IAM

Avant d'utiliser IAM pour gérer l'accès à Amazon EC2 Auto Scaling, apprenez-en davantage sur les fonctionnalités IAM qui peuvent être utilisées avec Amazon EC2 Auto Scaling.

Fonctionnalités IAM que vous pouvez utiliser avec Amazon EC2 Auto Scaling

Fonction IAM	Prise en charge d'Amazon EC2 Auto Scaling
Politiques basées sur l’identité	Oui
Politiques basées sur les ressources	Non
Actions de politique	Oui
Ressources de politique	Oui
Clés de condition de politique (spécifiques au service)	Oui
ACL	Non
ABAC (identifications dans les politiques)	Partielle
Informations d’identification temporaires	Oui
Fonctions de service	Oui
Rôles liés à un service	Oui

Pour obtenir une vue d'ensemble de la façon dont Amazon EC2 Auto Scaling et d'autres appareils Services AWS fonctionnent avec la plupart des fonctionnalités IAM, consultez Services AWS le guide de l'utilisateur d'IAM concernant leur compatibilité avec IAM.

Politiques basées sur l'identité pour Amazon EC2 Auto Scaling

Prend en charge les politiques basées sur l’identité	Oui

Les politiques basées sur l’identité sont des documents de politique d’autorisations JSON que vous pouvez attacher à une identité telle qu’un utilisateur, un Groupes d’utilisateurs IAM ou un rôle IAM. Ces politiques contrôlent quel type d’actions des utilisateurs et des rôles peuvent exécuter, sur quelles ressources et dans quelles conditions. Pour découvrir comment créer une politique basée sur l’identité, veuillez consulter Création de politiques IAM dans le Guide de l’utilisateur IAM.

Avec les politiques IAM basées sur l’identité, vous pouvez spécifier des actions et ressources autorisées ou refusées, ainsi que les conditions dans lesquelles les actions sont autorisées ou refusées. Vous ne pouvez pas spécifier le principal dans une politique basée sur une identité car celle-ci s’applique à l’utilisateur ou au rôle auquel elle est attachée. Pour découvrir tous les éléments que vous utilisez dans une politique JSON, consultez Références des éléments de politique JSON IAM dans le Guide de l’utilisateur IAM.

Politiques dans Amazon EC2 Auto Scaling basées sur une ressource

Prend en charge les politiques basées sur les ressources	Non

Les politiques basées sur les ressources sont des documents de politique JSON que vous attachez à une ressource. Des politiques basées sur les ressources sont, par exemple, les politiques de confiance de rôle IAM et des politiques de compartiment Amazon S3. Dans les services qui sont compatibles avec les politiques basées sur les ressources, les administrateurs de service peuvent les utiliser pour contrôler l’accès à une ressource spécifique. Pour la ressource dans laquelle se trouve la politique, cette dernière définit quel type d’actions un principal spécifié peut effectuer sur cette ressource et dans quelles conditions. Vous devez spécifier un principal dans une politique basée sur les ressources. Les principaux peuvent inclure des comptes, des utilisateurs, des rôles, des utilisateurs fédérés ou. Services AWS

Pour permettre un accès intercompte, vous pouvez spécifier un compte entier ou des entités IAM dans un autre compte en tant que principal dans une politique basée sur les ressources. L'ajout d'un principal entre comptes à une politique basée sur les ressources ne représente qu'une partie de l'instauration de la relation d'approbation. Lorsque le principal et la ressource sont différents Comptes AWS, un administrateur IAM du compte sécurisé doit également accorder à l'entité principale (utilisateur ou rôle) l'autorisation d'accéder à la ressource. Pour ce faire, il attache une politique basée sur une identité à l'entité. Toutefois, si une politique basée sur des ressources accorde l'accès à un principal dans le même compte, aucune autre politique basée sur l'identité n'est requise. Pour plus d'informations, consultez Différence entre les rôles IAM et les politiques basées sur une ressource dans le Guide de l'utilisateur IAM.

Actions de politique pour Amazon EC2 Auto Scaling

Prend en charge les actions de politique	Oui

Les administrateurs peuvent utiliser les politiques AWS JSON pour spécifier qui a accès à quoi. C'est-à-dire, quel principal peut effectuer des actions sur quelles ressources et dans quelles conditions.

L'élément `Action` d'une politique JSON décrit les actions que vous pouvez utiliser pour autoriser ou refuser l'accès à une politique. Les actions de stratégie portent généralement le même nom que l'opération AWS d'API associée. Il existe quelques exceptions, telles que les actions avec autorisations uniquement qui n'ont pas d'opération API correspondante. Certaines opérations nécessitent également plusieurs actions dans une politique. Ces actions supplémentaires sont nommées actions dépendantes.

Intégration d'actions dans une stratégie afin d'accorder l'autorisation d'exécuter les opérations associées.

Pour accéder à une liste d'actions Amazon EC2 Auto Scaling, consultez Action définies par Amazon EC2 Auto Scaling dans la Référence d'autorisation.

Les actions de politique dans Amazon EC2 Auto Scaling utilisent le préfixe suivant avant l'action :

```
autoscaling
```

Pour indiquer plusieurs actions dans une seule déclaration, séparez-les par des virgules.

```
"Action": [
  "autoscaling:action1",
  "autoscaling:action2"
]
```

Vous pouvez aussi spécifier plusieurs actions à l'aide de caractères génériques (*). Par exemple, pour spécifier toutes les actions qui commencent par le mot `Describe`, incluez l'action suivante :

```
"Action": "autoscaling:Describe*"
```

Ressources sur les politiques pour Amazon EC2 Auto Scaling

Prend en charge les ressources de politique	Oui

Les administrateurs peuvent utiliser les politiques AWS JSON pour spécifier qui a accès à quoi. C'est-à-dire, quel principal peut effectuer des actions sur quelles ressources et dans quelles conditions.

L'élément de politique JSON `Resource` indique le ou les objets pour lesquels l'action s'applique. Les instructions doivent inclure un élément `Resource` ou `NotResource`. Il est recommandé de définir une ressource à l'aide de son Amazon Resource Name (ARN). Vous pouvez le faire pour des actions qui prennent en charge un type de ressource spécifique, connu sous la dénomination autorisations de niveau ressource.

Pour les actions qui ne sont pas compatibles avec les autorisations de niveau ressource, telles que les opérations de liste, utilisez un caractère générique (*) afin d'indiquer que l'instruction s'applique à toutes les ressources.

```
"Resource": "*"
```

Vous pouvez utiliser les ARN pour identifier les groupes Auto Scaling et lancer des configurations auxquelles la politique IAM s'applique.

Un groupe Auto Scaling comprend l'ARN suivant.

```
"Resource": "arn:aws:autoscaling:region:account-
id:autoScalingGroup:uuid:autoScalingGroupName/asg-name"
```

Une configuration du lancement comprend l'ARN suivant.

```
"Resource": "arn:aws:autoscaling:region:account-
id:launchConfiguration:uuid:launchConfigurationName/lc-name"
```

Pour spécifier un groupe Auto Scaling avec l'action `CreateAutoScalingGroup`, vous devez remplacer l'UUID par un caractère générique (*) comme suit :

```
"Resource": "arn:aws:autoscaling:region:account-
id:autoScalingGroup:*:autoScalingGroupName/asg-name"
```

Pour spécifier une configuration de lancement avec l'action `CreateLaunchConfiguration`, vous devez remplacer l'UUID par un caractère générique (*) comme suit :

```
"Resource": "arn:aws:autoscaling:region:account-
id:launchConfiguration:*:launchConfigurationName/lc-name"
```

Pour de plus d'informations sur les types de ressources Amazon EC2 Auto Scaling, veuillez consulter Ressources définies par Amazon EC2 Auto Scaling dans la Référence de l'autorisation de service. Pour savoir les actions avec lesquelles vous pouvez spécifier l'ARN de chaque ressource, consultez Actions définies par Amazon EC2 Auto Scaling.

> **Note**
>
> Pour obtenir un exemple de politique IAM qui utilise des ARN afin de contrôler l'accès aux groupes Auto Scaling, consultez Contrôler les groupes Auto Scaling pouvant être supprimés.

Certaines actions Amazon EC2 Auto Scaling ne prennent pas en charge les autorisations de niveau ressource. Pour les actions qui ne prennent pas en charge les autorisations au niveau de la ressource, vous devez utiliser (*) comme ressource.

Les actions Amazon EC2 Auto Scaling suivantes ne prennent pas en charge les autorisations de niveau ressource :

- `DescribeAccountLimits`
- `DescribeAdjustmentTypes`

- `DescribeAutoScalingGroups`
- `DescribeAutoScalingInstances`
- `DescribeAutoScalingNotificationTypes`
- `DescribeInstanceRefreshes`
- `DescribeLaunchConfigurations`
- `DescribeLifecycleHooks`
- `DescribeLifecycleHookTypes`
- `DescribeLoadBalancers`
- `DescribeLoadBalancerTargetGroups`
- `DescribeMetricCollectionTypes`
- `DescribeNotificationConfigurations`
- `DescribePolicies`
- `DescribeScalingActivities`
- `DescribeScalingProcessTypes`
- `DescribeScheduledActions`
- `DescribeTags`
- `DescribeTerminationPolicyTypes`
- `DescribeWarmPool`

Politique des clés de condition pour Amazon EC2 Auto Scaling

Prend en charge les clés de condition de politique spécifiques au service	Oui

Les administrateurs peuvent utiliser les politiques AWS JSON pour spécifier qui a accès à quoi. C'est-à-dire, quel principal peut effectuer des actions sur quelles ressources et dans quelles conditions.

L'élément `Condition` (ou le bloc `Condition`) vous permet de spécifier des conditions lorsqu'une instruction est appliquée. L'élément `Condition` est facultatif. Vous pouvez créer des expressions conditionnelles qui utilisent des opérateurs de condition, tels que les signes égal ou inférieur à, pour faire correspondre la condition de la politique aux valeurs de la demande.

Si vous spécifiez plusieurs éléments `Condition` dans une instruction, ou plusieurs clés dans un seul élément `Condition`, AWS les évalue à l'aide d'une opération `AND` logique. Si vous spécifiez plusieurs valeurs pour une seule clé de condition, AWS évalue la condition à l'aide d'une `OR` opération logique. Toutes les conditions doivent être remplies avant que les autorisations associées à l'instruction ne soient accordées.

Vous pouvez aussi utiliser des variables d'espace réservé quand vous spécifiez des conditions. Par exemple, vous pouvez accorder à un utilisateur IAM l'autorisation d'accéder à une ressource uniquement si elle est balisée avec son nom d'utilisateur IAM. Pour plus d'informations, consultez Éléments d'une politique IAM : variables et identifications dans le Guide de l'utilisateur IAM.

AWS prend en charge les clés de condition globales et les clés de condition spécifiques au service. Pour voir toutes les clés de condition AWS globales, voir les clés de contexte de condition AWS globales dans le guide de l'utilisateur IAM.

Amazon EC2 Auto Scaling prend en charge les clés de condition suivantes que vous pouvez utiliser pour contrôler l'accès aux actions assurées et appliquer la configuration des groupes Auto Scaling :

- `autoscaling:InstanceTypes`
- `autoscaling:LaunchConfigurationName`
- `autoscaling:LaunchTemplateVersionSpecified`
- `autoscaling:LoadBalancerNames`
- `autoscaling:MaxSize`
- `autoscaling:MinSize`
- `autoscaling:ResourceTag/key-name: tag-value`
- `autoscaling:TargetGroupARNs`
- `autoscaling:VPCZoneIdentifiers`

Les clés de condition suivantes sont spécifiques à la création de demandes de configuration de lancement :

- `autoscaling:ImageId`
- `autoscaling:InstanceType`
- `autoscaling:MetadataHttpEndpoint`
- `autoscaling:MetadataHttpPutResponseHopLimit`
- `autoscaling:MetadataHttpTokens`

- `autoscaling:SpotPrice`

Amazon EC2 Auto Scaling prend également en charge les clés de condition globales suivantes que vous pouvez utiliser pour définir des autorisations en fonction des balises figurant dans la demande ou présentes dans le groupe Auto Scaling. Pour plus d'informations, consultez Baliser des groupes et des instances Auto Scaling.

- `aws:RequestTag/key-name: tag-value`
- `aws:ResourceTag/key-name: tag-value`
- `aws:TagKeys: [tag-key, ...]`

Pour savoir avec quelles actions de l'API Amazon EC2 Auto Scaling vous pouvez utiliser une clé de condition, consultez Action définies par Amazon EC2 Auto Scaling dans la Référence d'autorisation. Pour de plus amples informations sur les clés de condition Amazon EC2 Auto Scaling, consultez Clés de condition pour Amazon EC2 Auto Scaling.

> **Note**
>
> Pour obtenir des exemples de politiques IAM utilisant des clés de condition pour contrôler l'accès aux actions prises en charge et appliquer la configuration des groupes Auto Scaling, consultez les ressources suivantes :
>
> - Demander un modèle de lancement et un numéro de version— Cet exemple indique qu'un modèle de lancement et le numéro de version du modèle de lancement doivent être spécifiés lors de la création ou de la mise à jour de groupes Auto Scaling.
> - Contrôler la taille des groupes Auto Scaling qui peuvent être créés— Cet exemple impose des contraintes sur les valeurs possibles pour les `MaxSize` propriétés `MinSize` et lors de la création ou de la mise à jour de groupes Auto Scaling avec une balise spécifique.
> - Contrôler les politiques de mise à l'échelle pouvant être supprimées— Cet exemple indique que la suppression des politiques de dimensionnement n'est autorisée que pour les groupes Auto Scaling sans balise spécifique.

ACLs dans Amazon EC2 Auto Scaling

Prend en charge les listes ACL	Non

Les listes de contrôle d'accès (ACL) vérifient quels principaux (membres de compte, utilisateurs ou rôles) ont l'autorisation d'accéder à une ressource. Les listes de contrôle d'accès sont similaires aux politiques basées sur les ressources, bien qu'elles n'utilisent pas le format de document de politique JSON.

ABAC avec Amazon EC2 Auto Scaling

Prise en charge d'ABAC (identifications dans les politiques)	Partielle

Le contrôle d'accès basé sur les attributs (ABAC) est une politique d'autorisation qui définit des autorisations en fonction des attributs. Dans AWS, ces attributs sont appelés balises. Vous pouvez associer des balises aux entités IAM (utilisateurs ou rôles) et à de nombreuses AWS ressources. L'étiquetage des entités et des ressources est la première étape d'ABAC. Vous concevez ensuite des politiques ABAC pour autoriser des opérations quand l'identification du principal correspond à celle de la ressource à laquelle il tente d'accéder.

L'ABAC est utile dans les environnements qui connaissent une croissance rapide et pour les cas où la gestion des politiques devient fastidieuse.

Pour contrôler l'accès basé sur des étiquettes, vous devez fournir les informations d'étiquette dans l'élément de condition d'une politique utilisant les clés de condition `aws:ResourceTag/key-name`, `aws:RequestTag/key-name` ou `aws:TagKeys`.

Si un service prend en charge les trois clés de condition pour tous les types de ressources, alors la valeur pour ce service est Oui. Si un service prend en charge les trois clés de condition pour certains types de ressources uniquement, la valeur est Partielle.

Pour plus d'informations sur l'ABAC, consultez Qu'est-ce que le contrôle d'accès basé sur les attributs (ABAC) ? dans le Guide de l'utilisateur IAM. Pour accéder à un didacticiel décrivant les étapes de configuration de l'ABAC, consultez Utilisation du contrôle d'accès basé sur les attributs (ABAC) dans le Guide de l'utilisateur IAM.

ABAC est possible pour les ressources qui prennent en charge les balises. Toutefois, toutes les ressources ne prennent pas en charge les balises. Les configurations de lancement et les politiques de mise à l'échelle ne prennent pas en charge les balises. Cependant, les groupes Auto Scaling le font.

Pour plus d'informations, consultez Baliser des groupes et des instances Auto Scaling.

Utilisation d'informations d'identification temporaires avec Amazon EC2 Auto Scaling

Prend en charge les informations d’identification temporaires	Oui

Certains Services AWS ne fonctionnent pas lorsque vous vous connectez à l'aide d'informations d'identification temporaires. Pour plus d'informations, y compris celles qui Services AWS fonctionnent avec des informations d'identification temporaires, consultez Services AWS la section relative à l'utilisation d'IAM dans le guide de l'utilisateur d'IAM.

Vous utilisez des informations d'identification temporaires si vous vous connectez à l' AWS Management Console aide d'une méthode autre qu'un nom d'utilisateur et un mot de passe. Par exemple, lorsque vous accédez à AWS l'aide du lien d'authentification unique (SSO) de votre entreprise, ce processus crée automatiquement des informations d'identification temporaires. Vous créez également automatiquement des informations d’identification temporaires lorsque vous vous connectez à la console en tant qu’utilisateur, puis changez de rôle. Pour plus d’informations sur le changement de rôle, consultez Changement de rôle (console) dans le Guide de l’utilisateur IAM.

Vous pouvez créer manuellement des informations d'identification temporaires à l'aide de l' AWS API AWS CLI or. Vous pouvez ensuite utiliser ces informations d'identification temporaires pour y accéder AWS. AWS recommande de générer dynamiquement des informations d'identification temporaires au lieu d'utiliser des clés d'accès à long terme. Pour plus d’informations, consultez Informations d’identification de sécurité temporaires dans IAM.

Fonctions du service pour Amazon EC2 Auto Scaling

Prend en charge les fonctions de service	Oui

Une fonction du service est un rôle IAM qu’un service endosse pour accomplir des actions en votre nom. Un administrateur IAM peut créer, modifier et supprimer une fonction du service à partir d’IAM. Pour plus d’informations, consultez Création d’un rôle pour la délégation d’autorisations à un Service AWS dans le Guide de l’utilisateur IAM.

Lorsque vous créez un hook de cycle de vie qui informe une rubrique Amazon SNS ou une file d'attente Amazon SQS, vous devez spécifier une fonction afin de d'autoriser Amazon EC2 Auto Scaling à accéder Amazon SNS ou Amazon SQS en votre nom. Utilisez la console IAM pour

configurer la fonction de service pour votre hook de cycle de vie. La console vous aide à créer une fonction avec un ensemble d'autorisations suffisantes, à l'aide d'une stratégie gérée. Pour plus d'informations, consultez Recevoir des notifications à l'aide d'Amazon SNS et Recevoir des notifications à l'aide d'Amazon SQS.

Lorsque vous créez un groupe Auto Scaling, vous pouvez éventuellement transmettre un rôle de service pour permettre aux instances Amazon EC2 d'accéder à d'autres en votre Services AWS nom. La fonction de service pour les instances Amazon EC2 (également appelé profil d'instance Amazon EC2 pour un modèle de lancement ou une configuration de lancement) est un type de fonction de service unique assigné à chaque instance EC2 dans un groupe Auto Scaling lors du lancement de l'instance. Vous pouvez utiliser la console IAM AWS CLI pour créer ou modifier ce rôle de service. Pour plus d'informations, consultez Rôle IAM pour les applications qui s'exécutent sur des instances Amazon EC2.

> **Warning**
>
> La modification des autorisations d'une fonction de service peut porter préjudice à la fonctionnalité Amazon EC2 Auto Scaling. Ne modifiez des rôles de service que lorsque Amazon EC2 Auto Scaling vous le recommande.

Rôles liés à un service pour Amazon EC2 Auto Scaling

Prend en charge les rôles liés à un service.	Oui

Un rôle lié à un service est un type de rôle de service lié à un. Service AWS Le service peut endosser le rôle afin d'effectuer une action en votre nom. Les rôles liés au service apparaissent dans votre Compte AWS fichier et appartiennent au service. Un administrateur IAM peut consulter, mais ne peut pas modifier, les autorisations concernant les rôles liés à un service.

Pour plus d'informations sur la création ou la gestion des rôles liés à un service Amazon EC2 Auto Scaling, consultez Rôles liés à un service pour Amazon EC2 Auto Scaling.

Autorisations API Amazon EC2 Auto Scaling

Vous devez autoriser les utilisateurs IAM à appeler les actions de l'API Amazon EC2 Auto Scaling dont ils ont besoin, comme décrit dans Actions de politique pour Amazon EC2 Auto Scaling. En outre,

pour certaines actions Amazon EC2 Auto Scaling, vous devez autoriser les utilisateurs à appeler des actions spécifiques depuis d'autres AWS API.

Autorisations requises provenant d’autres API AWS

Outre les autorisations d'API Amazon EC2 Auto Scaling, les utilisateurs doivent disposer des autorisations suivantes provenant d'autres AWS API pour effectuer correctement l'action associée.

Créer un groupe Auto Scaling (`autoscaling:CreateAutoScalingGroup`)

- `iam:CreateServiceLinkedRole`— Pour créer le rôle lié au service par défaut s'il n'existe pas encore.
- `iam:PassRole`— Pour transmettre un rôle IAM au service ou aux instances EC2 lors du lancement. Nécessaire lorsqu’un rôle lié à un service autre que celui par défaut, un rôle IAM pour un hook de cycle de vie ou un modèle de lancement spécifiant un profil d’instance (un conteneur pour un rôle IAM) est fourni.
- `ec2:RunInstances`— Pour lancer des instances lorsqu'un modèle de lancement est fourni.
- `ec2:CreateTags`— Pour baliser les instances et les volumes lors du lancement lorsqu'un modèle de lancement avec une spécification de balise est fourni.

Créer un hook de cycle de vie (`autoscaling:PutLifecycleHook`)

- `iam:PassRole`— Pour transmettre un rôle IAM au service. Nécessaire lorsqu’un rôle IAM est fourni.

Associer un groupe cible VPC Lattice (`autoscaling:AttachTrafficSources`)

- `vpc-lattice:RegisterTargets`— Pour enregistrer automatiquement les instances auprès du groupe cible.

Détacher un groupe cible en VPC Lattice (`autoscaling:DetachTrafficSources`)

- `vpc-lattice:DeregisterTargets`— Pour désenregistrer automatiquement les instances auprès du groupe cible.

Créez une configuration du lancement (`autoscaling:CreateLaunchConfiguration`)

- `ec2:DescribeImages`
- `ec2:DescribeInstances`
- `ec2:DescribeInstanceAttribute`
- `ec2:DescribeKeyPairs`
- `ec2:DescribeSecurityGroups`

- `ec2:DescribeSpotInstanceRequests`
- `ec2:DescribeVpcClassicLink`
- `iam:PassRole`— Pour transmettre un rôle IAM aux instances EC2 lors du lancement. Nécessaire lorsqu'une configuration de lancement spécifie un profil d'instance (un conteneur pour un rôle IAM).

AWS politiques gérées pour Amazon EC2 Auto Scaling

Une politique AWS gérée est une politique autonome créée et administrée par AWS. AWS les politiques gérées sont conçues pour fournir des autorisations pour de nombreux cas d'utilisation courants afin que vous puissiez commencer à attribuer des autorisations aux utilisateurs, aux groupes et aux rôles.

N'oubliez pas que les politiques AWS gérées peuvent ne pas accorder d'autorisations de moindre privilège pour vos cas d'utilisation spécifiques, car elles sont accessibles à tous les AWS clients. Nous vous recommandons de réduire encore les autorisations en définissant des politiques gérées par le client qui sont propres à vos cas d'utilisation.

Vous ne pouvez pas modifier les autorisations définies dans les politiques AWS gérées. Si les autorisations définies dans une politique AWS gérée sont AWS mises à jour, la mise à jour affecte toutes les identités principales (utilisateurs, groupes et rôles) auxquelles la politique est attachée. AWS est le plus susceptible de mettre à jour une politique AWS gérée lorsqu'une nouvelle politique Service AWS est lancée ou lorsque de nouvelles opérations d'API sont disponibles pour les services existants.

Pour plus d'informations, consultez la section Politiques gérées par AWS dans le Guide de l'utilisateur IAM.

Politiques gérées par Amazon EC2 Auto Scaling

Vous pouvez associer les politiques gérées suivantes à vos identités AWS Identity and Access Management (IAM) (utilisateurs ou rôles). Chaque politique accorde l'accès à tout ou partie des actions d'API pour Amazon EC2 Auto Scaling.

- AutoScalingConsoleFullAccès — Accorde un accès complet à Amazon EC2 Auto Scaling à l'aide du. AWS Management Console Cette politique fonctionne lorsque vous utilisez des configurations de lancement, mais pas avec des modèles de lancement.

- AutoScalingConsoleReadOnlyAccess— Accorde un accès en lecture seule à Amazon EC2 Auto Scaling à l'aide du. AWS Management Console Cette politique fonctionne lorsque vous utilisez des configurations de lancement, mais pas avec des modèles de lancement.
- AutoScalingFullAccess— Accorde un accès complet à Amazon EC2 Auto Scaling pour les identités IAM qui ont besoin d'un accès complet à Amazon EC2 Auto Scaling depuis AWS CLI le ou les SDK, mais pas d'accès. AWS Management Console
- AutoScalingReadOnlyAccès : accorde un accès en lecture seule à Amazon EC2 Auto Scaling pour les identités IAM qui appellent uniquement AWS CLI les SDK.

Lorsque vous utilisez des modèles de lancement à partir de la console, vous devez appliquer des autorisations supplémentaires spécifiques aux modèles de lancement, abordés dans Support de modèle de lancement. La console Amazon EC2 Auto Scaling a besoin d'autorisations pour des actions `ec2` afin de pouvoir afficher des informations sur les modèles de lancement et les instances de lancement à l'aide de modèles de lancement.

AutoScalingServiceRoleStratégie AWS gérée par des politiques

Cette politique est associée à un rôle lié à un service qui permet à Amazon EC2 Auto Scaling d'effectuer des actions en votre nom. Pour plus d'informations, consultez Rôles liés à un service pour Amazon EC2 Auto Scaling.

Pour consulter les autorisations associées à cette politique, consultez la section AutoScalingServiceRolePolitique dans la référence des stratégies AWS gérées.

Amazon EC2 Auto Scaling met à jour les politiques gérées AWS

Consultez les informations relatives aux mises à jour des politiques AWS gérées pour Amazon EC2 Auto Scaling depuis que ce service a commencé à suivre ces modifications. Pour recevoir des alertes automatiques sur les modifications apportées à cette page, abonnez-vous au flux RSS sur la page de l'historique des documents Amazon EC2 Auto Scaling.

Modification	Description	Date
Amazon EC2 Auto Scaling ajoute des autorisations à son rôle lié au service	La `AutoScalingServiceRolePolicy` politique accorde désormais l'autorisation d'appeler l'action d'API	29 février 2024

Modification	Description	Date
	Amazon EC2 GetSecurityGroupsForVpc pour obtenir tous les groupes de sécurité d'un VPC afin d'améliorer la validation, et l'action d'GetInstanceTypesFromInstanceRequirementsAPI Amazon EC2 pour obtenir des informations sur les types d'instances qui répondent à un certain ensemble d'exigences d'instance. Pour plus d'informations, consultez Rôles liés à un service pour Amazon EC2 Auto Scaling.	

Modification	Description	Date
Amazon EC2 Auto Scaling ajoute des autorisations à son rôle lié au service	La politique `AutoScalingServiceRolePolicy` octroie désormais des autorisations au service pour accéder aux actions d'API dont il a besoin pour une intégration avec VPC Lattice. • Actions `GetTargetGroup` et `ListTargetGroup` . Nécessaire pour récupérer des informations sur des groupes cibles de VPC Lattice. • Actions `RegisterTargets` et `DeregisterTargets` . Nécessaire pour enregistrer et annuler l'enregistrement des instances des groupes cibles de VPC Lattice. • `ListTargets` . Permet à Amazon EC2 Auto Scaling de récupérer des informations d'état des instances enregistrées auprès des groupes cibles de VPC Lattice. Pour plus d'informations, consultez Rôles liés à un service pour Amazon EC2 Auto Scaling.	6 décembre 2022

Modification	Description	Date
Amazon EC2 Auto Scaling ajoute des autorisations à son rôle lié au service	Pour permettre l'utilisation d'un AWS Systems Manager paramètre comme alias pour un ID d'AMI lors de la création d'un modèle de lancement, la `AutoScalingServiceRolePolicy` politique autorise désormais l'appel de l'action AWS Systems Manager GetParametersAPI. Pour plus d'informations, consultez Rôles liés à un service pour Amazon EC2 Auto Scaling.	28 mars 2022
Amazon EC2 Auto Scaling ajoute des autorisations à son rôle lié au service	Pour prendre en charge le dimensionnement prédictif, la `AutoScalingServiceRolePolicy` politique inclut désormais l'autorisation d'appeler l'action CloudWatch GetMetricData API. Pour plus d'informations, consultez Rôles liés à un service pour Amazon EC2 Auto Scaling.	19 mai 2021
Amazon EC2 Auto Scaling a commencé à assurer le suivi des modifications	Amazon EC2 Auto Scaling a commencé à suivre les modifications apportées à ses politiques AWS gérées.	19 mai 2021

Rôles liés à un service pour Amazon EC2 Auto Scaling

Amazon EC2 Auto Scaling utilise des rôles liés à un service pour les autorisations qui lui sont nécessaires pour l'appel d'autres Services AWS en votre compte. Un rôle lié à un service est un type unique de rôle IAM directement lié à un. Service AWS

Les rôles liés à un service offrent une manière sécurisée d'accorder des autorisations aux autres Services AWS , car seul le service lié peut assumer un rôle lié à un service. Pour plus d'informations, consultez Utilisation des rôles liés à un service dans le Guide de l'utilisateur IAM. Les rôles liés à un service permettent également à tous les appels d'API d'être visibles. AWS CloudTrail Cela facilite le suivi et la vérification des exigences, car vous pouvez suivre toutes les actions exécutées par Amazon EC2 Auto Scaling en votre nom. Pour plus d'informations, consultez Enregistrez les appels d'API Amazon EC2 Auto Scaling avec AWS CloudTrail.

Les sections suivantes décrivent comment créer et gérer des rôles liés à un service Amazon EC2 Auto Scaling. Commencez par configurer les autorisations de manière à permettre à une identité IAM (comme un utilisateur ou un rôle) de créer, modifier ou supprimer un rôle lié à un service. Pour plus d'informations, consultez Utilisation des rôles liés à un service dans le Guide de l'utilisateur IAM.

Table des matières

Présentation

Il existe deux types de rôles Amazon EC2 Auto Scaling liés à un service :

- Le rôle lié au service par défaut pour votre compte, nommé. AWSServiceRoleForAutoScaling Ce rôle est automatiquement affecté à vos groupes Auto Scaling, sauf si vous désignez un autre rôle lié à un service.

- ***Rôle lié à un service avec un suffixe personnalisé que vous spécifiez lors de la création du rôle, AWSServiceRoleForAutoScaling par exemple _ mysuffix.***

Les autorisations d'un rôle lié à un service avec suffixe personnalisé sont identiques à ceux du rôle lié à un service par défaut. Dans les deux cas, vous ne pouvez pas modifier les rôles, et vous ne pouvez pas les supprimer s'ils sont toujours en cours d'utilisation par un groupe Auto Scaling. La seule différence est le suffixe de nom de rôle.

Vous pouvez spécifier l'un ou l'autre rôle lorsque vous modifiez vos politiques AWS Key Management Service clés pour permettre aux instances lancées par Amazon EC2 Auto Scaling d'être chiffrées avec votre clé gérée par le client. Toutefois, si vous prévoyez d'accorder un accès granulaire à une clé spécifique gérée par le client, vous devez utiliser un rôle lié à un service avec suffixe personnalisé. L'utilisation d'un rôle lié à un service avec suffixe personnalisé vous permet les actions suivantes :

- Renforcer le contrôle sur la clé gérée par le client
- La possibilité de savoir quel groupe Auto Scaling a effectué un appel d'API dans vos CloudTrail journaux

Si vous créez des clés gérées par le client auxquelles tous les utilisateurs n'ont pas accès, procédez comme suit pour autoriser le suffixe personnalisé à l'utilisation d'un rôle lié à un service :

1. Crée un rôle lié à un service avec un suffixe personnalisé. Pour plus d'informations, consultez Créer un rôle lié à un service (manuel).
2. Attribuer au rôle lié à un service l'accès à une clé gérée par le client. Pour plus d'informations sur la politique de clé qui autorise la clé à être utilisée par un rôle lié à un service, consultez Politique de AWS KMS clé requise pour une utilisation avec des volumes chiffrés.
3. Donnez aux utilisateurs accès au rôle lié au service que vous avez créé. Pour plus d'informations sur la création d'une politique IAM, consultez Contrôler quel rôle lié à un service peut être transmis (en utilisant) PassRole. Si les utilisateurs essaient de spécifier un rôle lié au service sans autorisation de transmettre ce rôle au service, ils reçoivent une erreur.

Autorisations accordées par le rôle lié à un service

Amazon EC2 Auto Scaling utilise le rôle lié au service nommé AWSServiceRoleForAutoScalingou votre suffixe personnalisé lié au service.

Le rôle lié à un service approuve le fait que le service suivant endosse le rôle :

- `autoscaling.amazonaws.com`

La politique d'autorisation des rôles permet à Amazon EC2 Auto Scaling d'effectuer les actions suivantes : AutoScalingServiceRolePolicy

- `ec2`— Créez, décrivez, modifiez, démarrez/arrêtez et arrêtez des instances EC2.
- `iam`— Transférez les rôles IAM aux instances EC2 afin que les applications exécutées sur les instances puissent accéder aux informations d'identification temporaires pour le rôle.
- `iam`— Créez le rôle AWSServiceRoleForEC2Spotlié au service pour permettre à Amazon EC2 Auto Scaling de lancer des instances Spot en votre nom.
- `elasticloadbalancing`— Enregistrez et désenregistrez des instances avec Elastic Load Balancing et vérifiez l'état de santé des cibles enregistrées.
- `cloudwatch`— Créez, décrivez, modifiez et supprimez des CloudWatch alarmes pour les politiques de dimensionnement et récupérez les métriques utilisées pour le dimensionnement prédictif.
- `sns`— Publiez des notifications sur Amazon SNS lorsque des instances sont lancées ou mises hors service.
- `events`— Créez, décrivez, mettez à jour et supprimez EventBridge des règles en votre nom.
- `ssm`— Lit les paramètres depuis le Parameter Store lorsque vous utilisez un paramètre Systems Manager comme alias pour un ID d'AMI dans un modèle de lancement.
- `vpc-lattice`— Enregistrez et désenregistrez les instances avec VPC Lattice et vérifiez l'état de santé des cibles enregistrées.

Créer un rôle lié à un service (automatique)

Amazon EC2 Auto Scaling crée le rôle AWSServiceRoleForAutoScalinglié au service pour vous la première fois que vous créez un groupe Auto Scaling, sauf si vous créez manuellement un rôle lié au service avec un suffixe personnalisé et que vous le spécifiez lors de la création du groupe.

> **Important**
>
> Vous devez disposer des autorisations IAM de création du rôle lié au service. Dans le cas contraire, la création automatique échoue. Pour plus d'informations, consultez Autorisations de rôles liés à un service dans le Guide de l'utilisateur IAM et Créer un rôle lié à un service dans ce guide.

Amazon EC2 Auto Scaling a commencé à prendre en charge les rôles liés à un service en mars 2018. Si vous avez créé un groupe Auto Scaling avant cette date, Amazon EC2 Auto Scaling a créé le AWSServiceRoleForAutoScalingrôle dans votre compte. Pour plus d'informations, consultez Un nouveau rôle est apparu dans mon Compte AWS dans le Guide de l'utilisateur IAM.

Créer un rôle lié à un service (manuel)

Pour créer un rôle lié à un service (console)

1. Ouvrez la console IAM à l'adresse https://console.aws.amazon.com/iam/.
2. Dans le panneau de navigation, choisissez Roles (Rôles), puis Create role (Créer un rôle).
3. Pour Select trusted entity (Sélectionner une entité de confiance), choisissez service AWS .
4. Pour Choose the service that will use this role (Choisir le service qui utilisera ce rôle), choisissez EC2 Auto Scaling et le cas d'utilisation EC2 Auto Scaling.
5. Sélectionnez Next: Permissions (Suivant : autorisations), Next: Tags (Suivant : balises), puis Next: Review (Suivant : vérifier). Remarque : vous ne pouvez pas attacher des balises à des rôles liés à un service lors de la création.
6. ***Sur la page Révision, laissez le champ Nom du rôle vide pour créer un rôle lié au service portant ce nom AWSServiceRoleForAutoScaling, ou entrez un suffixe pour créer un rôle lié au service avec le suffixe _. AWSServiceRoleForAutoScaling***
7. (Facultatif) Dans le champ Role description (Description du rôle), modifiez la description du nouveau rôle lié à un service.
8. Sélectionnez Create role (Créer un rôle).

Pour créer un rôle lié à un service (AWS CLI)

Utilisez la commande de CLI create-service-linked-role suivante pour créer un rôle lié à un service pour Amazon EC2 Auto Scaling avec le suffixe « _ ». AWSServiceRoleForAutoScaling

```
aws iam create-service-linked-role --aws-service-name autoscaling.amazonaws.com --
custom-suffix suffix
```

La sortie de cette commande contient l'ARN du rôle lié à un service, que vous pouvez utiliser pour accorder au rôle lié à un service l'accès à vos clés gérées par le client.

```
{
    "Role": {
        "RoleId": "ABCDEF0123456789ABCDEF",
        "CreateDate": "2018-08-30T21:59:18Z",
        "RoleName": "AWSServiceRoleForAutoScaling_suffix",
        "Arn": "arn:aws:iam::123456789012:role/aws-service-role/
autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling_suffix",
        "Path": "/aws-service-role/autoscaling.amazonaws.com/",
        "AssumeRolePolicyDocument": {
            "Version": "2012-10-17",
            "Statement": [
                {
                    "Action": [
                        "sts:AssumeRole"
                    ],
                    "Principal": {
                        "Service": [
                            "autoscaling.amazonaws.com"
                        ]
                    },
                    "Effect": "Allow"
                }
            ]
        }
    }
}
```

Pour plus d'informations, consultez Création d'un rôle lié à un service dans le Guide de l'utilisateur IAM.

Modifier le rôle lié à un service

Vous ne pouvez pas modifier les rôles liés à un service qui sont créés pour Amazon EC2 Auto Scaling. Une fois que vous avez créé un rôle lié à un service, vous ne pouvez pas modifier le nom du rôle ou ses autorisations. Néanmoins, vous pouvez modifier la description du rôle. Pour plus d'informations, consultez Modification d'un rôle lié à un service dans le Guide de l'utilisateur IAM.

Suppression du rôle lié à un service

Si vous n'utilisez pas un groupe Auto Scaling, nous vous recommandons de supprimer son rôle lié à un service. La suppression du rôle vous évite d'avoir une entité qui n'est pas utilisée, ou surveillée et gérée activement.

Vous pouvez supprimer un rôle lié à un service uniquement après la suppression préalable des ressources dépendantes connexes. Cela vous évite de révoquer involontairement les autorisations Amazon EC2 Auto Scaling sur vos ressources. Si un rôle lié à un service est utilisé avec plusieurs groupes Auto Scaling, vous devez supprimer tous les groupes Auto Scaling qui utilisent le rôle lié à un service avant de pouvoir le supprimer. Pour plus d’informations, consultez Supprimer votre infrastructure Auto Scaling.

Vous pouvez utiliser IAM pour supprimer le rôle lié à un service. Pour plus d'informations, veuillez consulter Suppression d'un rôle lié à un service dans le Guide de l'utilisateur IAM.

Si vous supprimez le rôle AWSServiceRoleForAutoScalinglié à un service, Amazon EC2 Auto Scaling le crée à nouveau lorsque vous créez un groupe Auto Scaling et ne spécifiez aucun autre rôle lié au service.

Régions prises en charge pour les rôles liés à un service Amazon EC2 Auto Scaling

Amazon EC2 Auto Scaling prend en charge l'utilisation de rôles liés à un service partout Régions AWS où le service est disponible.

Exemples de politiques Amazon EC2 Auto Scaling basées sur l'identité

Par défaut, un nouvel utilisateur n' Compte AWS est pas autorisé à faire quoi que ce soit. Un administrateur IAM doit créer et assigner des politiques IAM qui accordent à une identité IAM (utilisateur ou rôle, par exemple) l'autorisation d'effectuer des actions d'API dans Amazon EC2 Auto Scaling.

Pour savoir comment créer une stratégie IAM à partir de ces exemples de documents de stratégie JSON, consultez Création de stratégies dans l'onglet JSON dans le Guide de l'utilisateur IAM.

Un exemple de politique d'autorisation est exposé ci-dessous.

```
{
   "Version": "2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": [
          "autoscaling:CreateAutoScalingGroup",
          "autoscaling:UpdateAutoScalingGroup",
          "autoscaling:DeleteAutoScalingGroup"
      ],
      "Resource": "*",
      "Condition": {
          "StringEquals": { "autoscaling:ResourceTag/purpose": "testing" }
      }
   },
   {
      "Effect": "Allow",
      "Action": "autoscaling:Describe*",
      "Resource": "*"
   }]
}
```

Cet exemple de politique accorde les autorisations de créer, mettre à jour et supprimer des groupes Auto Scaling, mais uniquement si le groupe utilise la balise **`purpose=testing`**. Comme les actions `Describe` ne prennent pas en charge les autorisations au niveau des ressources, vous devez les spécifier dans une instruction distincte sans condition. Pour lancer des instances avec un modèle de lancement, l'utilisateur doit également disposer de l'autorisation `ec2:RunInstances`. Pour plus d'informations, consultez Support de modèle de lancement.

> **Note**
>
> Vous pouvez créer vos propres politiques IAM personnalisées pour autoriser ou refuser les autorisations à des identités IAM (utilisateurs ou rôles) d'exécuter des actions Amazon EC2 Auto Scaling. Vous pouvez attacher ces politiques personnalisées aux identités IAM qui nécessitent les autorisations spécifiées. Les exemples suivants présentent des autorisations pour quelques cas d'utilisation courants.
>
> Certaines actions d'API Amazon EC2 Auto Scaling vous permettent d'inclure des groupes Auto Scaling spécifiques dans votre politique qui peuvent être créés ou modifiés par l'action. Vous pouvez restreindre les ressources cibles pour ces actions en spécifiant des ARN de

groupe Auto Scaling individuels. Toutefois, nous vous recommandons d'utiliser des politiques basées sur des balises qui autorisent (ou refusent) les actions sur les groupes Auto Scaling avec une balise spécifique.

Exemples

- Contrôler la taille des groupes Auto Scaling qui peuvent être créés
- Contrôler les clés de balise et les valeurs de balise pouvant être utilisées
- Contrôler les groupes Auto Scaling pouvant être supprimés
- Contrôler les politiques de mise à l'échelle pouvant être supprimées
- Contrôler l'accès aux actions d'actualisation des instances
- Créer un rôle lié à un service
- Contrôler quel rôle lié à un service peut être transmis (en utilisant) PassRole

Contrôler la taille des groupes Auto Scaling qui peuvent être créés

La politique suivante accorde les autorisations de créer et de mettre à jour tous les groupes Auto Scaling avec la balise **`environment=development`**, à condition que le demandeur ne spécifie pas une taille minimale inférieure à **`1`** ou une taille maximale supérieure à **`10`**. Dans la mesure du possible, utilisez des balises pour contrôler l'accès aux groupes Auto Scaling de votre compte.

```
{
   "Version": "2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": [
          "autoscaling:CreateAutoScalingGroup",
          "autoscaling:UpdateAutoScalingGroup"
      ],
      "Resource": "*",
      "Condition": {
          "StringEquals": { "autoscaling:ResourceTag/environment": "development" },
          "NumericGreaterThanEqualsIfExists": { "autoscaling:MinSize": 1 },
          "NumericLessThanEqualsIfExists": { "autoscaling:MaxSize": 10 }
      }
   }]
}
```

Par ailleurs, si vous n'utilisez pas de balises pour contrôler l'accès aux groupes Auto Scaling, vous pouvez utiliser les ARN pour identifier les groupes Auto Scaling auxquels la Politique IAM s'applique.

Un groupe Auto Scaling comprend l'ARN suivant.

```
"Resource": "arn:aws:autoscaling:region:account-
id:autoScalingGroup:*:autoScalingGroupName/my-asg"
```

Vous pouvez également spécifier plusieurs ARN en les intégrant à une liste. Pour en savoir plus sur la spécification des ARN des ressources Amazon EC2 Auto Scaling dans l'élément `Resource`, consultez Ressources sur les politiques pour Amazon EC2 Auto Scaling.

Contrôler les clés de balise et les valeurs de balise pouvant être utilisées

Vous pouvez également utiliser des conditions dans vos politiques IAM pour contrôler les clés de balise et les valeurs de balise qui peuvent être appliquées aux groupes Auto Scaling. Pour accorder les autorisations de créer ou baliser un groupe Auto Scaling uniquement si le demandeur spécifie des balises spécifiques, utilisez la clé de condition `aws:RequestTag`. Pour autoriser uniquement des clés de balise spécifiques, utilisez la clé de condition `aws:TagKeys` avec le modificateur `ForAllValues`.

La politique suivante exige que le demandeur spécifie une balise avec la clé **`environment`** dans la demande. La valeur **`"?*"`** impose qu'une valeur existe pour la clé de balise. Pour utiliser un caractère générique, vous devez utiliser l'opérateur de condition `StringLike`.

```
{
   "Version": "2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": [
          "autoscaling:CreateAutoScalingGroup",
          "autoscaling:CreateOrUpdateTags"
      ],
      "Resource": "*",
      "Condition": {
          "StringLike": { "aws:RequestTag/environment": "?*" }
      }
   }]
}
```

La politique suivante spécifie que le demandeur peut uniquement marquer les groupes Auto Scaling avec les balises **`purpose=webserver`** et **`cost-center=cc123`**, et autorise uniquement les balises **`purpose`** et **`cost-center`** (aucune autre balise ne peut être spécifiée).

```
{
   "Version": "2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": [
          "autoscaling:CreateAutoScalingGroup",
          "autoscaling:CreateOrUpdateTags"
      ],
      "Resource": "*",
      "Condition": {
          "StringEquals": {
              "aws:RequestTag/purpose": "webserver",
              "aws:RequestTag/cost-center": "cc123"
          },
          "ForAllValues:StringEquals": { "aws:TagKeys": ["purpose", "cost-center"] }
      }
   }]
}
```

La politique suivante exige que le demandeur spécifie au moins une balise dans la demande, et autorise uniquement les clés **`cost-center`** et **`owner`**.

```
{
   "Version": "2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": [
          "autoscaling:CreateAutoScalingGroup",
          "autoscaling:CreateOrUpdateTags"
      ],
      "Resource": "*",
      "Condition": {
          "ForAnyValue:StringEquals": { "aws:TagKeys": ["cost-center", "owner"] }
      }
   }]
}
```

> **Note**
>
> Pour les conditions, la clé de condition n'est pas sensible à la casse et la valeur de la condition est sensible à la casse. Par conséquent pour forcer la sensibilité à la casse d'une clé de balise, utilisez la clé de condition `aws:TagKeys`, où la clé de balise est indiquée comme une valeur dans la condition.

Contrôler les groupes Auto Scaling pouvant être supprimés

La politique suivante autorise la suppression d'un groupe Auto Scaling uniquement si le groupe comporte la balise **`environment=development`**.

```
{
   "Version": "2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": "autoscaling:DeleteAutoScalingGroup",
      "Resource": "*",
      "Condition": {
          "StringEquals": { "aws:ResourceTag/environment": "development" }
      }
   }]
}
```

Sinon, si vous n'utilisez pas de clés de condition pour contrôler l'accès aux groupes Auto Scaling, vous pouvez spécifier les ARN des ressources dans l'élément `Resource` pour contrôler l'accès à la place.

La politique suivante autorise les utilisateurs à utiliser l'action API `DeleteAutoScalingGroup`, mais uniquement pour les groupes Auto Scaling dont le nom commence par **`devteam-`**.

```
{
   "Version": "2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": "autoscaling:DeleteAutoScalingGroup",
      "Resource": "arn:aws:autoscaling:region:account-
id:autoScalingGroup:*:autoScalingGroupName/devteam-*"
   }]
```

```
}
```

Vous pouvez également spécifier plusieurs ARN en les intégrant à une liste. L'inclusion de l'UUID permet de s'assurer que l'accès est accordé au groupe Auto Scaling spécifique. L'UUID d'un nouveau groupe est différent de celui d'un groupe supprimé avec le même nom.

```
"Resource": [
    "arn:aws:autoscaling:region:account-
id:autoScalingGroup:uuid:autoScalingGroupName/devteam-1",
    "arn:aws:autoscaling:region:account-
id:autoScalingGroup:uuid:autoScalingGroupName/devteam-2",
    "arn:aws:autoscaling:region:account-
id:autoScalingGroup:uuid:autoScalingGroupName/devteam-3"
]
```

Contrôler les politiques de mise à l'échelle pouvant être supprimées

La politique suivante accorde les autorisations d'utiliser l'action `DeletePolicy` pour supprimer une politique de mise à l'échelle. Cependant, elle refuse également l'action si le groupe Auto Scaling cible dispose de la balise **`environment=production`**. Dans la mesure du possible, utilisez des balises pour contrôler l'accès aux groupes Auto Scaling de votre compte.

```
{
   "Version": "2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": "autoscaling:DeletePolicy",
      "Resource": "*"
   },
   {
      "Effect": "Deny",
      "Action": "autoscaling:DeletePolicy",
      "Resource": "*",
      "Condition": {
          "StringEquals": { "autoscaling:ResourceTag/environment": "production" }
      }
   }]
}
```

Contrôler l’accès aux actions d’actualisation des instances

La politique suivante autorise le démarrage, la restauration et l’annulation d’une actualisation d’instance uniquement si le groupe Auto Scaling concerné comporte la balise **`environment=testing`**. Comme les actions `Describe` ne prennent pas en charge les autorisations au niveau des ressources, vous devez les spécifier dans une instruction distincte sans condition.

```
{
   "Version": "2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": [
          "autoscaling:StartInstanceRefresh",
          "autoscaling:CancelInstanceRefresh",
          "autoscaling:RollbackInstanceRefresh"
      ],
      "Resource": "*",
      "Condition": {
          "StringEquals": { "autoscaling:ResourceTag/environment": "testing" }
      }
   },
   {
      "Effect": "Allow",
      "Action": "autoscaling:DescribeInstanceRefreshes",
      "Resource": "*"
   }]
}
```

Pour spécifier la configuration souhaitée dans l’appel `StartInstanceRefresh`, les utilisateurs peuvent avoir besoin de certaines autorisations associées, telles que :

- ec2 : RunInstances — Pour lancer des instances EC2 à l'aide d'un modèle de lancement, l'utilisateur doit avoir l'`ec2:RunInstances`autorisation requise dans une politique IAM. Pour plus d’informations, consultez Support de modèle de lancement.
- ec2 : CreateTags — Pour lancer des instances EC2 à partir d'un modèle de lancement qui ajoute des balises aux instances et aux volumes lors de leur création, l'utilisateur doit disposer de l'`ec2:CreateTags`autorisation prévue dans une politique IAM. Pour plus d’informations, consultez Autorisations requises pour baliser des instances et des volumes.

- iam : PassRole — Pour lancer des instances EC2 à partir d'un modèle de lancement contenant un profil d'instance (un conteneur pour un rôle IAM), l'utilisateur doit également disposer de l'`iam:PassRole`autorisation prévue dans une politique IAM. Pour plus d'informations, et un exemple de politique IAM, consultez Rôle IAM pour les applications qui s'exécutent sur des instances Amazon EC2.
- ssm : GetParameters — Pour lancer des instances EC2 à partir d'un modèle de lancement utilisant un AWS Systems Manager paramètre, l'utilisateur doit également disposer de l'`ssm:GetParameters`autorisation dans une politique IAM. Pour plus d'informations, consultez Utiliser des AWS Systems Manager paramètres plutôt que des ID d'AMI dans les modèles de lancement.

Créer un rôle lié à un service

Amazon EC2 Auto Scaling a besoin d'autorisations pour créer un rôle lié à un service la première fois qu'un de vos utilisateurs appelle des actions d'API Amazon Compte AWS EC2 Auto Scaling. Si le rôle lié à un service n'existe pas déjà, Amazon EC2 Auto Scaling le crée dans votre compte. Le rôle lié à un service donne des autorisations à Amazon EC2 Auto Scaling afin qu'il puisse Services AWS appeler d'autres personnes en votre nom.

Pour que cette création de rôle automatique aboutisse, les utilisateurs doivent disposer des autorisations nécessaires pour l'action `iam:CreateServiceLinkedRole`.

```
"Action": "iam:CreateServiceLinkedRole"
```

L'exemple suivant illustre une politique d'autorisations qui permet à un utilisateur de créer un rôle lié au service Amazon EC2 Auto Scaling pour Amazon EC2 Auto Scaling.

```
{
   "Version": "2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": "iam:CreateServiceLinkedRole",
      "Resource": "arn:aws:iam::*:role/aws-service-role/
autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling",
      "Condition": {
          "StringLike": { "iam:AWSServiceName":"autoscaling.amazonaws.com" }
      }
   }]
}
```

Contrôler quel rôle lié à un service peut être transmis (en utilisant) PassRole

Les utilisateurs qui créent ou mettent à jour des groupes Auto Scaling et qui spécifient un rôle lié à un service avec suffixe personnalisé dans la demande ont besoin de l'autorisation `iam:PassRole`.

Vous pouvez utiliser cette `iam:PassRole` autorisation pour protéger la sécurité des clés gérées par vos AWS KMS clients si vous autorisez différents rôles liés au service à accéder à différentes clés. En fonction des besoins de votre organisation, vous pouvez avoir une clé pour l'équipe de développement, une autre pour l'équipe assurance qualité et encore une autre pour l'équipe financière. Créez d'abord un rôle lié à un service ayant accès à la clé requise, par exemple un rôle lié à un service nommé. AWSServiceRoleForAutoScaling_devteamkeyaccess Attachez ensuite la politique à une identité IAM, telle qu'un utilisateur ou un rôle.

La politique suivante accorde les autorisations de transmettre le rôle **`AWSServiceRoleForAutoScaling_devteamkeyaccess`**à un groupe Auto Scaling dont le nom commence par **`devteam-`**. Si l'identité IAM qui crée le groupe Auto Scaling essaie de spécifier un rôle lié à un service différent, elle reçoit une erreur. S'ils choisissent de ne pas spécifier de rôle lié à un service, le AWSServiceRoleForAutoScalingrôle par défaut est utilisé à la place.

```
{
   "Version": "2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": "iam:PassRole",
      "Resource": "arn:aws:iam::account-id:role/aws-service-role/
autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling_devteamkeyaccess",
      "Condition": {
          "StringEquals": { "iam:PassedToService": [ "autoscaling.amazonaws.com" ] },
          "StringLike": { "iam:AssociatedResourceARN":
 [ "arn:aws:autoscaling:region:account-
id:autoScalingGroup:*:autoScalingGroupName/devteam-*" ] }
      }
   }]
}
```

Pour plus d'informations sur les rôles liés à un service avec suffixe personnalisé, consultez Rôles liés à un service pour Amazon EC2 Auto Scaling.

Prévention du problème de l’adjoint confus entre services

Le problème de député confus est un problème de sécurité dans lequel une entité qui n’est pas autorisée à effectuer une action peut contraindre une entité plus privilégiée à le faire.

En AWS, l'usurpation d'identité interservices peut entraîner la confusion des adjoints. L’usurpation d’identité entre services peut se produire lorsqu’un service (le service appelant) appelle un autre service (le service appelé). Le service appelant peut être manipulé et ses autorisations utilisées pour agir sur les ressources d’un autre client auxquelles on ne serait pas autorisé d’accéder autrement.

Pour éviter cela, AWS fournit des outils qui vous aident à protéger vos données pour tous les services auprès des principaux fournisseurs de services qui ont obtenu l'accès aux ressources de votre compte. Nous vous recommandons d'utiliser les clés de contexte de condition globales `aws:SourceArn` et `aws:SourceAccount` dans les politiques d'approbation pour les rôles de service Amazon EC2 Auto Scaling. Ces clés limitent les autorisations qu'Amazon EC2 Auto Scaling accorde à un autre service pour la ressource.

Les valeurs des `SourceAccount` champs `SourceArn` et sont définies lorsqu'Amazon EC2 Auto Scaling utilise AWS Security Token Service (AWS STS) pour assumer un rôle en votre nom.

Pour utiliser les clés de condition globales `aws:SourceArn` ou `aws:SourceAccount`, définissez comme valeur l'Amazon Resource Name (ARN) ou le compte de la ressource qu'Amazon EC2 Auto Scaling stocke. Dans la mesure du possible, utilisez `aws:SourceArn`, qui est plus spécifique. Définissez comme valeur l'ARN ou un modèle d'ARN avec des caractères génériques (*) pour les parties inconnues de l'ARN. Si vous ne connaissez pas l'ARN de la ressource, utilisez `aws:SourceAccount` à la place.

L'exemple suivant montre comment utiliser les clés de contexte de condition globale `aws:SourceArn` et `aws:SourceAccount` dans Amazon EC2 Auto Scaling pour éviter le problème de l'adjoint confus.

Exemple : utilisation des clés de condition **`aws:SourceArn`** et **`aws:SourceAccount`**

Un rôle qu'un service endosse pour effectuer des actions en votre nom s'appelle un rôle de service. Si vous souhaitez créer des hooks de cycle de vie qui envoient des notifications à un autre endroit qu'Amazon EventBridge, vous devez créer un rôle de service pour permettre à Amazon EC2 Auto Scaling d'envoyer des notifications à une rubrique Amazon SNS ou à une file d'attente Amazon SQS en votre nom. Si vous souhaitez qu'un seul groupe Auto Scaling soit associé à l'accès interservice, vous pouvez spécifier la politique d'approbation de la fonction du service comme suit.

Cet exemple de politique d'approbation utilise des déclarations de condition pour limiter la capacité `AssumeRole` sur la fonction du service uniquement aux actions qui affectent le groupe Auto Scaling spécifié dans le compte spécifié. Les conditions `aws:SourceArn` et `aws:SourceAccount` sont évaluées indépendamment. Toute demande d'utilisation de la fonction du service doit répondre aux deux conditions.

Avant d'utiliser cette politique, remplacez la région, l'ID de compte, l'UUID, et le nom du groupe par des valeurs valides de votre compte.

```
{
  "Version": "2012-10-17",
  "Statement": {
    "Sid": "ConfusedDeputyPreventionExamplePolicy",
    "Effect": "Allow",
    "Principal": {
      "Service": "autoscaling.amazonaws.com"
    },
    "Action": "sts:AssumeRole",
    "Condition": {
      "ArnLike": {
        "aws:SourceArn":
 "arn:aws:autoscaling:region:account_id:autoScalingGroup:uuid:autoScalingGroupName/my-
asg"
      },
      "StringEquals": {
        "aws:SourceAccount": "account_id"
      }
    }
  }
}
```

Dans l'exemple précédent :

- L'élément `Principal` spécifie le principal de service du service (`autoscaling.amazonaws.com`).
- L'élément `Action` indique l'action `sts:AssumeRole`.
- L'élément `Condition` spécifie les clés de condition `aws:SourceArn` et `aws:SourceAccount` globales. L'ARN de la source inclut l'ID de compte. Il n'est donc pas nécessaire d'utiliser `aws:SourceAccount` avec `aws:SourceArn`.

Informations supplémentaires

Pour plus d'informations, veuillez consulter la rubrique Clés de contexte de condition globales AWS, Le problème d'adjoint confus, et la Modification d'une politique d'approbation de rôle (console)dans le Guide de l'utilisateur IAM.

Support de modèle de lancement

Amazon EC2 Auto Scaling prend en charge l'utilisation de modèles de lancement Amazon EC2 avec vos groupes Auto Scaling. Nous vous recommandons d'autoriser les utilisateurs à créer des groupes Auto Scaling à partir de modèles de lancement, car cela leur permet d'utiliser les dernières fonctionnalités d'Amazon EC2 Auto Scaling et d'Amazon EC2. Par exemple, les utilisateurs doivent spécifier un modèle de lancement pour utiliser une politique d'instances mixtes.

Vous pouvez utiliser la politique `AmazonEC2FullAccess` pour donner aux utilisateurs un accès complet aux ressources Amazon EC2 Auto Scaling, aux modèles de lancement et aux autres ressources EC2 dans leur compte. Vous pouvez également créer vos propres politiques IAM personnalisées pour accorder aux utilisateurs des autorisations précises d'utiliser des modèles de lancement, comme décrit dans cette rubrique.

Exemple de politique que vous pouvez personnaliser pour votre propre usage

Voici un exemple de politique d'autorisations de base que vous pouvez adapter à votre propre usage. La politique accorde les autorisations de créer, mettre à jour et supprimer tous les groupes Auto Scaling, mais uniquement si le groupe utilise la balise **`purpose=testing`**. Elle donne ensuite l'autorisation d'exécuter toutes les actions de type `Describe`. Comme les actions `Describe` ne prennent pas en charge les autorisations au niveau des ressources, vous devez les spécifier dans une instruction distincte sans condition.

Les identités IAM (utilisateurs ou rôles) disposant de cette politique sont autorisées à créer ou mettre à jour un groupe Auto Scaling à l'aide d'un modèle de lancement, car ils ont également l'autorisation d'utiliser l'action `ec2:RunInstances`.

```
{
    "Version": "2012-10-17",
    "Statement": [
          {
            "Effect": "Allow",
            "Action": [
                "autoscaling:CreateAutoScalingGroup",
```

```
                "autoscaling:UpdateAutoScalingGroup",
                "autoscaling:DeleteAutoScalingGroup"
            ],
            "Resource": "*",
            "Condition": {
                "StringEquals": { "autoscaling:ResourceTag/purpose": "testing" }
            }
        },
        {
            "Effect": "Allow",
            "Action": [
                "autoscaling:Describe*",
                "ec2:RunInstances"
            ],
            "Resource": "*"
        }
    ]
}
```

Les utilisateurs qui créent ou mettent à jour des groupes Auto Scaling peuvent avoir besoin de certaines autorisations connexes, telles que :

- ec2 : CreateTags — Pour ajouter des balises aux instances et aux volumes lors de leur création, l'utilisateur doit avoir l'`ec2:CreateTags`autorisation requise dans une politique IAM. Pour plus d'informations, consultez Autorisations requises pour baliser des instances et des volumes.
- iam : PassRole — Pour lancer des instances EC2 à partir d'un modèle de lancement contenant un profil d'instance (un conteneur pour un rôle IAM), l'utilisateur doit également disposer de l'`iam:PassRole`autorisation prévue dans une politique IAM. Pour plus d'informations, et un exemple de politique IAM, consultez Rôle IAM pour les applications qui s'exécutent sur des instances Amazon EC2.
- ssm : GetParameters — Pour lancer des instances EC2 à partir d'un modèle de lancement utilisant un AWS Systems Manager paramètre, l'utilisateur doit également disposer de l'`ssm:GetParameters`autorisation dans une politique IAM. Pour plus d'informations, consultez Utiliser des AWS Systems Manager paramètres plutôt que des ID d'AMI dans les modèles de lancement.

Ces autorisations relatives aux actions à effectuer lors du lancement d'instances sont vérifiées lorsque l'utilisateur interagit avec un groupe Auto Scaling. Pour plus d'informations, consultez Validation des autorisations pour `ec2:RunInstances` et `iam:PassRole`.

Les exemples suivants illustrent des déclarations de politique que vous pouvez utiliser pour contrôler l'accès dont les utilisateurs IAM disposent pour utiliser des modèles de lancement.

Rubriques

Demander des modèles de lancement dotés d'une balise spécifique

Lorsque vous accordez des autorisations `ec2:RunInstances`, vous pouvez spécifier que les utilisateurs ne peuvent utiliser que des modèles de lancement dotés de balises ou d'identifiants spécifiques pour limiter les autorisations lors du lancement d'instances avec modèle de lancement. Vous pouvez également contrôler l'AMI et les autres ressources auxquelles toute personne utilisant des modèles de lancement peut faire référence et utiliser lors du lancement d'instances en spécifiant des autorisations supplémentaires au niveau des ressources pour l'appel à `RunInstances`.

L'exemple suivant restreint les autorisations pour l'action `ec2:RunInstances` avec les modèles de lancement qui se trouvent dans la région spécifiée et qui disposent de la balise **`purpose=testing`**. Cela permet également aux utilisateurs d'accéder aux ressources spécifiées dans un modèle de lancement : AMI, types d'instances, volumes, paires de clés, interfaces réseau et groupes de sécurité.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "ec2:RunInstances",
            "Resource": "arn:aws:ec2:region:account-id:launch-template/*",
            "Condition": {
                "StringEquals": { "aws:ResourceTag/purpose": "testing" }
```

```
            }
        },
        {
            "Effect": "Allow",
            "Action": "ec2:RunInstances",
            "Resource": [
                "arn:aws:ec2:region::image/ami-*",
                "arn:aws:ec2:region:account-id:instance/*",
                "arn:aws:ec2:region:account-id:subnet/*",
                "arn:aws:ec2:region:account-id:volume/*",
                "arn:aws:ec2:region:account-id:key-pair/*",
                "arn:aws:ec2:region:account-id:network-interface/*",
                "arn:aws:ec2:region:account-id:security-group/*"
            ]
        }
    ]
}
```

Pour plus d'informations sur l'utilisation de politiques basées sur des balises avec les modèles de lancement, consultez la section Contrôler l'accès aux modèles de lancement avec des autorisations IAM dans le guide de l'utilisateur Amazon EC2.

Demander un modèle de lancement et un numéro de version

Vous pouvez également utiliser les autorisations IAM pour imposer qu'un modèle de lancement et le numéro de version du modèle de lancement soient spécifiés lors de la création ou de la mise à jour de groupes Auto Scaling.

Dans l'exemple suivant, les utilisateurs peuvent créer et mettre à jour des groupes Auto Scaling uniquement si un modèle de lancement et le numéro de version du modèle de lancement sont spécifiés. Si les utilisateurs disposant de cette politique omettent le numéro de version du modèle de lancement pour spécifier `$Latest` ou `$Default`, ou s'ils essaient d'utiliser une configuration du lancement à la place, l'action échoue.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "autoscaling:CreateAutoScalingGroup",
                "autoscaling:UpdateAutoScalingGroup"
```

```
            ],
            "Resource": "*",
            "Condition": {
                "Bool": { "autoscaling:LaunchTemplateVersionSpecified": "true" }
            }
        },
        {
            "Effect": "Deny",
            "Action": [
                "autoscaling:CreateAutoScalingGroup",
                "autoscaling:UpdateAutoScalingGroup"
            ],
            "Resource": "*",
            "Condition": {
                "Null": { "autoscaling:LaunchConfigurationName": "false" }
            }
        }
    ]
}
```

Demander l'utilisation du service de métadonnées d'instance version 2 (IMDSv2)

Pour plus de sécurité, vous pouvez définir les autorisations de vos utilisateurs de sorte à exiger l'utilisation d'un modèle de lancement qui nécessite IMDSv2. Pour plus d'informations, consultez Configuration du service de métadonnées d'instance dans le guide de l'utilisateur Amazon EC2.

L'exemple suivant indique que les utilisateurs ne peuvent pas appeler l'action `ec2:RunInstances`, sauf si l'instance est également définie pour exiger l'utilisation d'IMDSv2 (indiqué par `"ec2:MetadataHttpTokens":"required"`).

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "RequireImdsV2",
            "Effect": "Deny",
            "Action": "ec2:RunInstances",
            "Resource": "arn:aws:ec2:*:*:instance/*",
            "Condition": {
                "StringNotEquals": { "ec2:MetadataHttpTokens": "required" }
            }
        }
    ]
```

```
}
```

> **Tip**
>
> Pour forcer le lancement des instances Auto Scaling de remplacement qui utilisent un nouveau modèle de lancement ou une nouvelle version d'un modèle de lancement avec les options de métadonnées d'instance configurées, vous pouvez démarrer une actualisation d'instance. Pour plus d'informations, consultez Mise à jour des instances Auto Scaling.

Limiter l'accès aux ressources Amazon EC2

L'exemple suivant montre comment contrôler la configuration des instances qu'un utilisateur peut lancer en limitant l'accès aux ressources Amazon EC2. Pour spécifier des autorisations au niveau des ressources pour les ressources spécifiées dans un modèle de lancement, vous devez inclure les ressources dans l'instruction d'action `RunInstances`.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "ec2:RunInstances",
            "Resource": [
                "arn:aws:ec2:region:account-id:launch-template/*",
                "arn:aws:ec2:region::image/ami-04d5cc9b88example",
                "arn:aws:ec2:region:account-id:subnet/subnet-1a2b3c4d",
                "arn:aws:ec2:region:account-id:volume/*",
                "arn:aws:ec2:region:account-id:key-pair/*",
                "arn:aws:ec2:region:account-id:network-interface/*",
                "arn:aws:ec2:region:account-id:security-group/sg-903004f88example"
            ]
        },
        {
            "Effect": "Allow",
            "Action": "ec2:RunInstances",
            "Resource": "arn:aws:ec2:region:account-id:instance/*",
            "Condition": {
                "StringEquals": { "ec2:InstanceType": ["t2.micro", "t2.small"] }
            }
        }
```

```
    ]
}
```

Cet exemple contient deux instructions :

- La première instruction exige que les utilisateurs lancent des instances dans un sous-réseau spécifique (**`subnet-1a2b3c4d`**), en utilisant un groupe de sécurité spécifique (**`sg-903004f88example`**) et une AMI spécifique (**`ami-04d5cc9b88example`**). Cela permet également aux utilisateurs d'accéder aux ressources spécifiées dans un modèle de lancement : interfaces réseau, paires de clés, et volumes.
- La seconde instruction permet aux utilisateurs de lancer des instances uniquement à l'aide des types d'instance **`t2.micro`** et **`t2.small`**, ce que vous pourriez faire pour contrôler les coûts.

  Notez toutefois qu'il n'existe actuellement aucun moyen efficace d'empêcher complètement les utilisateurs autorisés à lancer des instances avec un modèle de lancement de lancer d'autres types d'instances. Cela est dû au fait qu'un type d'instance spécifié dans un modèle de lancement peut être remplacé pour utiliser des types d'instance définis à l'aide d'une sélection de type d'instance basée sur les attributs.

Pour obtenir la liste complète des autorisations au niveau des ressources que vous pouvez utiliser pour contrôler la configuration des instances qu'un utilisateur peut lancer, consultez la section Actions, ressources et clés de condition pour Amazon EC2 dans la Référence de l'autorisation de service.

Autorisations requises pour baliser des instances et des volumes

L'exemple suivant permet aux utilisateurs de baliser les instances et les volumes lors de leur création. Cette politique est nécessaire si des balises sont spécifiées dans le modèle de lancement. Pour plus d'informations, consultez la section Accorder l'autorisation de baliser les ressources lors de leur création dans le guide de l'utilisateur Amazon EC2.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "ec2:CreateTags",
            "Resource": "arn:aws:ec2:region:account-id:*/*",
            "Condition": {
```

```
                "StringEquals": { "ec2:CreateAction": "RunInstances" }
            }
        }
    ]
}
```

Autorisations de modèle de lancement supplémentaires

Vous devez accorder aux utilisateurs de votre console des autorisations pour les actions `ec2:DescribeLaunchTemplates` et `ec2:DescribeLaunchTemplateVersions`. Sans ces autorisations, les données du modèle de lancement ne peuvent pas se charger dans l'assistant de groupe Auto Scaling, et les utilisateurs ne peuvent pas utiliser l'assistant pour lancer des instances via un modèle de lancement. Vous pouvez spécifier ces actions supplémentaires dans l'élément `Action` d'une instruction de politique IAM.

Validation des autorisations pour **`ec2:RunInstances`** et **`iam:PassRole`**

Les utilisateurs peuvent spécifier la version du modèle de lancement utilisée par leur groupe Auto Scaling. En fonction de leurs autorisations, il peut s'agir d'une version numérotée spécifique ou de la version `$Latest` ou `$Default` du modèle de lancement. Si c'est le dernier cas, faites particulièrement attention. Cela peut annuler les autorisations pour `ec2:RunInstances` et `iam:PassRole` que vous aviez l'intention de restreindre.

Cette section explique le scénario d'utilisation de la dernière version ou de la version par défaut du modèle de lancement avec un groupe Auto Scaling.

Lorsqu'un utilisateur appelle les API `CreateAutoScalingGroup`, `UpdateAutoScalingGroup` ou `StartInstanceRefresh`, Amazon EC2 Auto Scaling vérifie ses autorisations par rapport à la version du modèle de lancement qui est la version la plus récente ou la version par défaut à ce moment-là avant de traiter la demande. Cela valide les autorisations pour les actions à effectuer lors du lancement d'instances, telles que les actions `ec2:RunInstances` et `iam:PassRole`. Pour ce faire, nous lançons un appel d'exécution à RunInstancessec Amazon EC2 afin de valider si l'utilisateur dispose des autorisations requises pour effectuer l'action, sans réellement faire la demande. Lorsqu'une réponse est renvoyée, elle est lue par Amazon EC2 Auto Scaling. Si les autorisations de l'utilisateur n'autorisent pas une action donnée, Amazon EC2 Auto Scaling fait échouer la demande et renvoie à l'utilisateur une erreur contenant des informations sur l'autorisation manquante.

Une fois la vérification et la demande initiales terminées, chaque fois que les instances sont lancées, Amazon EC2 Auto Scaling les lance avec la version la plus récente ou par défaut, même si elle a

changé, en utilisant les autorisations de son rôle lié au service. Cela signifie qu'un utilisateur utilisant le modèle de lancement peut potentiellement le mettre à jour pour transmettre un rôle IAM à une instance même s'il n'en dispose pas de l'autorisation `iam:PassRole`.

Utilisez la clé de condition `autoscaling:LaunchTemplateVersionSpecified` si vous souhaitez limiter le nombre de personnes ayant accès à la configuration des groupes pour utiliser la version `$Latest` ou `$Default`. Cela garantit que le groupe Auto Scaling n'accepte qu'une version numérotée spécifique lorsqu'un utilisateur appelle les API `CreateAutoScalingGroup` et `UpdateAutoScalingGroup`. Pour un exemple indiquant comment ajouter cette clé de condition à une politique IAM, consultez Demander un modèle de lancement et un numéro de version.

Pour les groupes Auto Scaling configurés pour utiliser la version du modèle de lancement `$Latest` ou `$Default`, envisagez de limiter le nombre de personnes autorisées à créer et à gérer les versions du modèle de lancement, y compris l'action `ec2:ModifyLaunchTemplate`permettant à un utilisateur de spécifier la version du modèle de lancement par défaut. Pour plus d'informations, consultez la section Contrôler les autorisations de version dans le guide de l'utilisateur Amazon EC2.

Ressources connexes

Pour en savoir plus sur les autorisations permettant de consulter, de créer et de supprimer des modèles de lancement et des versions de modèles de lancement, consultez la section Contrôler l'accès aux modèles de lancement avec des autorisations IAM dans le guide de l'utilisateur Amazon EC2.

Pour en savoir plus sur les autorisations au niveau des ressources que vous pouvez utiliser pour contrôler l'accès à l'appel `RunInstances`, consultez la section Actions, ressources et clés de condition pour Amazon EC2 dans la Référence de l'autorisation de service.

Rôle IAM pour les applications qui s'exécutent sur des instances Amazon EC2

Les applications qui s'exécutent sur les instances Amazon EC2 requièrent des informations d'identification afin d'accéder aux autres Services AWS. Pour fournir ces informations d'identification en toute sécurité, utilisez un rôle IAM. Le rôle fournit des autorisations temporaires que l'application peut utiliser lorsqu'elle accède à d'autres ressources AWS . Les autorisations du rôle déterminent ce que l'application est autorisée à faire.

Pour les instances d'un groupe Auto Scaling, vous devez créer une configuration ou un modèle de lancement et choisir un profil d'instance à associer aux instances. Un profil d'instance est un

conteneur pour un rôle IAM qui permet à Amazon EC2 de transmettre le rôle IAM à une instance lors du lancement de l'instance. Créez d'abord un rôle IAM doté de toutes les autorisations requises pour accéder aux AWS ressources. Créez ensuite le profil d'instance et affectez-lui le rôle.

> **Note**
>
> En tant que bonne pratique, nous vous recommandons vivement de créer le rôle de manière à ce qu'il dispose des autorisations minimales nécessaires pour accéder aux autres Services AWS rôles requis par votre application.

Table des matières

Prérequis

Créez le rôle IAM que votre application s'exécutant sur Amazon EC2 peut endosser. Choisissez les autorisations appropriées, de manière à ce que l'application à laquelle est accordée par la suite le rôle puisse procéder aux appels d'API dont elle a besoin.

Si vous utilisez la console IAM au lieu du AWS CLI ou de l'un des AWS SDK, la console crée automatiquement un profil d'instance et lui donne le même nom que le rôle auquel il correspond.

Pour créer un rôle IAM (console)

1. Ouvrez la console IAM à l’adresse https://console.aws.amazon.com/iam/.
2. Dans le panneau de navigation de gauche, sélectionnez Roles (Rôles).
3. Sélectionnez Create role (Créer un rôle).
4. Pour Select trusted entity (Sélectionner une entité de confiance), choisissez service AWS .
5. Pour votre cas d'utilisation, sélectionnez EC2, puis Next (Suivant).
6. Si possible, sélectionnez la politique à utiliser pour la politique d'autorisations ou choisissez Create policy (Créer une politique) pour ouvrir un nouvel onglet de navigateur et créer une nouvelle politique de bout en bout. Pour plus d’informations, consultez Création de politiques IAM dans le Guide de l’utilisateur IAM. Une fois la politique créée, fermez cet onglet et revenez

à l'onglet initial. Cochez la case en regard des stratégies d'autorisations que vous souhaitez octroyer au service.

7. (Facultatif) Définissez une limite d'autorisations. Il s'agit d'une fonctionnalité avancée disponible pour les rôles de service. Pour plus d'informations, consultez Limites d'autorisations pour des entités IAM dans le Guide de l'utilisateur IAM.
8. Choisissez Suivant.
9. Sur la page Name, review, and create (Nommer, réviser et créer), pour Role name (Nom de rôle), saisissez un nom de rôle vous permettant d'identifier l'objectif de ce rôle. Ce nom doit être unique au sein de votre Compte AWS. Étant donné que d'autres AWS ressources peuvent faire référence au rôle, vous ne pouvez pas modifier le nom du rôle une fois celui-ci créé.
10. Passez en revue les informations du rôle, puis choisissez Create role (Créer un rôle).

Autorisations IAM

Utilisez une politique basée sur l'identité IAM pour contrôler l'accès à votre nouveau rôle IAM. L'autorisation `iam:PassRole` est requise pour l'utilisateur IAM (utilisateur ou rôle) qui crée ou met à jour un groupe Auto Scaling à l'aide d'un modèle de lancement qui spécifie un profil d'instance.

L'exemple de politique suivant accorde les autorisations de transmettre uniquement des rôles IAM dont le nom commence par **`qateam-`**.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "iam:PassRole",
            "Resource": "arn:aws:iam::account-id:role/qateam-*",
            "Condition": {
                "StringEquals": {
                    "iam:PassedToService": [
                        "ec2.amazonaws.com",
                        "ec2.amazonaws.com.cn"
                    ]
                }
            }
        }
    ]
}
```

> **Important**
>
> Pour obtenir des informations sur la manière dont Amazon EC2 Auto Scaling valide les autorisations pour l'action `iam:PassRole` d'un groupe Auto Scaling qui utilise un modèle de lancement, consultez Validation des autorisations pour `ec2:RunInstances` et `iam:PassRole`.

Création d'un modèle de lancement

Lorsque vous créez le modèle de lancement à l' AWS Management Console aide de la section Détails avancés, sélectionnez le rôle dans le profil d'instance IAM. Pour plus d'informations, consultez Créer un modèle de lancement à l'aide de paramètres avancés.

Lorsque vous créez le modèle de lancement à l'aide de la commande create-launch-template du AWS CLI, spécifiez le nom du profil d'instance de votre rôle IAM, comme indiqué dans l'exemple suivant.

```
aws ec2 create-launch-template --launch-template-name my-lt-with-instance-profile --
version-description version1 \
--launch-template-data
 '{"ImageId":"ami-04d5cc9b88example","InstanceType":"t2.micro","IamInstanceProfile":
{"Name":"my-instance-profile"}}'
```

Consultez aussi

Pour plus d'informations permettant de découvrir et d'utiliser les rôles IAM pour Amazon EC2, consultez :

- Rôles IAM pour Amazon EC2 dans le guide de l'utilisateur Amazon EC2
- Utilisation de profils d'instance et Utilisation d'un rôle IAM pour accorder des autorisations à des applications s'exécutant sur des instances Amazon EC2 dans le Guide de l'utilisateur IAM

Validation de la conformité pour Amazon EC2 Auto Scaling

Pour savoir si un programme Services AWS de conformité Service AWS s'inscrit dans le champ d'application de programmes de conformité spécifiques, consultez Services AWS la section de

conformité et sélectionnez le programme de conformité qui vous intéresse. Pour des informations générales, voir Programmes de AWS conformité Programmes AWS de .

Vous pouvez télécharger des rapports d'audit tiers à l'aide de AWS Artifact. Pour plus d'informations, voir Téléchargement de rapports dans AWS Artifact .

Votre responsabilité en matière de conformité lors de l'utilisation Services AWS est déterminée par la sensibilité de vos données, les objectifs de conformité de votre entreprise et les lois et réglementations applicables. AWS fournit les ressources suivantes pour faciliter la mise en conformité :

- Guides de démarrage rapide sur la sécurité et la conformité : ces guides de déploiement abordent les considérations architecturales et indiquent les étapes à suivre pour déployer des environnements de base axés sur AWS la sécurité et la conformité.
- Architecture axée sur la sécurité et la conformité HIPAA sur Amazon Web Services : ce livre blanc décrit comment les entreprises peuvent créer des applications AWS conformes à la loi HIPAA.

  > **Note**
  >
  > Tous ne Services AWS sont pas éligibles à la loi HIPAA. Pour plus d'informations, consultez le HIPAA Eligible Services Reference.

- AWS Ressources de https://aws.amazon.com/compliance/resources/ de conformité — Cette collection de classeurs et de guides peut s'appliquer à votre secteur d'activité et à votre région.
- AWS Guides de conformité destinés aux clients — Comprenez le modèle de responsabilité partagée sous l'angle de la conformité. Les guides résument les meilleures pratiques en matière de sécurisation Services AWS et décrivent les directives relatives aux contrôles de sécurité dans de nombreux cadres (notamment le National Institute of Standards and Technology (NIST), le Payment Card Industry Security Standards Council (PCI) et l'Organisation internationale de normalisation (ISO)).
- Évaluation des ressources à l'aide des règles du guide du AWS Config développeur : le AWS Config service évalue dans quelle mesure les configurations de vos ressources sont conformes aux pratiques internes, aux directives du secteur et aux réglementations.
- AWS Security Hub— Cela Service AWS fournit une vue complète de votre état de sécurité interne AWS. Security Hub utilise des contrôles de sécurité pour évaluer vos ressources AWS et vérifier votre conformité par rapport aux normes et aux bonnes pratiques du secteur de la sécurité. Pour

obtenir la liste des services et des contrôles pris en charge, consultez Référence des contrôles Security Hub.

- Amazon GuardDuty — Cela Service AWS détecte les menaces potentielles qui pèsent sur vos charges de travail Comptes AWS, vos conteneurs et vos données en surveillant votre environnement pour détecter toute activité suspecte et malveillante. GuardDuty peut vous aider à répondre à diverses exigences de conformité, telles que la norme PCI DSS, en répondant aux exigences de détection des intrusions imposées par certains cadres de conformité.
- AWS Audit Manager— Cela vous Service AWS permet d'auditer en permanence votre AWS utilisation afin de simplifier la gestion des risques et la conformité aux réglementations et aux normes du secteur.

Conformité PCI DSS

Amazon EC2 Auto Scaling prend en charge le traitement, le stockage et la transmission des données de cartes bancaires par un commerçant ou un fournisseur de services et a été validé comme étant conforme à la norme PCI (Payment Card Industry) DSS (Data Security Standard). Pour plus d'informations sur la norme PCI DSS, notamment sur la manière de demander une copie du Package de AWS conformité PCI, consultez la section PCI DSS niveau 1.

Pour plus d'informations sur la mise en conformité avec la norme PCI DSS pour vos AWS charges de travail, reportez-vous au guide de conformité suivant :

- Norme de sécurité des données de l'industrie des cartes de paiement (PCI DSS) 3.2.1 sur AWS

Amazon EC2 Auto Scaling et points de terminaison d'un VPC d'interface

Vous pouvez améliorer la posture de sécurité de votre VPC en configurant Amazon EC2 Auto Scaling de sorte à utiliser un point de terminaison de VPC d'interface. Les points de terminaison de l'interface sont alimentés par AWS PrivateLink une technologie qui vous permet d'accéder de manière privée aux API Amazon EC2 Auto Scaling en limitant tout le trafic réseau entre votre VPC et Amazon EC2 Auto Scaling vers le réseau. AWS Avec les points de terminaison d'interface, vous n'avez pas non plus besoin d'une passerelle Internet, d'un appareil NAT ou d'une passerelle privée virtuelle.

Vous n'êtes pas obligé de le configurer AWS PrivateLink, mais c'est recommandé. Pour plus d'informations sur les points AWS PrivateLink de terminaison VPC, consultez Qu'est-ce que c'est ? AWS PrivateLink dans le AWS PrivateLink Guide.

Rubriques

- Création d'un point de terminaison d'un VPC d'interface
- Créer une politique de point de terminaison de VPC

Création d'un point de terminaison d'un VPC d'interface

Création d'un point de terminaison pour Amazon EC2 Auto Scaling à l'aide du nom de service suivant :

```
com.amazonaws.region.autoscaling
```

Pour plus d'informations, consultez la section Accès à un AWS service à l'aide d'un point de terminaison VPC d'interface dans le AWS PrivateLink Guide.

Vous n'avez pas besoin de modifier les paramètres Amazon EC2 Auto Scaling. Amazon EC2 Auto Scaling appelle d'autres AWS services en utilisant soit des points de terminaison de service, soit des points de terminaison VPC d'interface privée, selon ceux utilisés.

Créer une politique de point de terminaison de VPC

Vous pouvez attacher une politique à votre point de terminaison de VPC pour contrôler l'accès à l'API Amazon EC2 Auto Scaling. La politique spécifie :

- Le principal qui peut exécuter des actions.
- Les actions qui peuvent être effectuées.
- La ressource sur laquelle les actions peuvent être effectuées.

L'exemple suivant montre une politique de point de terminaison de VPC qui refuse à tout le monde l'autorisation de supprimer une politique de dimensionnement via le point de terminaison. L'exemple de politique accorde également à tout le monde l'autorisation d'effectuer toutes les autres actions.

```
{
   "Statement": [
```

```
        {
            "Action": "*",
            "Effect": "Allow",
            "Resource": "*",
            "Principal": "*"
        },
        {
            "Action": "autoscaling:DeleteScalingPolicy",
            "Effect": "Deny",
            "Resource": "*",
            "Principal": "*"
        }
    ]
}
```

Pour plus d'informations, consultez la section Contrôler l'accès aux points de terminaison VPC à l'aide des politiques relatives aux points de terminaison dans le Guide.AWS PrivateLink

Résoudre les problèmes d'Amazon EC2 Auto Scaling

Amazon EC2 Auto Scaling fournit des erreurs spécifiques et descriptives pour vous aider à résoudre les problèmes. Vous trouverez les messages d'erreur dans la description des activités de mise à l'échelle.

Rubriques

Récupérer un message d'erreur à partir d'activités de mise à l'échelle

Pour récupérer un message d'erreur à partir de la description des activités de mise à l'échelle, utilisez la commande describe-scaling-activities. Vous avez un enregistrement des activités de mise à l'échelle qui remonte à 6 semaines. Les activités de mise à l'échelle sont classées par heure de début, les dernières activités de mise à l'échelle étant répertoriées en premier.

> **Note**
>
> Les activités de mise à l'échelle sont également affichées dans la console Amazon EC2 Auto Scaling, dans l'onglet Activity (Activités), dans l'historique des activités pour le groupe Auto Scaling.

Pour afficher les activités de mise à l'échelle d'un groupe Auto Scaling spécifique, utilisez la commande suivante.

```
aws autoscaling describe-scaling-activities --auto-scaling-group-name my-asg
```

Voici un exemple de réponse, dans lequel `StatusCode` contient le statut actuel de l'activité et `StatusMessage` contient le message d'erreur.

```
{
    "Activities": [
        {
            "ActivityId": "3b05dbf6-037c-b92f-133f-38275269dc0f",
            "AutoScalingGroupName": "my-asg",
            "Description": "Launching a new EC2 instance: i-003a5b3ffe1e9358e.  Status
 Reason: Instance failed to complete user's Lifecycle Action: Lifecycle Action with
 token e85eb647-4fe0-4909-b341-a6c42d8aba1f was abandoned: Lifecycle Action Completed
 with ABANDON Result",
            "Cause": "At 2021-01-11T00:35:52Z a user request created an
 AutoScalingGroup changing the desired capacity from 0 to 1.  At 2021-01-11T00:35:53Z
 an instance was started in response to a difference between desired and actual
 capacity, increasing the capacity from 0 to 1.",
            "StartTime": "2021-01-11T00:35:55.542Z",
            "EndTime": "2021-01-11T01:06:31Z",
            "StatusCode": "Cancelled",
            "StatusMessage": "Instance failed to complete user's Lifecycle Action:
 Lifecycle Action with token e85eb647-4fe0-4909-b341-a6c42d8aba1f was abandoned:
 Lifecycle Action Completed with ABANDON Result",
            "Progress": 100,
            "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":\"us-
west-2b\"...}",
            "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:283179a2-
f3ce-423d-93f6-66bb518232f7:autoScalingGroupName/my-asg"
        },
      ...
    ]
}
```

Pour obtenir une description des champs de la sortie, consultez Activité dans la Référence de l'API Amazon EC2 Auto Scaling.

Pour afficher les activités de mise à l'échelle d'un groupe supprimé

Pour afficher les activités de mise à l'échelle après la suppression du groupe Auto Scaling, ajoutez l'option `--include-deleted-groups` à la commande describe-scaling-activities comme suit.

```
aws autoscaling describe-scaling-activities --auto-scaling-group-name my-asg --include-
deleted-groups
```

Voici un exemple de réponse avec une activité de mise à l'échelle pour un groupe supprimé.

```
{
    "Activities": [
        {
            "ActivityId": "e1f5de0e-f93e-1417-34ac-092a76fba220",
            "AutoScalingGroupName": "my-asg",
            "Description": "Launching a new EC2 instance.  Status Reason: Your Spot
 request price of 0.001 is lower than the minimum required Spot request fulfillment
 price of 0.0031. Launching EC2 instance failed.",
            "Cause": "At 2021-01-13T20:47:24Z a user request update of AutoScalingGroup
 constraints to min: 1, max: 5, desired: 3 changing the desired capacity from 0 to 3.
  At 2021-01-13T20:47:27Z an instance was started in response to a difference between
 desired and actual capacity, increasing the capacity from 0 to 3.",
            "StartTime": "2021-01-13T20:47:30.094Z",
            "EndTime": "2021-01-13T20:47:30Z",
            "StatusCode": "Failed",
            "StatusMessage": "Your Spot request price of 0.001 is lower than the
 minimum required Spot request fulfillment price of 0.0031. Launching EC2 instance
 failed.",
            "Progress": 100,
            "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":\"us-
west-2b\"...}",
            "AutoScalingGroupState": "Deleted",
            "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:283179a2-
f3ce-423d-93f6-66bb518232f7:autoScalingGroupName/my-asg"
        },
     ...
    ]
}
```

Désactiver les activités de dimensionnement

Les options suivantes s'offrent à vous si vous devez étudier un problème sans interférer avec les politiques de dimensionnement ou les actions planifiées :

- Empêchez toutes les politiques de dimensionnement dynamique et les actions planifiées d'apporter des modifications à la capacité souhaitée du groupe en suspendant les `ScheduledActions` processus `AlarmNotification` et. Pour plus d’informations, consultez Suspendre et reprendre les processus Amazon EC2 Auto Scaling.
- Désactivez les politiques de dimensionnement dynamique individuelles afin qu'elles ne modifient pas la capacité souhaitée du groupe en réponse aux changements de charge. Pour plus d’informations, consultez Désactiver une politique de mise à l'échelle pour un groupe Auto Scaling.
- Mettez à jour les politiques de dimensionnement de suivi des cibles individuelles afin de les étendre uniquement (ajouter de la capacité) en désactivant la partie évolutive de la politique. Cette méthode empêche la réduction de la capacité souhaitée par le groupe, mais permet de l'augmenter lorsque la charge augmente. Pour plus d’informations, consultez Politiques de suivi des objectifs et d'échelonnement pour Amazon EC2 Auto Scaling.
- Mettez à jour votre politique de dimensionnement prédictif en mode prévision uniquement. En mode prévisions uniquement, la mise à l'échelle prédictive continuera de générer des prévisions, mais elle n'augmentera pas automatiquement la capacité. Pour plus d’informations, consultez Création d'une politique de dimensionnement prédictive.

Ressources supplémentaires pour la résolution des problèmes

Les pages suivantes fournissent des informations supplémentaires pour la résolution des problèmes liés à Amazon EC2 Auto Scaling.

- Vérifier une activité de mise à l'échelle pour un groupe Auto Scaling
- Afficher des graphiques de surveillance dans la console Amazon EC2 Auto Scaling
- Surveillance de l’état des instances dans un groupe Auto Scaling
- Considérations et restrictions relatives aux hooks de cycle de vie
- Effectuer une action de cycle de vie
- Fournir une connectivité réseau pour vos instances Auto Scaling à l'aide d'Amazon VPC
- Supprimer temporairement des instances du groupe Auto Scaling
- Désactiver une politique de mise à l'échelle pour un groupe Auto Scaling
- Suspendre et reprendre les processus Amazon EC2 Auto Scaling
- Contrôler les instances à scalabilité automatique à résilier pendant une mise à l'échelle horizontale
- Supprimer votre infrastructure Auto Scaling

- Quotas pour les ressources et les groupes Auto Scaling

Les AWS ressources suivantes peuvent également vous être utiles :

- Rubriques relatives à Amazon EC2 Auto Scaling dans le centre de connaissances AWS
- Questions relatives à Amazon EC2 Auto Scaling sur Re:Post AWS
- Articles publiés sur Amazon EC2 Auto Scaling sur le Compute Blog AWS
- Résolution des problèmes CloudFormation dans le guide de AWS CloudFormation l'utilisateur

La résolution des problèmes nécessite souvent une requête et une recherche itératives par un expert ou par une communauté d'assistants. Si vous continuez à rencontrer des problèmes après avoir essayé les suggestions de cette section, contactez AWS Support (dans le AWS Management Console, cliquez sur Support, Support Center) ou posez une question sur AWS Re:post en utilisant le tag Amazon EC2 Auto Scaling.

Dépanner Amazon EC2 Auto Scaling : échecs de lancement d'instance EC2

Cette page fournit des informations sur les instances EC2 dont le lancement échoue, les causes potentielles et les étapes à suivre pour résoudre le problème.

Pour récupérer un message d'erreur, consultez Récupérer un message d'erreur à partir d'activités de mise à l'échelle.

Lorsque des instances EC2 échouent lors du lancement, un ou plusieurs des messages d'erreur suivants peuvent s'afficher :

Problèmes de lancement

- La configuration demandée n'est actuellement pas prise en charge.
- Le groupe de sécurité <nom du groupe de sécurité> n'existe pas. Échec du lancement de l'instance EC2.
- La paire de clés <paire de clés associée à l'instance EC2> n'existe pas. Échec du lancement de l'instance EC2.
- Le type d'instance demandé (<type d'instance>) n'est pas pris en charge dans la zone de disponibilité demandée (<zone de disponibilité de l'instance>)...

- Votre prix de demande Spot de 0,015 est inférieur au prix minimum requis d'exécution de la demande Spot de 0,0735...
- Nom de périphérique non valide <nom de périphérique> / Chargement de nom de périphérique non valide. Échec du lancement de l'instance EC2.
- La valeur (<nom associé au périphérique de stockage de l'instance>) pour le paramètre virtualName n'est pas valide... Échec du lancement de l'instance EC2.
- Les mappages de périphérique de stockage en mode bloc EBS ne sont pas pris en charge pour les AMI de stockage d'instance.
- Les groupes de placement ne peuvent pas être utilisés avec des instances de type <type d'instance>. Échec du lancement de l'instance EC2.
- Cliente. InternalError: erreur du client au lancement.
- Nous ne possédons actuellement pas suffisamment de capacité <type d'instance> dans la zone de disponibilité que vous avez demandée… Échec du lancement de l'instance EC2.
- La réservation demandée ne dispose pas d'une capacité compatible et disponible suffisante pour cette demande. Échec du lancement de l'instance EC2.
- Votre réservation de bloc de capacité <ID réserve> n'est pas encore active. Échec du lancement de l'instance EC2.
- Il n'y a pas de capacité ponctuelle disponible qui correspond à votre demande. Échec du lancement de l'instance EC2.
- <nombre d'instances> instance(s) sont déjà en cours d'exécution. Échec du lancement de l'instance EC2.

La configuration demandée n'est actuellement pas prise en charge.

Cause : Certaines options de votre modèle de lancement ou de votre configuration de lancement ne sont peut-être pas compatibles avec le type d'instance, ou la configuration de l'instance n'est peut-être pas prise en charge dans AWS la région ou les zones de disponibilité que vous avez demandées.

Solution : essayez une autre configuration d'instance. Pour rechercher un type d'instance répondant à vos besoins, consultez la section Trouver un type d'instance Amazon EC2 dans le guide de l'utilisateur Amazon EC2.

Pour plus d'informations sur la solution de ce problème, consultez les informations suivantes :

- Assurez-vous que vous avez choisi une AMI qui est prise en charge par votre type d'instance. Par exemple, si le type d'instance utilise un processeur AWS Graviton basé sur ARM au lieu d'un processeur Intel Xeon, vous avez besoin d'une AMI compatible ARM. Pour plus d'informations sur le choix d'un type d'instance compatible, consultez la section Compatibilité pour la modification du type d'instance dans le guide de l'utilisateur Amazon EC2.
- Vérifiez que le type d'instance est disponible dans vos régions et zones de disponibilité demandées. Les types d'instance de la nouvelle génération peuvent ne pas encore être disponibles dans une région ou une zone de disponibilité donnée. Les types d'instance de génération ancienne peuvent ne pas être disponibles dans les régions et zones de disponibilité les plus récentes. Pour rechercher les types d'instance offerts par emplacement (région ou zone de disponibilité), utilisez la commande describe-instance-type-offerings. Pour plus d'informations, consultez la section Trouver un type d'instance Amazon EC2 dans le guide de l'utilisateur Amazon EC2.
- Si vous utilisez des instances dédiées ou des hôtes dédiés, assurez-vous que vous avez choisi un type d'instance pris en charge en tant qu'Instance dédiée ou Hôte dédié.

Le groupe de sécurité <nom du groupe de sécurité> n'existe pas. Échec du lancement de l'instance EC2.

Cause : le groupe de sécurité spécifié dans le modèle de lancement ou la configuration du lancement peut avoir été supprimé.

Solution :

1. Utilisez la commande describe-security-groups pour obtenir la liste des groupes de sécurité associés au compte.
2. Dans la liste, sélectionnez le groupe de sécurité à utiliser. Pour créer un groupe de sécurité à la place, utilisez la commande create-security-group.
3. Créez un modèle de lancement ou une configuration du lancement.
4. Mettez à jour votre groupe Auto Scaling avec le nouveau modèle de lancement ou la nouvelle configuration du lancement à l'aide de la commande update-auto-scaling-group.

La paire de clés <paire de clés associée à l'instance EC2> n'existe pas. Échec du lancement de l'instance EC2.

Cause : la paire de clés utilisée lors du lancement de l'instance peut avoir été supprimée.

Solution :

1. Utilisez la commande describe-key-pairs pour obtenir la liste des paires de clés disponibles pour vous.
2. Dans la liste, sélectionnez la paire de clés à utiliser. Pour créer une paire de clés à la place, utilisez la commande create-key-pair.
3. Créez un modèle de lancement ou une configuration du lancement.
4. Mettez à jour votre groupe Auto Scaling avec le nouveau modèle de lancement ou la nouvelle configuration du lancement à l'aide de la commande update-auto-scaling-group.

Le type d'instance demandé (<type d'instance>) n'est pas pris en charge dans la zone de disponibilité demandée (<zone de disponibilité de l'instance>)...

Message d’erreur : le type d’instance demandé (<type d’instance>) n’est pas pris en charge dans la zone de disponibilité demandée (<zone de disponibilité de l’instance>)…Échec du lancement de l’instance EC2.

Cause : les zones de disponibilité indiquées dans le groupe Auto Scaling ne prennent pas en charge le type d’instance que vous avez choisi.

Solution :

1. Vérifiez quelles zones de disponibilité prennent en charge le type d’instance que vous avez choisi à l’aide de la commande describe-instance-type-offerings ou depuis la console Amazon EC2 en vérifiant la valeur des zones de disponibilité dans le volet réseau de la page Types d’instances.
2. Mettez à jour ou supprimez le sous-réseau pour toutes les zones non prises en charge dans les paramètres de votre groupe Auto Scaling à l’aide de la commande update-auto-scaling-group. Pour plus d’informations, consultez Ajouter et supprimer des zones de disponibilité.

Votre prix de demande Spot de 0,015 est inférieur au prix minimum requis d'exécution de la demande Spot de 0,0735...

Cause : le prix Spot maximum de votre demande est inférieur au prix Spot pour le type d'instance que vous avez sélectionné.

Solution : envoyez une nouvelle demande avec un prix Spot maximum plus élevé (éventuellement le prix à la demande). Auparavant, le prix Spot que vous aviez payé était basé sur des enchères. Aujourd'hui, vous payez le prix Spot actuel. En fixant votre prix maximum plus élevé, cela donne au service Amazon EC2 Spot plus de chances de lancer et de maintenir la quantité de capacité requise.

Nom de périphérique non valide <nom de périphérique> / Chargement de nom de périphérique non valide. Échec du lancement de l'instance EC2.

Cause 1 : les mappages de périphérique de stockage en mode bloc dans le modèle de lancement ou la configuration du lancement peuvent contenir des noms de périphérique de stockage en mode bloc indisponibles ou non pris en charge actuellement.

Solution :

1. Vérifiez quels noms de périphériques sont disponibles pour votre configuration d'instance spécifique. Pour plus de détails sur la dénomination des appareils, consultez la section Noms des appareils sur les instances Linux dans le guide de l'utilisateur Amazon EC2.
2. Créez manuellement une instance Amazon EC2 qui ne fait pas partie du groupe Auto Scaling et examinez le problème. Si la configuration du nommage des périphériques de stockage en mode bloc entre en conflit avec les noms d'Amazon Machine Image (AMI), l'instance échouera lors du lancement. Pour plus d'informations, consultez Bloquer les mappages d'appareils dans le guide de l'utilisateur Amazon EC2.
3. Après avoir confirmé que votre instance a été lancée avec succès, utilisez la commande describe-volumes pour voir comment les volumes sont exposés à l'instance.
4. Créez un modèle de lancement ou une configuration du lancement avec le nom du périphérique répertorié dans la description du volume.
5. Mettez à jour votre groupe Auto Scaling avec le nouveau modèle de lancement ou la nouvelle configuration du lancement à l'aide de la commande update-auto-scaling-group.

La valeur (<nom associé au périphérique de stockage de l'instance>) pour le paramètre virtualName n'est pas valide... Échec du lancement de l'instance EC2.

Cause : le format spécifié pour le nom virtuel associé au périphérique de stockage en mode bloc est incorrect.

Solution :

1. Créez un modèle de lancement ou une configuration du lancement en spécifiant le nom du périphérique dans le paramètre `virtualName`. Pour plus d'informations sur le format du nom de l'appareil, consultez la section Dénomination des appareils sur les instances Linux dans le guide de l'utilisateur Amazon EC2.
2. Mettez à jour votre groupe Auto Scaling avec le nouveau modèle de lancement ou la nouvelle configuration du lancement à l'aide de la commande update-auto-scaling-group.

Les mappages de périphérique de stockage en mode bloc EBS ne sont pas pris en charge pour les AMI de stockage d'instance.

Cause : les mappages de périphérique de stockage en mode bloc spécifiés dans le modèle de lancement ou la configuration du lancement ne sont pas pris en charge sur l'instance.

Solution :

1. Créez un modèle de lancement ou une configuration du lancement avec des mappages de périphérique de stockage en mode bloc pris en charge par le type d'instance. Pour plus d'informations, consultez la section Bloquer le mappage des appareils dans le guide de l'utilisateur Amazon EC2.
2. Mettez à jour votre groupe Auto Scaling avec le nouveau modèle de lancement ou la nouvelle configuration du lancement à l'aide de la commande update-auto-scaling-group.

Les groupes de placement ne peuvent pas être utilisés avec des instances de type <type d'instance>. Échec du lancement de l'instance EC2.

Cause : le groupe de placement du cluster contient un type d'instance qui n'est pas valide.

Solution :

1. Pour plus d'informations sur les types d'instances valides pris en charge par les groupes de placement, consultez la section Groupes de placement dans le guide de l'utilisateur Amazon EC2.
2. Suivez les instructions détaillées dans Groupes de placement pour créer un groupe de placement.
3. Sinon, créez un modèle de lancement ou une configuration du lancement avec le type d'instance pris en charge.

4. Mettez à jour le groupe Auto Scaling avec un nouveau groupe de placement, un nouveau modèle de lancement ou une nouvelle configuration du lancement grâce à la commande update-auto-scaling-group.

Cliente. InternalError: erreur du client au lancement.

Problème : Amazon EC2 Auto Scaling essaie de lancer une instance dotée d'un volume EBS chiffré, mais le rôle lié au service n'a pas accès à la clé gérée par le AWS KMS client utilisée pour le chiffrer. Pour plus d'informations, consultez Politique de AWS KMS clé requise pour une utilisation avec des volumes chiffrés.

Cause 1 : vous avez besoin d'une politique de clé qui autorise l'utilisation de la clé gérée par le client pour le rôle lié au service approprié.

Solution 1 : autorisez le rôle lié à un service à utiliser la clé gérée par le client de la manière suivante :

1. Déterminez quel rôle lié à un service utiliser pour ce groupe Auto Scaling.
2. Mettez à jour la politique de clé sur la clé gérée par le client et autorisez le rôle lié à un service à utiliser la clé gérée par le client.
3. Mettez à jour le groupe Auto Scaling afin qu'il utilise le rôle lié à un service.

Pour obtenir un exemple de politique de clé qui permet au rôle lié au service d'utiliser la clé gérée par le client, voir Exemple 1 : sections de la politique de clé qui autorisent l'accès à la clé gérée par le client.

Cause 2 : Si la clé gérée par le client et le groupe Auto Scaling se trouvent dans AWS des comptes différents, vous devez configurer l'accès entre comptes à la clé gérée par le client afin d'autoriser l'utilisation de la clé gérée par le client au rôle lié au service approprié.

Solution 2 : autorisez le rôle lié au service dans le compte externe à utiliser la clé gérée par le client dans le compte local comme suit :

1. Mettez à jour la politique de clé sur la clé gérée par le client pour autoriser le compte de groupe Auto Scaling à accéder à la clé gérée par le client.
2. Définissez un rôle ou un utilisateur IAM; dans le compte du groupe Auto Scaling capable de créer un octroi.
3. Déterminez quel rôle lié à un service utiliser pour ce groupe Auto Scaling.

4. Créez un octroi à la clé gérée par le client avec le rôle lié à un service en tant que principal bénéficiaire.
5. Mettez à jour le groupe Auto Scaling afin qu'il utilise le rôle lié à un service.

Pour plus d'informations, consultez Exemple 2 : sections de la politique de clé autorisant l'accès entre comptes à la clé gérée par le client.

Solution 3 : utilisez une clé gérée par le client dans le même compte AWS que le groupe Auto Scaling.

1. Copiez et rechiffrez l'instantané avec une autre clé gérée par le client qui appartient au même compte que le groupe Auto Scaling.
2. Autorisez le rôle lié à un service à utiliser la clé gérée par le client. Consultez les étapes pour obtenir la solution 1.

Nous ne possédons actuellement pas suffisamment de capacité <type d'instance> dans la zone de disponibilité que vous avez demandée… Échec du lancement de l'instance EC2.

Message d'erreur : nous ne possédons actuellement pas suffisamment de capacité <type d'instance> dans la zone de disponibilité demandée (<zone de disponibilité demandée>). Le système s'occupera d'allouer de la capacité supplémentaire. Vous pouvez actuellement bénéficier de la capacité <type d'instance> en ne spécifiant pas de zone de disponibilité dans la demande ou en choisissant <liste des zones de disponibilité qui prennent actuellement en charge le type d'instance>. Échec du lancement de l'instance EC2.

Cause : la combinaison type d'instance et zone de disponibilité demandée n'est pas prise en charge pour le moment.

Solution : pour résoudre le problème, essayez ce qui suit :

- Attendez quelques minutes qu'Amazon EC2 Auto Scaling trouve de la capacité pour ce type d'instance dans d'autres zones de disponibilité activées.
- Étendez votre groupe Auto Scaling à d'autres zones de disponibilité. Pour plus d'informations, consultez Ajouter et supprimer des zones de disponibilité.

- Suivez les meilleures pratiques d'utilisation d'un ensemble diversifié de types d'instance afin de ne pas dépendre d'un type d'instance spécifique. Pour plus d'informations, consultez Groupes Auto Scaling combinant plusieurs types d'instances et options d'achat.

La réservation demandée ne dispose pas d'une capacité compatible et disponible suffisante pour cette demande. Échec du lancement de l'instance EC2.

Cause 1 : vous avez atteint la limite du nombre d'instances que vous pouvez lancer avec une réserve de capacité à la demande `targeted`.

Solution 1 : augmentez le nombre d'instances que vous pouvez lancer avec la réserve de capacité à la demande `targeted` ou utilisez un groupe de réserves de capacité afin que tout ce qui dépasse la capacité réservée soit lancé en tant que capacité à la demande normale. Pour plus d'informations, consultez Utilisez les réserves de capacité à la demande pour réserver de la capacité dans des zones de disponibilité spécifiques.

Cause 2 : vous avez atteint la limite du nombre d'instances que vous pouvez lancer dans un bloc de capacité.

Avec les blocs de capacité, vous êtes limité par la quantité de capacité initialement achetée. Si le nombre de lancements est plus élevé que prévu et que vous utilisez toute la capacité disponible, cela entraîne l'échec des lancements. Les instances résiliées sont soumises à un long processus de nettoyage avant d'être complètement résiliées. Pendant ce temps, ils ne peuvent pas être réutilisés. Cela peut également provoquer d'échec des lancements. Pour plus d'informations, consultez Utilisation Capacity Blocks pour les charges de travail liées à l'apprentissage automatique.

Solution 2 : pour résoudre le problème, essayez ce qui suit :

- Conservez la demande telle quelle. Si une instance Capacity Block prend fin, vous devez attendre plusieurs minutes pour que l'instance finisse de s'arrêter et que la capacité soit de nouveau disponible. Amazon EC2 Auto Scaling continue à effectuer automatiquement la demande de lancement jusqu'à ce que la capacité devienne disponible.
- Assurez-vous d'acheter une capacité suffisante pour faire face à votre charge de travail maximale afin de ne pas rencontrer cette erreur fréquemment.

Votre réservation de bloc de capacité <ID réserve> n’est pas encore active. Échec du lancement de l'instance EC2.

Cause : le bloc de capacité indiqué n’est pas encore actif.

Solution : suivez l’approche recommandée pour les blocs de capacité et utilisez le dimensionnement planifié. Cela vous permet de vous assurer d’augmenter la capacité souhaitée de votre groupe Auto Scaling uniquement lorsque la réservation est active et de la diminuer avant la fin de la réservation.

Il n'y a pas de capacité ponctuelle disponible qui correspond à votre demande. Échec du lancement de l'instance EC2.

Cause : à l'heure actuelle, il n'y a pas assez de capacité de rechange pour répondre à votre demande d'instances Spot.

Solution : pour résoudre le problème, essayez ce qui suit :

- Attendez quelques minutes, car la capacité peut changer fréquemment. Amazon EC2 Auto Scaling continue à effectuer automatiquement la demande de lancement jusqu’à ce que la capacité devienne disponible.
- Étendez votre groupe Auto Scaling à d’autres zones de disponibilité. Pour plus d’informations, consultez Ajouter et supprimer des zones de disponibilité.
- Suivez les meilleures pratiques d’utilisation d’un ensemble diversifié de types d’instance afin de ne pas dépendre d’un type d’instance spécifique. Pour plus d’informations, consultez Groupes Auto Scaling combinant plusieurs types d'instances et options d'achat.

<nombre d'instances> instance(s) sont déjà en cours d'exécution. Échec du lancement de l'instance EC2.

Cause : vous avez atteint la limite du nombre d'instances que vous pouvez lancer dans une région. Lorsque vous créez votre AWS compte, nous fixons des limites par défaut quant au nombre d'instances que vous pouvez exécuter par région.

Solution : pour résoudre le problème, essayez ce qui suit :

- Vos limites actuelles ne sont pas adaptées à vos besoins, vous pouvez demander une augmentation des quotas par région. Pour plus d'informations, consultez les quotas de service Amazon EC2 dans le guide de l'utilisateur Amazon EC2.

- Envoyez une nouvelle demande avec un nombre réduit d'instances (que vous pouvez augmenter à un stade ultérieur).

Résoudre les problèmes d'Amazon EC2 Auto Scaling : AMI

Cette page fournit des informations sur les problèmes associés aux AMI, les causes potentielles et les étapes à suivre pour résoudre le problème.

Pour récupérer un message d'erreur, consultez Récupérer un message d'erreur à partir d'activités de mise à l'échelle.

Lorsque des instances EC2 échouent lors du lancement à cause de problèmes avec l'AMI, un ou plusieurs des messages d'erreur suivants peuvent s'afficher.

Problèmes AMI

- L'ID d'AMI <ID de l'AMI> n'existe pas. Échec du lancement de l'instance EC2.
- L'AMI <ID d'AMI> est en attente et ne peut pas être exécutée. Échec du lancement de l'instance EC2.
- Nom de périphérique non valide <nom périphérique>. Échec du lancement de l'instance EC2.
- L’architecture « arm64 » du type d’instance indiqué ne correspond pas à l’architecture « x86_64 » de l’AMI indiquée... Le lancement de l’instance EC2 a échoué.
- L’AMI <ID AMI> est en attente et ne peut pas être exécutée. Échec du lancement de l'instance EC2.

> **Important**
>
> AWS permet de partager une AMI en privé avec un autre AWS compte en modifiant les autorisations de l'AMI. Si une AMI est rendue privée sans être partagée, cela peut entraîner une erreur d’autorisation lors du lancement de nouvelles instances. Pour plus d'informations sur le partage d'AMI privées, consultez la section Partager une AMI avec des AWS comptes spécifiques dans le guide de l'utilisateur Amazon EC2.

L'ID d'AMI <ID de l'AMI> n'existe pas. Échec du lancement de l'instance EC2.

- Cause : l'AMI peut avoir été supprimée après la création du modèle de lancement ou de la configuration du lancement.
- Solution :
  1. Créez un modèle de lancement ou une configuration du lancement avec une AMI valide.
  2. Mettez à jour votre groupe Auto Scaling avec le nouveau modèle de lancement ou la nouvelle configuration du lancement à l'aide de la commande update-auto-scaling-group.

L'AMI <ID d'AMI> est en attente et ne peut pas être exécutée. Échec du lancement de l'instance EC2.

Cause : vous venez peut-être de créer l'AMI (en prenant un instantané d'une instance en cours d'exécution ou de toute autre façon), et elle peut ne pas être encore disponible.

Solution : vous devez attendre que l'AMI soit disponible pour ensuite créer le modèle de lancement ou la configuration du lancement.

Nom de périphérique non valide <nom périphérique>. Échec du lancement de l'instance EC2.

Cause : lorsque vous attachez un volume EBS à une instance EC2, vous devez fournir un nom de périphérique valide pour le volume. L'AMI sélectionnée doit prendre en charge ce nom de périphérique.

Solution :

1. Créez un modèle de lancement ou une configuration du lancement et spécifiez correctement le nom du périphérique pour l'AMI. La convention de dénomination recommandée varie en fonction du type de virtualisation de l'AMI. Pour plus d'informations, consultez la section Noms des appareils dans le guide de l'utilisateur Amazon EC2.
2. Mettez à jour votre groupe Auto Scaling avec le nouveau modèle de lancement ou la nouvelle configuration du lancement à l'aide de la commande update-auto-scaling-group.

L’architecture « arm64 » du type d’instance indiqué ne correspond pas à l’architecture « x86_64 » de l’AMI indiquée... Le lancement de l’instance EC2 a échoué.

Cause 1 : si l’architecture de l’AMI et le type d’instance utilisé dans votre modèle de lancement ou votre configuration de lancement ne sont pas identiques, une erreur s’affiche lorsqu’Amazon EC2 Auto Scaling tente de lancer une instance à l’aide de la configuration d’instance incompatible.

Solution 1 :

1. Vérifiez l’architecture de votre AMI à l’aide de la commande describe-images ou à partir de la console Amazon EC2 en consultant la valeur Architecture dans le volet de détails de la page Amazon Machine Images (AMI).
2. Trouvez un type d’instance ayant la même architecture que votre AMI à l’aide de la commande describe-instance-types ou depuis la console Amazon EC2 en consultant la colonne Architecture de l’écran Types d’instances. Pour plus d'informations sur le choix d'un type d'instance compatible, consultez la section Compatibilité pour la modification du type d'instance dans le guide de l'utilisateur Amazon EC2.
3. Créez un nouveau modèle de lancement ou une configuration de lancement avec un type d’instance doté de la même architecture que celle de votre AMI.
4. Mettez à jour votre groupe Auto Scaling avec le nouveau modèle de lancement ou la nouvelle configuration du lancement à l'aide de la commande update-auto-scaling-group.

Cause 2 : Amazon EC2 Auto Scaling essaye de lancer un type d’instance indiqué dans la politique relative aux instances mixtes de votre groupe Auto Scaling, mais le type d’instance n’a pas la même architecture que l’AMI indiquée dans votre modèle de lancement.

Solution 1 : n’ajoutez pas les types d’instance dotés d’architectures différentes dans votre politique d’instances mixtes.

1. Vérifiez l’architecture de votre AMI à l’aide de la commande describe-images ou à partir de la console Amazon EC2 en consultant la valeur Architecture dans le volet de détails de la page Amazon Machine Images (AMI).
2. Vérifiez l’architecture de chaque type d’instance que vous avez l’intention d’ajouter dans votre politique d’instances mixtes à l’aide de la commande describe-instance-types ou depuis la console Amazon EC2 en consultant la colonne Architecture de l’écran Types d’instances. Pour plus

d'informations sur le choix des types d'instances compatibles, consultez la section Compatibilité pour la modification du type d'instance dans le guide de l'utilisateur Amazon EC2.

3. Mettez à jour ou supprimez les types d'instance incompatibles de votre groupe Auto Scaling à l'aide de la commande update-auto-scaling-group.

Solution 2 : pour lancer des instances Arm (Graviton2) et x86_64 (Intel) dans le même groupe Auto Scaling, vous devez utiliser des modèles de lancement pris en charge par une AMI compatible Arm et une AMI compatible Intel x86, respectivement, pour correspondre aux types d'instances définis dans votre politique d'instances mixtes.

1. Vérifiez l'architecture de l'AMI dans votre modèle de lancement existant à l'aide de la commande describe-images ou à partir de la console Amazon EC2 en consultant la valeur Architecture dans le volet de détails de la page Amazon Machine Images (AMI).
2. Créez un nouveau modèle de lancement à l'aide d'une AMI qui correspond à l'autre architecture que vous souhaitez utiliser.
3. Mettez à jour votre groupe Auto Scaling pour qu'il remplace le modèle de lancement existant et indiquez le nouveau modèle de lancement pour chaque type d'instance compatible à l'aide de la commande update-auto-scaling-group. Pour plus d'informations, consultez Utiliser un modèle de lancement différent pour un type d'instance.

L'AMI <ID AMI> est en attente et ne peut pas être exécutée. Échec du lancement de l'instance EC2.

Cause : vous essayez de lancer des instances à partir d'une AMI qui a été désactivée. Pour plus d'informations, consultez la section Désactiver une AMI dans le guide de l'utilisateur Amazon EC2.

Solution :

1. Créez un modèle de lancement ou une configuration de lancement et indiquez une AMI non désactivée.
2. Mettez à jour votre groupe Auto Scaling avec le nouveau modèle de lancement ou la nouvelle configuration du lancement à l'aide de la commande update-auto-scaling-group.

Résoudre les problèmes Amazon EC2 Auto Scaling : équilibreur de charge

Cette page fournit des informations sur les problèmes causés par l'équilibreur de charge associé au groupe Auto Scaling, les causes potentielles et les étapes à suivre pour résoudre le problème.

Pour récupérer un message d'erreur, consultez Récupérer un message d'erreur à partir d'activités de mise à l'échelle.

Lorsque des instances EC2 échouent lors du lancement à cause de problèmes avec l'équilibreur de charge associé au groupe Auto Scaling, un ou plusieurs des messages d'erreur suivants peuvent s'afficher.

Problèmes d'équilibreur de charge

- Un ou plusieurs groupes cibles introuvables. Échec de la validation de la configuration de l'équilibreur de charge.
- Impossible de trouver Load Balancer <your load balancer>. Échec de la validation de la configuration de l'équilibreur de charge.
- Il n'existe aucun équilibreur de charge ACTIF nommé <nom de l'équilibreur de charge>. Échec de la mise à jour de la configuration de l'équilibreur de charge.
- L'instance EC2 <ID d'instance> ne se trouve pas dans le VPC. Échec de la mise à jour de la configuration de l'équilibreur de charge.

> **Note**
>
> Vous pouvez utiliser Reachability Analyzer pour résoudre les problèmes de connectivité en vérifiant si les instances de votre groupe Auto Scaling sont accessibles via l'équilibreur de charge. Pour en savoir plus sur les différents problèmes de mauvaise configuration du réseau qui sont automatiquement détectés par Reachability Analyzer, consultez les codes d'explication de Reachability Analyzer dans le Guide de l'utilisateur de Reachability Analyzer.

Un ou plusieurs groupes cibles introuvables. Échec de la validation de la configuration de l'équilibreur de charge.

Problème : lorsque votre groupe Auto Scaling lance des instances, Amazon EC2 Auto Scaling essaye de vérifier que les ressources Elastic Load Balancing associées au groupe Auto Scaling existent. Lorsqu'un groupe cible est introuvable, l'activité de mise à l'échelle échoue et vous obtenez l'erreur `One or more target groups not found. Validating load balancer configuration failed.`.

Cause 1 : un groupe cible associé à votre groupe Auto Scaling a été supprimé.

Solution : vous pouvez créer un groupe Auto Scaling sans groupe cible ou supprimer le groupe cible non utilisé du groupe Auto Scaling en utilisant la console Amazon EC2 Auto Scaling ou la commande detach-load-balancer-target-groups.

Cause 2 : le groupe cible existe, mais un problème est survenu lors de la tentative de spécification de l'ARN du groupe cible pendant la création du groupe Auto Scaling. Les ressources ne sont pas créées dans le bon ordre.

Solution 2 : créez un groupe Auto Scaling et spécifiez le groupe cible à la fin.

Impossible de trouver Load Balancer <your load balancer>. Échec de la validation de la configuration de l'équilibreur de charge.

Problème : lorsque votre groupe Auto Scaling lance des instances, Amazon EC2 Auto Scaling essaye de vérifier que les ressources Elastic Load Balancing associées au groupe Auto Scaling existent. Lorsqu'un Classic Load Balancer est introuvable, l'activité de mise à l'échelle échoue et vous obtenez l'erreur `Cannot find Load Balancer <your load balancer>. Validating load balancer configuration failed.`.

Cause 1 : le Classic Load Balancer a été supprimé.

Solution 1 : vous pouvez créer un groupe Auto Scaling sans l'équilibreur de charge ou supprimer l'équilibreur de charge non utilisé du groupe Auto Scaling en utilisant la console Amazon EC2 Auto Scaling ou la commande detach-load-balancers.

Cause 2 : le Classic Load Balancer existe, mais il y a eu un problème lors de la spécification du nom de l'équilibreur de charge pendant la création du groupe Auto Scaling. Les ressources ne sont pas créées dans le bon ordre.

Solution 2 : créez un groupe Auto Scaling et spécifiez le nom de l'équilibreur de charge à la fin.

Il n'existe aucun équilibreur de charge ACTIF nommé <nom de l'équilibreur de charge>. Échec de la mise à jour de la configuration de l'équilibreur de charge.

Cause : l'équilibreur de charge spécifié peut avoir été supprimé.

Solution : vous pouvez créer un équilibreur de charge et ensuite créer un nouveau groupe Auto Scaling ou créer un groupe Auto Scaling sans équilibreur de charge.

L'instance EC2 <ID d'instance> ne se trouve pas dans le VPC. Échec de la mise à jour de la configuration de l'équilibreur de charge.

Cause : l'instance spécifiée n'existe pas dans le VPC.

Solution : vous pouvez supprimer l'équilibreur de charge associé à l'instance ou créer un groupe Auto Scaling.

Résoudre les problèmes d'Amazon EC2 Auto Scaling : modèles de lancement

Utilisez les informations suivantes pour identifier et résoudre les problèmes courants que vous pouvez rencontrer lorsque vous essayez de spécifier un modèle de lancement avec votre groupe Auto Scaling.

Impossible de lancer des instances

Si vous ne parvenez pas à lancer des instances avec un modèle de lancement déjà spécifié, vérifiez les points suivants pour un dépannage général :Dépanner Amazon EC2 Auto Scaling : échecs de lancement d'instance EC2.

Vous devez utiliser un modèle de lancement complet valide (valeur non valide)

Problème: lorsque vous essayez de spécifier un modèle de lancement pour un groupe Auto Scaling, vous obtenez l'erreur `You must use a valid fully-formed launch template`. Vous

pouvez rencontrer cette erreur car les valeurs du modèle de lancement sont uniquement vérifiées lorsqu'un groupe Auto Scaling utilisant le modèle de lancement est créé ou mis à jour.

Cause 1: si vous recevez une erreur `You must use a valid fully-formed launch template`, il existe des problèmes qui font qu'Amazon EC2 Auto Scaling considère que le modèle de lancement n'est pas valide. Il s'agit d'une erreur générique qui peut avoir plusieurs causes différentes.

Solution 1: essayez les étapes suivantes pour résoudre les problèmes :

1. Faites attention à la deuxième partie du message d'erreur pour obtenir plus d'informations. À la suite d'erreur `You must use a valid fully-formed launch template`, consultez le message d'erreur plus spécifique qui identifie le problème que vous devrez résoudre.
2. Si vous ne parvenez pas à en trouver la cause, testez votre modèle de lancement avec la commande run-instances. Utilisez l'option `--dry-run`, comme indiqué dans l'exemple suivant. Cela vous permet de reproduire le problème et de fournir des informations sur sa cause.

   ```
   aws ec2 run-instances --launch-template LaunchTemplateName=my-template,Version='1' --
   dry-run
   ```

3. Si une valeur n'est pas valide, vérifiez que la ressource spécifiée existe et qu'elle est correcte. Par exemple, lorsque vous spécifiez une paire de clés Amazon EC2, la ressource doit exister dans votre compte et dans la région dans laquelle vous créez ou mettez à jour votre groupe Auto Scaling.
4. Si les informations attendues sont manquantes, vérifiez vos paramètres et ajustez le modèle de lancement selon vos besoins.
5. Après avoir effectué vos modifications, exécutez à nouveau la commande run-instances avec l'option `--dry-run` pour vérifier que votre modèle de lancement utilise des valeurs valides.

Pour plus d'informations, consultez Créer un modèle de lancement pour un groupe Auto Scaling.

Vous n'êtes pas autorisé à utiliser le modèle de lancement (autorisations insuffisantes)

Problème: lorsque vous essayez de spécifier un modèle de lancement pour un groupe Auto Scaling, vous obtenez l'erreur `You are not authorized to use launch template`.

Cause 1 : si vous essayez d'utiliser un modèle de lancement et que les informations d'identification IAM que vous utilisez ne disposent pas des autorisations suffisantes, vous recevez une erreur indiquant que vous n'êtes pas autorisé à utiliser le modèle de lancement.

Solution 1 : pour résoudre le problème, essayez ce qui suit :

- Vérifiez que les Informations d'identification IAM que vous utilisez pour effectuer la demande disposent des autorisations pour appeler les actions de l'API EC2 dont vous avez besoin, y compris l'action `ec2:RunInstances`. Si vous avez spécifié des balises dans votre modèle de lancement, vous devez également disposer d'une autorisation pour utiliser l'action `ec2:CreateTags`.
- Vous pouvez également vérifier que les informations d'identification IAM que vous utilisez pour effectuer la demande sont affectées à la stratégie `AmazonEC2FullAccess`. Cette politique AWS gérée accorde un accès complet à toutes les ressources Amazon EC2 et aux services associés, notamment Amazon EC2 CloudWatch Auto Scaling et Elastic Load Balancing.

Pour plus d'informations sur les autorisations requises pour utiliser les modèles de lancement, y compris des exemples de politiques IAM, consultez la section Contrôler l'accès aux modèles de lancement avec des autorisations IAM dans le guide de l'utilisateur Amazon EC2. Pour d'autres exemples de politiques IAM, consultez Support de modèle de lancement.

Cause 2 : si vous tentez d'utiliser un modèle de lancement qui spécifie un profil d'instance, vous devez avoir l'autorisation IAM de transmettre le rôle IAM associé au profil d'instance.

Solution 2 : vérifiez que les informations d'identification IAM que vous utilisez pour effectuer la demande disposent de l'autorisation `iam:PassRole` appropriée pour transmettre le rôle spécifié au service Amazon EC2 Auto Scaling. Pour plus d'informations, et un exemple de politique IAM, consultez Rôle IAM pour les applications qui s'exécutent sur des instances Amazon EC2. Pour d'autres rubriques de dépannage relatives aux profils d'instance, consultez Résolution d'Amazon EC2 et IAM dans le Guide de l'utilisateur IAM.

Cause 3 : Si vous essayez d'utiliser un modèle de lancement qui spécifie une AMI dans un autre Compte AWS, et que l'AMI est privée et n'est pas partagée avec celle que Compte AWS vous utilisez, vous recevez un message d'erreur indiquant que vous n'êtes pas autorisé à utiliser le modèle de lancement.

Solution 3 : vérifiez que les autorisations sur l'AMI incluent le compte que vous utilisez. Pour plus d'informations, consultez la section Partager une AMI avec des informations spécifiques Comptes AWS dans le guide de l'utilisateur Amazon EC2.

Informations connexes

Les ressources connexes suivantes peuvent s'avérer utiles lors de l'utilisation de ce service.

Ressource	Description
Référence d'API Amazon EC2 Auto Scaling	La documentation de chaque opération d'API présente les paramètres de la demande et la réponse XML, et fournit des liens vers des rubriques de référence du kit SDK spécifiques au langage.
autoscaling dans Référence des commandes AWS CLI	Descriptions des AWS CLI commandes que vous pouvez utiliser pour travailler avec les groupes Auto Scaling.
AWS Tools for PowerShell Référence de l'applet de commande	Les AWS outils vous PowerShell permettent de scripter des opérations sur vos AWS ressource s à partir de la ligne de PowerShell commande.
Créer un groupe Auto Scaling avec AWS CloudFormation	La ressource AWS::AutoScaling::AutoScali ngGroup AWS::AutoScaling::AutoScaling vous permet de créer, de modéliser et de gérer vos groupes Auto Scaling sans actions manuelles.
Quotas et points de terminaisons Amazon EC2 Auto Scaling dans le Références générales AWS	Informations sur les régions et les points de terminaison Amazon EC2 Auto Scaling.
Page produit	Page web principale d'informations sur Amazon EC2 Auto Scaling.
AWS Re : Publier	AWS service géré de questions et réponses (Q & A) offrant des réponses participatives et révisées par des experts à vos questions techniques.

Ressource	Description
Création d'une AMI dans le guide de l'utilisateur Amazon EC2	Découvrez comment créer une Amazon Machine Image (AMI) personnalisée à partir d'une instance personnalisée.
Comment se connecter à votre instance Linux dans le guide de l'utilisateur Amazon EC2	Découvrez comment vous connecter aux instances Linux que vous lancez.
Comment se connecter à votre instance Windows dans le guide de l'utilisateur Amazon EC2	Découvrez comment vous connecter aux instances Windows que vous lancez.
Création d'une alarme de facturation pour surveiller vos AWS frais estimés dans le guide de CloudWatch l'utilisateur Amazon	Découvrez comment surveiller vos frais estimés à l'aide de CloudWatch.
Guide de l'utilisateur Application Auto Scaling	Découvrez comment configurer le dimension nement automatique pour des ressources évolutives pour Amazon Web Services au-delà d'Amazon EC2.

Les ressources générales suivantes sont disponibles pour vous aider à en savoir plus sur AWS.

- Cours et ateliers — Liens vers des cours spécialisés et basés sur des rôles, ainsi que des ateliers à votre rythme pour vous aider à perfectionner vos AWS compétences et à acquérir une expérience pratique.
- AWS Centre pour développeurs : découvrez les didacticiels, téléchargez des outils et découvrez les événements AWS destinés aux développeurs.
- AWS Outils de développement : liens vers des outils de développement, des SDK, des boîtes à outils IDE et des outils de ligne de commande pour le développement et la gestion AWS d'applications.
- Centre de ressources pour la mise en route : découvrez comment configurer votre application Compte AWS, rejoindre la AWS communauté et lancer votre première application.
- Tutoriels pratiques — Suivez les step-by-step didacticiels pour lancer votre première application sur AWS.

- AWS Livres blancs : liens vers une liste complète de livres AWS blancs techniques, traitant de sujets tels que l'architecture, la sécurité et l'économie, rédigés par des architectes de AWS solutions ou d'autres experts techniques.
- AWS Support Centre — Le centre de création et de gestion de vos AWS Support dossiers. Comprend également des liens vers d'autres ressources utiles, telles que des forums, des FAQ techniques, l'état de santé du service et AWS Trusted Advisor.
- AWS Support— La principale page Web contenant des informations sur AWS Support un one-on-one canal d'assistance à réponse rapide pour vous aider à créer et à exécuter des applications dans le cloud.
- Contactez-nous : point de contact central pour toute question relative à la facturation AWS , à votre compte, aux événements, à des abus ou à d'autres problèmes.
- AWS Conditions du site — Informations détaillées sur nos droits d'auteur et notre marque commerciale ; votre compte, votre licence et l'accès au site ; et d'autres sujets.

Historique du document

Le tableau ci-dessous décrit les ajouts majeurs à la documentation Amazon EC2 Auto Scaling depuis juillet 2018. Pour recevoir les notifications sur les mises à jour de cette documentation, vous pouvez vous abonner au Flux RSS.

Modification	Description	Date
Mise à jour de sécurité IAM	La politique AutoScalingServiceRolePolicygérée accorde désormais des autorisations supplémentaires à Amazon EC2 (`ec2:GetSecurityGroupsForVpc` etec2:GetInstanceTypesFromInstanceRequirements).	29 février 2024
Hibernation en piscine chaude prise en charge en supplément Régions AWS	Vous pouvez désormais mettre en veille prolongée les instances d'un pool de chaleur dans deux régions supplémentaires : AWS GovCloud (USA Est) et AWS GovCloud (USA Ouest). Pour plus d'informations, consultez Groupes d'instances pré-initialisées pour Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	26 février 2024
Hibernation en piscine chaude prise en charge en supplément Régions AWS	Vous pouvez désormais mettre en veille prolongée des instances dans un pool de chaleur dans deux régions supplémentaires : Europe	21 février 2024

	(Zurich) et Moyen-Orient (Émirats arabes unis). Pour plus d'informations, consultez Groupes d'instances pré-initialisées pour Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	
Support pour l'utilisation de paramètres entre comptes	Vous pouvez désormais utiliser un AWS Systems Manager paramètre partagé par un autre Compte AWS avec Amazon EC2 Auto Scaling. Pour plus d'informations, consultez la section Utiliser des AWS Systems Manager paramètres plutôt que des ID d'AMI dans les modèles de lancement du manuel Amazon EC2 Auto Scaling User Guide.	21 février 2024
Nouvelle option de protection des prix au comptant	Vous pouvez désormais définir votre seuil de protection des prix pour les instances Spot sous forme de pourcentage d'un prix à la demande lorsque vous utilisez la sélection du type d'instance basée sur les attributs. Pour plus d'informations, consultez la section Protection des prix dans le guide de l'utilisateur d'Amazon EC2 Auto Scaling.	29 janvier 2024

Politiques de maintenance des instances	Vous pouvez désormais utiliser une politique de maintenance des instances pour définir si les instances sont lancées avant ou après la résiliation des instances existantes lors d'événements entraînant le remplacement de vos instances, y compris une actualisation d'instance. Pour de plus amples informations, veuillez consulter Politique de maintenance des instances dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	15 novembre 2023
Blocs de capacité pour ML	Vous pouvez désormais lancer des instances dans un bloc de capacité en indiquant l'identifiant de réservation du bloc de capacité lorsque vous créez un modèle de lancement. Grâce aux blocs de capacité, vous pouvez réserver des instances GPU à une date ultérieure pour prendre en charge vos charges de travail de machine learning (ML) de courte durée. Pour plus d'informations, consultez la section Utiliser les blocs de capacité pour les charges de travail d'apprentissage automatique dans le guide de l'utilisateur Amazon EC2 Auto Scaling.	31 octobre 2023

Nouvelles fonctionnalités d'actualisation d'instance	Vous pouvez désormais configurer l'actualisation d'une instance pour définir son statut sur Échec et éventuellement revenir en arrière lorsqu'elle détecte qu'une CloudWatch alarme spécifiée est passée dans `ALARM` cet état. Pour plus d'informations, veuillez Annuler les modifications avec une restauration dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	31 juillet 2023
Modifications du guide	Une nouvelle rubrique sur le lancement des instances à la demande dans les réserves de capacité a été ajoutée au guide. Pour plus d'informations, consultez Utiliser les réserves de capacité à la demande dans des zones de disponibilité spécifiques dans le Guide de l'utilisateur Amazon EC2 Auto Scaling User Guide.	28 juillet 2023

Modifications du guide	Une nouvelle rubrique sur la migration de vos AWS CloudFormation piles depuis les configurations de lancement vers les modèles de lancement a été ajoutée au guide. Pour plus d'informations, consultez Migrer des piles AWS CloudFormation des configurations de lancement aux modèles de lancement dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	18 avril 2023

Assistance pour les nouvelles opérations d'API	Cette version ajoute trois nouvelles opérations d'API : `AttachTrafficSources`, `DetachTrafficSources` et `DescribeTrafficSources`. De plus, un nouveau champ, `TrafficSources`, a été ajouté aux résultats des opérations `DescribeAutoScalingGroups`. Un nouveau statut d'activité, `WaitingForConnectionDraining`, a été ajouté aux résultats des opérations `DescribeScalingActivities`. Amazon EC2 Auto Scaling prend également en charge une nouvelle valeur, `VPC_LATTICE`, pour le champ `HealthCheckType` dans les opérations `CreateAutoScalingGroup`, `UpdateAutoScalingGroup` et `DescribeAutoScalingGroups`. Pour de plus amples informations, veuillez consulter la Référence d'API Amazon EC2 Auto Scaling.	31 mars 2023

Assistance pour Amazon VPC Lattice	Il s’agit de la version de disponibilité générale de VPC Lattice pour Amazon EC2 Auto Scaling. Pour plus d’informations, consultez Acheminer le trafic vers votre groupe Auto Scaling avec un groupe cible VPC Lattice dans le Guide de l’utilisateur Amazon EC2 Auto Scaling.	31 mars 2023
Modifications du guide	La section contenant des AWS CLI exemples d'utilisation d'Elastic Load Balancing inclut désormais des exemples nouveaux et mis à jour. Pour plus d'informations, consultez Examples for work with Elastic Load Balancing with the AWS Command Line Interface (AWS CLI) dans le guide de l'utilisateur Amazon EC2 Auto Scaling.	31 mars 2023
Support supplémentaire pour la mise à l'échelle prédictive Régions AWS	Vous pouvez désormais créer des politiques de dimensionnement prédictif dans les régions du Moyen-Orient (EAU) et AWS GovCloud (USA Est). Pour plus d'informations, consultez Mise à l'échelle prédictive pour Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	16 mars 2023

Nouvelles fonctionnalités d'actualisation d'instance	Vous pouvez désormais choisir de résilier ou d'ignorer les instances en veille et de remplacer ou d'ignorer les instances protégées contre la mise à l'échelle horizontale, au lieu d'attendre qu'elles soient remplaçables. Vous pouvez également restaurer les modifications à la suite d'un échec de l'actualisation d'instance. Dans le cadre de cette mise à jour, la documentation a été étendue pour inclure des rubriques sur la restauration d'une actualisation d'instance, l'annulation d'une actualisation d'instance et la compréhension des valeurs par défaut des paramètres configurables d'une actualisation d'instance. Pour plus d'informations, consultez Remplacement des instances Auto Scaling en fonction d'une actualisation d'instance dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	10 février 2023

Support pour l'utilisation d'un AWS Systems Manager paramètre pour un ID d'AMI	Vous pouvez désormais utiliser un paramètre Systems Manager au lieu d'un ID d'AMI dans votre modèle de lancement. Pour plus d'informations, consultez Utilisation des paramètres AWS Systems Manager à la place des ID d'AMI dans les modèles de lancement dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	19 janvier 2023
Recommandations de mise à l'échelle prédictive	Vous pouvez désormais obtenir des recommandations pour évaluer et choisir la politique de mise à l'échelle prédictive appropriée depuis la console Amazon EC2 Auto Scaling. Pour plus d'informations, consultez Évaluer vos politiques de mise à l'échelle prédictive dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	18 janvier 2023
Prévisions de charge prédictive Scaling	Les prévisions générées par le dimensionnement prédictif sont désormais mises à jour toutes les six heures au lieu de tous les jours. Pour plus d'informations, consultez Mise à l'échelle prédictive pour Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	6 janvier 2023

Support pour les mathématiques CloudWatch métriques	Vous pouvez désormais utiliser une expression mathématique appliquée à une métrique lorsque vous créez des politiques de dimensionnement de suivi des cibles. Avec les mathématiques métriques, vous pouvez interroger plusieurs CloudWatch métriques et utiliser des expressions mathématiques pour créer de nouvelles séries chronologiques basées sur ces métriques. Pour de plus amples informations, veuillez consulter Créer une politique de mise à l'échelle du suivi de cible pour Amazon EC2 Auto Scaling dans le Guide de l'utilisateur d'Auto EC2 Auto Scaling.	8 décembre 2022
Mise à jour des autorisations de rôle lié à un service IAM	La politique `AutoScalingServiceRolePolicy` accorde désormais des autorisations supplémentaires à Amazon EC2 Auto Scaling. Pour de plus amples informations, consultez Politiques gérées par AWS pour Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	6 décembre 2022

Nouvelle politique d'allocation des instances Spot	Vous pouvez maintenant utiliser la stratégie d'allocation optimisée en termes de prix et de capacité pour demander des instances Spot à partir des groupes Spot qui présenten t le moins de risque d'être interrompus et dont le prix est le plus bas possible. Pour plus d'informations, consultez Stratégies d'allocation dans le guide de l'utilisateur Amazon EC2 Auto Scaling.	10 novembre 2022
Prise en charge de la mise à l'échelle prédictive dans la Région Asie-Pacifique (Jakarta)	Vous pouvez désormais créer des politiques de mise à l'échelle prédictive dans la Région en Asie-Paci fique (Jakarta). Pour plus d'informations, consultez Mise à l'échelle prédictive pour Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	13 octobre 2022

Prise en charge des métriques personnalisées pour une mise à l'échelle prédictive dans la console	Vous pouvez désormais utiliser des métriques personnalisées lorsque vous créez des politiques de mise à l'échelle prédictive depuis la console Amazon EC2 Auto Scaling. Pour plus d'informations, consultez Mise à l'échelle prédictive pour Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	13 octobre 2022
CloudWatch surveillance des mesures de mise à l'échelle prédictives	Vous pouvez désormais accéder aux données de surveillance pour une mise à l'échelle prédictive à l'aide de CloudWatch. Cela vous permet d'utiliser les mathématiques appliquées aux métriques pour créer de nouvelles séries chronologiques qui affichent la précision des données de prévision . Pour plus d'informations, consultez la section Surveiller les métriques de dimensionnement prédictif CloudWatch dans le guide de l'utilisateur Amazon EC2 Auto Scaling.	7 juillet 2022

Prise en charge de la mise à l'échelle prédictive dans la Région Asie-Pacifique (Osaka)	Vous pouvez désormais créer des politiques de mise à l'échelle prédictive dans la Région Asie-Pacifique (Osaka). Pour plus d'informations, consultez Mise à l'échelle prédictive pour Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	6 juillet 2022
La mise en veille prolongée pré-initialisée est prise en charge dans de nouvelles régions	Vous pouvez désormais mettre les instances pré-initialisées en veille prolongée dans quatre Régions supplémentaires : Afrique (Le Cap), Asie-Pacifique (Jakarta), Asie-Pacifique (Osaka) et Europe (Milan). Pour plus d'informations, consultez Groupes d'instances pré-initialisées pour Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	5 juillet 2022

Surveillance de l'état de mise à jour vers	Lorsque vous effectuez des surveillances de l'état, Amazon EC2 Auto Scaling vous aide désormais à réduire les temps d'arrêt pouvant survenir en raison de problèmes temporaires ou de surveillance de l'état mal configurée. Pour de plus amples informations, veuillez consulter la rubrique Comment Amazon EC2 Auto Scaling minimise les temps d'arrêt dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	21 mai 2022
Préparation d'instance par défaut	Vous pouvez désormais unifier tous les paramètres de préchauffage et de refroidissement d'un groupe Auto Scaling et optimiser les performances des politiques de dimensionnement qui évoluent en continu en activant le préchauffage d'instance par défaut. Pour de plus amples informations, veuillez consulter Définir la préparation d'instance par défaut pour un groupe Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	19 avril 2022

Modifications du guide	Un nouveau chapitre sur l'intégration à d'autres AWS services a été ajouté au guide. Pour plus d'informations, consultez Services AWS intégrés avec Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	29 mars 2022
Mise à jour des autorisations de rôle lié à un service IAM	La politique `AutoScalingServiceRolePolicy` accorde désormais des autorisations de lecture supplémentaires à Amazon EC2 Auto Scaling. Pour de plus amples informations, consultez Politiques gérées par AWS pour Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	28 mars 2022
Les métadonnées d'instance fournissent l'état du cycle de vie cible	Vous pouvez récupérer l'état du cycle de vie cible d'une instance Auto Scaling à partir des métadonnées de l'instance. Pour plus d'informations, consultez Récupérer l'état du cycle de vie cible par le biais des métadonnées d'instance dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	24 mars 2022

Support de la nouvelle fonctionnalité de groupe d'instances pré-initialisées	Vous pouvez désormais mettre en veille prolongée les instances dans un groupe d'instances pré-initialisées pour arrêter ces instances sans supprimer leur contenu de mémoire (RAM). Vous pouvez désormais également renvoyer des instances vers le groupe d'instances pré-initialisées lors de la l'échelle horizontale, au lieu de toujours résilier la capacité d'instance dont vous aurez besoin par la suite. Pour plus d'informations, consultez Groupes d'instances pré-initialisées pour les instances Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	24 février 2022
Modifications du guide	La console Amazon EC2 Auto Scaling a été mise à jour avec des options supplémentaires pour vous aider à démarrer une actualisation d'instance avec le saut de correspondance activé et une configuration souhaitée spécifiée. Pour de plus amples informations, veuillez consulter Lancer ou annuler une actualisation d'instance (console) dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	3 février 2022

Métriques personnalisées pour les politiques de mise à l'échelle prédictive	Vous pouvez désormais choisir d'utiliser des métriques personnalisées lorsque vous créez des politiques de mise à l'échelle prédictive. Vous pouvez également utiliser les mathématiques de métrique pour personnaliser davantage les métriques que vous incluez dans votre politique. Pour plus d'informations, consultez Configurations avancées des politiques de mise à l'échelle prédictive à l'aide de métriques personnalisées.	24 novembre 2021
Nouvelle stratégie d'allocation à la demande	Vous pouvez désormais choisir de lancer les instances à la demande en fonction du prix (les types d'instances les moins chers en premier) lorsque vous créez un groupe Auto Scaling qui utilise une politique d'instances mixtes. Pour plus d'informations, consultez Stratégies d'allocation dans le guide de l'utilisateur Amazon EC2 Auto Scaling.	27 octobre 2021

Sélection de type d'instance basée sur des attributs	Amazon EC2 Auto Scaling ajoute la prise en charge de la sélection du type d'instanc e basée sur des attributs. Au lieu de choisir manuellem ent les types d'instance, vous pouvez exprimer vos besoins en matière d'instance sous la forme d'un ensemble d'attributs, tels que vCPU, mémoire et stockage. Pour plus d'informations, consultez Création d'un groupe Auto Scaling à l'aide de la sélection du type d'instance basée sur les attributs dans le guide de l'utilisateur Amazon EC2 Auto Scaling.	27 octobre 2021
Prise en charge du filtrage des groupes par identification	Vous pouvez désormais filtrer vos groupes Auto Scaling à l'aide de filtres d'identification lorsque vous récupérez des informations sur vos groupes Auto Scaling à l'aide de la commande `describe-auto-scaling-groups` . Pour plus d'informations, consultez Utilisation des identifications pour filtrer les groupes Auto Scaling dans le guide de l'utilisateur Amazon EC2 Auto Scaling.	14 octobre 2021

Modifications du guide	La console Amazon EC2 Auto Scaling a été mise à jour pour vous aider à créer des politiques de résiliation personnalisées avec. AWS Lambda La documentation de la console a été révisée en conséquence. Pour plus d'informations, consultez Utilisation de différentes politiques de résiliation (console).	14 octobre 2021
Support de la copie des configurations de lancement pour lancer des modèles	Vous pouvez désormais copier toutes les configurations de lancement d'une AWS région vers de nouveaux modèles de lancement depuis la console Amazon EC2 Auto Scaling. Pour plus d'informations, consultez Copier des configurations pour lancer des modèles dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	9 août 2021

Développe la fonctionnalité d'actualisation de l'instance	Vous pouvez désormais inclure des mises à jour, telles qu'une nouvelle version d'un modèle de lancement , lors du remplacement d'instances en ajoutant la configuration souhaitée à la commande `start-instance-refresh` . En activant la fonction Ignorer la correspondance, vous pouvez éviter de remplacer les instances qui disposent déjà de la configuration souhaitée . Pour plus d'informations, consultez Remplacement des instances Auto Scaling en fonction d'une actualisation d'instance dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	5 août 2021
Support des politiques de résiliation personnalisées	Vous pouvez désormais créer des politiques de résiliation personnalisées avec AWS Lambda. Pour plus d'informations, consultez Création d'une politique de résiliation personnalisée avec Lambda. La documentation de spécification des politiques de résiliation a été mise à jour en conséquence.	29 juillet 2021

Modifications du guide	La console Amazon EC2 Auto Scaling a été mise à jour et améliorée avec des fonctionn alités supplémentaires vous permettant de créer des actions planifiées avec un fuseau horaire spécifié. La documentation de la Mise à l'échelle planifiée a été révisée en conséquence.	3 juin 2021
Volumes gp3 dans les configurations de lancement	Vous pouvez désormais spécifier des volumes gp3 dans les mappages de périphériques de stockage en mode bloc pour les configura tions de lancement.	2 juin 2021
Support de la mise à l'échelle prédictive	Vous pouvez désormais utiliser la mise à l'échelle prédictive pour mettre à l'échelle proactivement vos groupes Amazon EC2 Auto Scaling. Pour plus d'informations, consultez Mise à l'échelle prédictiv e pour Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling. Avec cette mise à jour, la politique AutoScali ngServiceRolePolicygérée inclut désormais l'autoris ation d'appeler l'action d'`cloudwatch:GetMetr icData` API.	19 mai 2021

Modifications du guide	Vous pouvez désormais accéder à des exemples de modèles pour les hooks du cycle de vie à partir de GitHub. Pour plus d'informations, consultez Hooks du cycle de vie d'Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	9 avril 2021
Support des groupes de démarrage à chaud	Vous pouvez désormais équilibrer les performances (minimiser les démarrages à froid) et les coûts (arrêter de surprovisionner la capacité des instances) pour les applications avec de longs délais de premier démarrage en ajoutant des groupes d'instances pré-initialisées aux groupes Auto Scaling. Pour plus d'informations, consultez Groupes d'instances pré-initialisées pour les instances Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	08 avril 2021

Support des points de contrôle	Vous pouvez à présent ajouter des points de contrôle à une actualisation d'instance pour remplacer les instances par phases et effectuer des vérifications sur les instances à certains stades Pour plus d'informations, consultez Ajout de points de contrôle à une actualisation d'instance dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	18 mars 2021
Modifications du guide	Documentation améliorée à utiliser EventBridge avec les événements et les hooks du cycle de vie Amazon EC2 Auto Scaling. Pour plus d'informations, consultez Using Amazon EC2 Auto Scaling with EventBridge and Tutorial : Configurer un hook de cycle de vie qui invoque une fonction Lambda dans le guide de l'utilisateur d'Amazon EC2 Auto Scaling.	18 mars 2021

Support des fuseaux horaires locaux	Vous pouvez désormais créer des actions planifiées récurrentes dans le fuseau horaire local en ajoutant l'option `--time-zone` à la commande put-scheduled-update-group-action. Si votre fuseau horaire observe l'heure d'été (DST), l'action récurrente s'ajuste automatiquement à l'heure d'été. Pour plus d'informations, consultez Mise à l'échelle planifiée dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	9 mars 2021
Développe les fonctionnalités pour les politiques d'instances mixtes	Vous pouvez désormais hiérarchiser les types d'instance de votre capacité Spot lorsque vous utilisez une politique d'instances mixtes. Amazon EC2 Auto Scaling tente de remplir les priorités sur la base du meilleur effort, mais optimise d'abord la capacité. Pour plus d'informations, consultez Groupes Auto Scaling avec types d'instance et options d'achat multiples dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	8 mars 2021

Mise à l'échelle des activités pour les groupes supprimés	Vous pouvez désormais afficher les activités de mise à l'échelle pour les groupes Auto Scaling supprimés en ajoutant l'option `--include-deleted-groups` à la commande describe-scaling-activities. Pour plus d'informations, consultez Résolution des problèmes Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	23 février 2021
Améliorations apportées à la console	Vous pouvez maintenant créer et attacher un Application Load Balancer ou Network Load Balancer à partir de la console Amazon EC2 Auto Scaling. Pour plus d'informations, consultez Création et attachement d'un nouvel équilibreur Application Load Balancer ou d'un équilibreur Network Load Balancer (console) dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	24 novembre 2020

Plusieurs interfaces réseau	Vous pouvez désormais configurer un modèle de lancement pour un groupe Auto Scaling spécifiant plusieurs interfaces réseau. Pour plus d'informations, consultez Interfaces réseau dans un VPC.	23 novembre 2020
Modèles de lancement	Plusieurs modèles de lancement peuvent désormais être utilisés avec les groupes Auto Scaling. Pour plus d'informations, consultez Spécification d'un modèle de lancement différent pour un type d'instance dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	19 novembre 2020

Gateway Load Balancers.	Guide mis à jour pour montrer comment attacher un équilibreur de charge de passerelle à un groupe Auto Scaling pour s'assurer que les instances d'appliance lancées par Amazon EC2 Auto Scaling sont automatiquement enregistrées et désenregistrées de l'équilibreur de charge. Pour plus d'informations, consultezTypes Elastic Load Balancing et Attachement d'un équilibreur de charge à votre groupe Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	10 novembre 2020
Durée de vie maximale de l'instance	Vous pouvez désormais réduire la durée de vie maximale de l'instance à une journée (86 400 secondes). Pour plus d'informations, consultez Remplacement d'instances Auto Scaling en fonction de la durée de vie maximale de l'instance dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	9 novembre 2020

Rééquilibrage de la capacité	Vous pouvez configurer votre groupe Auto Scaling pour lancer une Instance Spot de remplacement lorsque Amazon EC2 émet une recommandation de rééquilibrage. Pour plus d'informations, consultez Rééquilibrage de la capacité Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	4 novembre 2020
Instance Metadata Service Version 2	Vous pouvez exiger l'utilisation du Service des métadonnées d'instance Version 2, qui est une méthode orientée session de demande de métadonnées d'instance lors de l'utilisation de configurations de lancement. Pour plus d'informations, consultez Configuration des options de métadonnées d'instances dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	28 juillet 2020

Modifications du guide	Diverses améliorations dans les sections Contrôle des instances Auto Scaling qui sont résiliées pendant la mise à l'échelle horizontale, Surveillance de vos instances et groupes Auto Scaling, Modèles de lancement et Configurations de lancement du Guide de l'utilisateur Amazon EC2 Auto Scaling.	28 juillet 2020
Actualisation d'instance	Lancez une actualisation d'instance pour mettre à jour toutes les instances de votre groupe Auto Scaling lorsque vous modifiez la configura tion. Pour plus d'informations, consultez Remplacement des instances Auto Scaling en fonction d'une actualisation d'instance dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	16 juin 2020

Modifications du guide	Diverses améliorations concernant les sections Remplacement d'instances Auto Scaling en fonction de la durée de vie maximale de l'instance, Groupes Auto Scaling avec plusieurs types d'instance et options d'achat, Mise à l'échelle en fonction d'Amazon SQS et Balisage des groupes Auto Scaling et des instances dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	6 mai 2020
Modifications du guide	Diverses améliorations à la documentation IAM. Pour plus d'informations, consultez Support des modèles et Exemples de politique basée sur une identité Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	4 mars 2020

Politiques de mise à l'échelle uniques	Vous pouvez désormais désactiver et réactiver les politiques de mise à l'échelle. Cette fonctionnalité vous permet de désactiver temporairement une politique de mise à l'échelle tout en conservant les détails de configuration afin de pouvoir réactiver la politique ultérieur ement. Pour de plus amples informations, consultez Désactivation d'une politique de mise à l'échelle dans un groupe Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	18 février 2020
Ajout d'une fonctionnalité de notification	Amazon EC2 Auto Scaling vous envoie désormais des événements AWS Health Dashboard lorsque vos groupes Auto Scaling ne peuvent pas être redimensi onnés en raison de l'absence d'un groupe de sécurité ou d'un modèle de lancement . Pour plus d'informations, consultez Réception de notifications AWS Health Dashboard pour Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	12 février 2020

Modifications du guide	Diverses améliorations et corrections dans les sections Mode de fonctionnement d'Amazon EC2 Auto Scaling avec IAM, Exemples de politique Amazon EC2 Auto Scaling identity-based, Politique de clé CMK requise pour une utilisation avec des volumes chiffrés et Surveillance de vos instances et groupes dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	10 février 2020
Modifications du guide	Amélioration de la documentation pour les groupes Auto Scaling qui utilisent la pondération d'instance. Découvrez comment utiliser les politiques de mise à l'échelle lors de l'utilisation d'« unités de capacité » pour mesurer la capacité désirée. Pour plus d'informations, consultez Fonctionnement des politiques et Types d'ajustement de mise à l'échelle dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	6 février 2020

Nouveau chapitre « Sécurité »	Un nouveau chapitre Sécurité dans le Guide de l'utilisateur Amazon EC2 Auto Scaling vous aide à comprendr e comment appliquer le modèle de responsabilité partagée lorsque vous utilisez Amazon EC2 Auto Scaling. Dans le cadre de cette mise à jour, le chapitre « Contrôle d'accès à vos ressources Amazon EC2 Auto Scaling » du guide de l'utilisateur a été remplacé par une nouvelle section plus utile, Identity and access management pour Amazon EC2 Auto Scaling.	4 février 2020
Recommandations pour les types d'instances	AWS Compute Optimizer fournit des recommand ations relatives aux instances Amazon EC2 pour vous aider à améliorer les performances, à économiser de l'argent, ou les deux. Pour plus d'informa tions, consultez Réceptions de recommandations pour un type d'instance dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	3 décembre 2019

Hôtes dédiés et groupes de ressources hôte	Guide mis à jour pour montrer comment créer un modèle de lancement spécifiant un groupe de ressources hôte. Cela vous permet de créer un groupe Auto Scaling avec un modèle de lancement qui spécifie une AMI BYOL à utiliser sur des hôtes dédiés. Pour plus d'informations, consultez Création d'un modèle de lancement pour un groupe Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	3 décembre 2019
Support des points de terminaison Amazon VPC	Vous pouvez maintenant établir une connexion privée entre votre VPC et Amazon EC2 Auto Scaling. Pour plus d'informations, consultez Amazon EC2 Auto Scaling et les points de terminaison d'un VPC d'interface dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	22 novembre 2019

Durée de vie maximale de l'instance	Vous pouvez désormais remplacer automatiquement des instances en spécifiant la durée maximale pendant laquelle une instance peut être en service. Si des instances approchent de cette limite, Amazon EC2 Auto Scaling les remplace progressivement. Pour plus d'informations, consultez Remplacement d'instances Auto Scaling en fonction de la durée de vie maximale de l'instance dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	19 novembre 2019
Pondération d'instance	Pour les groupes Auto Scaling comportant plusieurs types d'instance, vous pouvez désormais spécifier le nombre d'unités de capacité avec lequel chaque type d'instance contribue à la capacité du groupe. Pour plus d'informations, consultez Pondération des instances pour les instances Amazon EC2 Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	19 novembre 2019

Nombre minimum de types d'instance	Vous n'avez plus besoin de spécifier des types d'instance supplémentaires pour des groupes d'instances Spot, à la demande et réservées. Pour tous les groupes Auto Scaling, le minimum est désormais un type d'instance. Pour plus d'informations, consultez Groupes Auto Scaling avec types d'instance et options d'achat multiples dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	16 septembre 2019
Support de la nouvelle politique d'allocation des instances Spot	Amazon EC2 Auto Scaling prend désormais en charge une nouvelle politique d'allocation des instances Spot « optimisée pour la capacité » qui répond à votre demande à l'aide de groupes d'instances Spot choisis de façon optimale en fonction de la capacité Spot disponible. Pour plus d'informations, consultez Groupes Auto Scaling avec types d'instance et options d'achat multiples dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	12 août 2019

Modifications du guide	Amélioration de la documentation Amazon EC2 Auto Scaling dans les rubriques Rôles liés à un service et Politique de clé CMK obligatoire à utiliser avec les volumes chiffrés.	1 août 2019
Support de l'amélioration du balisage	Amazon EC2 Auto Scaling ajoute désormais des balises aux instances Amazon EC2 dans le cadre du même appel d'API que celui qui lance les instances. Pour plus d'informations, consultez la rubrique relative au balisage des groupes et des instances Auto Scaling.	26 juillet 2019
Modifications du guide	Amélioration de la documentation Amazon EC2 Auto Scaling dans la rubrique Suspension et reprise des processus de mise à l'échelle. Mise à jour des exemples de politiques gérées par le client pour inclure un exemple de politique qui permet aux utilisateurs de transférer uniquement à Amazon EC2 Auto Scaling les rôles liés à un service avec suffixe personnalisé spécifique.	13 juin 2019

Support d'une nouvelle fonctionnalité Amazon EBS	Ajout du support d'une nouvelle fonctionnalité Amazon EBS dans la rubrique relative au modèle de lancement. Modifiez l'état de chiffrement d'un volume EBS lors de la restauration à partir d'un instantané. Pour plus d'informations, consultez Création d'un modèle de lancement pour un groupe Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	13 mai 2019
Modifications du guide	Amélioration de la documentation Amazon EC2 Auto Scaling dans les sections suivantes : Contrôle des instances Auto Scaling qui se terminent pendant la mise à l'échelle horizontale, Groupes Auto Scaling, Groupes Auto Scaling avec plusieurs types d'instance et options d'achat et Mise à l'échelle dynamique pour Amazon EC2 Auto Scaling.	12 mars 2019

Support de la combinaison des types d'instance et des options d'achat	Allouez et faites évoluer automatiquement les instances entre les options d'achat (instances ponctuelles, instances à la demande et Instances réservées) et les types d'instances au sein d'un même groupe Auto Scaling. Pour plus d'informations, consultez Groupes Auto Scaling avec types d'instance et options d'achat multiples dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	13 novembre 2018
Rubrique mise à jour pour la mise à l'échelle en fonction d'Amazon SQS	Mise à jour du guide pour expliquer comment vous pouvez utiliser les métriques personnalisées pour mettre à l'échelle un groupe Auto Scaling en réponse à une évolution de la demande à partir d'une file d'attente Amazon SQS. Pour plus d'informations, consultez Mise à l'échelle en fonction d'Amazon SQS dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	26 juillet 2018

Le tableau suivant décrit les modifications importantes apportées à la documentation Amazon EC2 Auto Scaling avant juillet 2018.

Fonctionnalité	Description	Date de publication
Support des politiques de suivi des objectifs et d'échelonnement	Configurez la mise à l'échelle dynamique pour votre application en quelques étapes seulement. Pour plus d'informations, consultez Politiques de suivi des objectifs et d'échelonnement pour Amazon EC2 Auto Scaling.	12 juillet 2017
Support des autorisations au niveau des ressources	Créez des politiques IAM pour contrôler l'accès au niveau des ressources. Pour plus d'informations, consultez Contrôle de l'accès à vos ressources Amazon EC2 Auto Scaling.	15 mai 2017
Surveillance des améliorations	Pour les métriques de groupe Auto Scaling, vous n'avez plus besoin d'activer la surveillance détaillée. Vous pouvez maintenant activer la collecte des mesures de groupe et afficher les graphiques de métriques dans l'onglet Monitoring (Surveillance) de la console. Pour plus d'informations, consultez Monitoring your Auto Scaling groups and instances using Amazon CloudWatch.	18 août 2016
Support des Application Load Balancers	Attachez un ou plusieurs groupes cibles à un groupe Auto Scaling nouveau ou existant. Pour plus d'informations, consultez Attachement d'un équilibreur de charge à votre groupe Auto Scaling.	11 août 2016
Événements pour hooks de cycle de vie	Amazon EC2 Auto Scaling envoie des événements EventBridge lorsqu'il appelle des hooks du cycle de vie. Pour plus d'informations, consultez Getting EventBridge when your Auto Scaling group scale.	24 février 2016
Protection d'instance	Empêche Amazon EC2 Auto Scaling de sélectionner des instances spécifiques pour la résiliation lors de la mise à l'échelle horizontale. Pour plus d'informations, consultez Protection d'instance.	07 décembre 2015

Fonctionnalité	Description	Date de publication
Politiques de mise à l'échelle d'étape	Créez une politique de mise à l'échelle qui vous permette de mettre à l'échelle en fonction de la valeur du seuil de l'alarme. Pour plus d'informations, consultez Types de politique de mise à l'échelle.	06 juillet 2015
Mettez à jour l'équilibreur de charge	Attachez un équilibreur de charge à un groupe Auto Scaling existant ou détachez-le. Pour plus d'informations, consultez Attachement d'un équilibreur de charge à votre groupe Auto Scaling.	11 juin 2015
Support pour ClassicLink	Liez des instances EC2-Classic dans le groupe Auto Scaling à un VPC, en permettant la communication entre ces instances EC2-Classic liées et les instances du VPC avec des adresses IP privées. Pour plus d'informations, consultez Liaison d'instances EC2-Classic à un VPC.	19 janvier 2015
Hooks de cycle de vie	Maintenez les instances nouvellement lancées ou résiliées en attente pendant que vous exécutez des actions sur elles. Pour plus d'informations, consultez Hooks du cycle de vie Amazon EC2 Auto Scaling.	30 juillet 2014
Détacher des instances	Détachez des instances d'un groupe Auto Scaling. Pour plus d'informations, consultez Détacher les instances EC2 de votre groupe Auto Scaling.	30 juillet 2014
Mettez les instances en veille	Mettez les instances qui sont en statut `InService` en `Standby`. Pour plus d'informations, consultez Suppression temporaire des instances du groupe Auto Scaling.	30 juillet 2014
Gérer les balises	Gérez vos groupes Auto Scaling à l'aide de la AWS Management Console. Pour plus d'informations, consultez la rubrique relative au balisage des groupes et des instances Auto Scaling.	01 mai 2014

Fonctionnalité	Description	Date de publication
Support pour les instances dédiées	Lancez des instances dédiées en spécifiant un attribut de location de placement lorsque vous créez une configuration de lancement. Pour plus d'informations, consultez Location de placement de l'instance.	23 avril 2014
Créez un groupe ou une configuration de lancement à partir d'une instance EC2	Créez un groupe Auto Scaling ou une configuration de lancement avec une instance EC2. Pour plus d'informations sur la création d'une configuration de lancement avec une instance EC2, consultez Création d'une configuration de lancement avec une instance EC2. Pour plus d'informations sur la création d'un groupe Auto Scaling avec une instance EC2, consultez Création d'un groupe Auto Scaling à l'aide d'une instance EC2.	02 janvier 2014
Attachez des instances	Activez la scalabilité automatique pour une instance EC2 en attachant l'instance à un groupe Auto Scaling existant. Pour plus d'informations, consultez Attacher des instances EC2 à un groupe Auto Scaling.	02 janvier 2014
Affichez les limites de compte	Affichez les limites sur les ressources Auto Scaling pour votre compte. Pour plus d'informations, consultez Limites Auto Scaling.	02 janvier 2014
Support de la console pour Amazon EC2 Auto Scaling	Accédez à Amazon EC2 Auto Scaling à l'aide du. AWS Management Console Pour plus d'informations, consultez Mise en route avec Amazon EC2 Auto Scaling.	10 décembre 2013
Attribuez une adresse IP publique	Attribuez une adresse IP publique à une instance lancée dans un VPC. Pour plus d'informations, consultez Lancement d'instances Auto Scaling dans un VPC.	19 septembre 2013

Fonctionnalité	Description	Date de publication
Politique de résiliation d'instance	Spécifiez une politique de résiliation d'instance à utiliser par Amazon EC2 Auto Scaling lors de la résiliation des instances EC2. Pour plus d'informations, consultez Contrôle des instances Auto Scaling à résilier pendant une diminution en charge des instances.	17 septembre 2012
Support des rôles IAM	Lancez des instances EC2 avec un profil d'instance IAM. Vous pouvez utiliser cette fonction pour attribuer des rôles IAM à vos instances, en permettant à vos applications d'accéder à d'autres services Amazon Web Services en toute sécurité. Pour plus d'informations, consultez Lancement d'instances Auto Scaling avec un rôle IAM.	11 juin 2012
Support des instances Spot	Lancez des instances Spot avec une configuration du lancement. Pour plus d'informations, consultez Demander des instances Spots pour des applications flexibles et tolérantes aux pannes.	7 juin 2012
Balisez les groupes et les instances	Balisez les groupes Auto Scaling et spécifiez que la balise s'applique aussi aux instances EC2 lancées après sa création. Pour plus d'informations, consultez la rubrique relative au balisage des groupes et des instances Auto Scaling.	26 janvier 2012

Fonctionnalité	Description	Date de publication
Support d'Amazon SNS	Utilisez Amazon SNS pour recevoir des notifications lorsqu'Amazon EC2 Auto Scaling lance des instances EC2 ou les résilie. Pour plus d'informations, consultez Réception de notifications SNS lorsque le groupe Auto Scaling évolue. Amazon EC2 Auto Scaling a également ajouté les nouvelles fonctions suivantes : • La capacité à configurer des activités de mise à l'échelle récurrentes avec la syntaxe cron. Pour plus d'informations, consultez l'opération API `PutScheduledUpdateGroupAction`. • Nouveau paramètre de configuration qui vous permet d'effectuer une mise à l'échelle sans ajouter l'instance lancée à l'équilibreur de charge (LoadBalancer). Pour plus d'informations, consultez le type de données API `ProcessType`. • L'indicateur `ForceDelete` dans l'opération `DeleteAutoScalingGroup` indique à Amazon EC2 Auto Scaling de supprimer le groupe Auto Scaling avec ses instances associées, sans attendre que ces dernières soient d'abord résiliées. Pour plus d'informations, consultez l'opération API `DeleteAutoScalingGroup`.	20 juillet 2011
Actions de mise à l'échelle planifiées	Support ajoutée des actions du mise à l'échelle planifié. Pour plus d'informations, consultez Mise à l'échelle planifiée dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	2 décembre 2010
Support pour Amazon VPC	Ajout du support pour Amazon VPC. Pour plus d'informations, consultez Lancement d'instances Auto Scaling dans un VPC.	2 décembre 2010

Fonctionnalité	Description	Date de publication
Support des clusters HPC	Ajout du support des clusters High Performance Computing (HPC)	2 décembre 2010
Support des surveillances de l'état	Ajout du support pour l'utilisation des surveillances de l'état Elastic Load Balancing avec des instances EC2 gérées par Amazon EC2 Auto Scaling Pour plus d'informations, consultez la section Health checks pour les instances d'un groupe Auto Scaling.	2 décembre 2010
Support pour les CloudWatch alarmes	Suppression de l'ancien mécanisme de déclenchement et refonte d'Amazon EC2 Auto Scaling pour utiliser CloudWatch la fonction d'alarme. Pour plus d'informations, consultez Mise à l'échelle dynamique dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.	2 décembre 2010
Suspension et reprise du mise à l'échelle	Support ajoutée pour l'interruption et la reprise des processus de mise à l'échelle.	2 décembre 2010
Support d'IAM	Ajout du support d'IAM. Pour plus d'informations, consultez Contrôle de l'accès à vos ressources Amazon EC2 Auto Scaling.	2 décembre 2010

Les traductions sont fournies par des outils de traduction automatique. En cas de conflit entre le contenu d'une traduction et celui de la version originale en anglais, la version anglaise prévaudra.