Utilisation des tables globales Amazon DynamoDB

AWS Directives prescriptives

AWS Directives prescriptives: Utilisation des tables globales Amazon DynamoDB

Table of Contents

Introduction 1
Présentation 2
Faits importants 2
Cas d'utilisation 4
Modes d'écriture 5
Mode Écrire dans n'importe quelle région (non primaire) 5
Mode Écrire dans une région (primaire unique) 8
Mode Écrire dans votre région (primaire mixte) 10
Stratégies de routage 13
Routage des demandes piloté par le client 14
Routage des demandes dans la couche de calcul 15
Routage des demandes avec Route 53 17
Routage des demandes avec Global Accelerator 18
processus d'évacuation 20
Évacuation d'une région active 20
Évacuation d'une région déconnectée 21
Planification de la capacité de débit 24
Liste de contrôle pour la préparation 26
FAQ 28
Quel est le coût des tables globales ? 28
Quelles sont les régions prises en charge par les tables globales ? 28
Comment les index secondaires globaux (GSI) sont-ils gérés avec les tables globales ? 29
Comment arrêter la réplication d’une table globale ? 29
Comment Amazon DynamoDB Streams interagit-il avec les tables globales ? 29
Comment les tables globales gèrent-elles les transactions ? 29
Comment les tables globales interagissent-elles avec le cache de DynamoDB Accelerator (DAX) ? 30
Les balises présentes sur les tables se propagent-elles ? 30
Dois-je sauvegarder des tables dans toutes les régions ou dans une seule ? 30
Comment déployer des tables globales avec AWS CloudFormation ? 30
Conclusion et ressources 32
Historique du document 33
Glossaire 34
..... 34

A 35
B 38
C 40
D 43
E 48
F 50
G 51
H 52
I 53
L 56
M 57
O 61
P 64
Q 67
R 67
S 70
T 74
U 75
V 76
W 76
Z 78
..... lxxix

Utilisation des tables globales Amazon DynamoDB

Jason Hunter, Amazon Web Services (AWS)

Mars 2024 (historique du document)

Les tables globales s'appuient sur l'étendue internationale d'Amazon DynamoDB pour vous fournir une base de données entièrement gérée, à régions multiples et à activités multiples, qui fournit des performances de lecture et d'écriture rapides et locales, pour des applications internationales et mises à l'échelle massivement. Les tables globales répliquent automatiquement vos tables DynamoDB parmi celles de votre choix. Régions AWS Aucune modification de l'application n'est requise, car les tables globales utilisent des API DynamoDB existantes. L'utilisation de tables globales n'entraîne ni frais initiaux ni engagement. Vous ne payez que les ressources que vous utilisez.

Ce guide explique comment utiliser efficacement les tables globales DynamoDB. Il fournit des informations clés sur les tables globales, explique les principaux cas d'utilisation de la fonctionnalité, présente une taxonomie de trois modèles d'écriture différents à prendre en compte, passe en revue les quatre principales options de routage des demandes que vous pourriez implémenter, explique comment évacuer une région active ou une région hors ligne, comment envisager la planification de la capacité de débit et fournit une liste des éléments à prendre en compte lors du déploiement de tables globales.

Ce guide s'inscrit dans le contexte plus large des déploiements AWS multirégionaux, tel que décrit dans le livre blanc sur les principes fondamentaux du AWS multirégional et dans les modèles de conception relatifs à la résilience des données avec vidéo. AWS

Table des matières

- Présentation
- Modes d'écriture
- Stratégies de routage
- Processus d'évacuation
- Planification de la capacité de débit
- Liste de contrôle pour la préparation
- FAQ
- Conclusion et ressources

Présentation des tables globales

Faits importants

- Il existe deux versions des tables globales : la version 2017.11.29 (ancienne version) (parfois appelée v1) et la version 2019.11.21 (actuelle) (parfois appelée v2). Ce guide se concentre exclusivement sur la version actuelle.
- DynamoDB (sans tables globales) est un service régional, ce qui signifie qu'il est hautement disponible et intrinsèquement résilient aux défaillances de l'infrastructure, y compris à la défaillance de l'ensemble d'une zone de disponibilité. Une table DynamoDB à région unique est conçue pour garantir une disponibilité de 99,99 %. Pour plus d'informations, consultez le contrat de niveau de service (SLA) de DynamoDB.
- Une table globale DynamoDB réplique ses données entre deux régions ou plus. Un tableau DynamoDB multirégional est conçu pour garantir une disponibilité de 99,999 %. Avec une planification appropriée, les tables globales peuvent aider à créer une architecture résiliente aux défaillances régionales.
- Les tables globales utilisent un modèle de réplication actif-actif. Du point de vue de DynamoDB, la table de chaque région dispose du même statut pour accepter les demandes de lecture et d'écriture. Après avoir reçu une demande d'écriture, la table de réplication locale reproduit l'opération d'écriture vers les autres régions distantes participantes en arrière-plan.
- Les éléments sont répliqués individuellement. Les éléments mis à jour au cours d'une même transaction peuvent ne pas être répliqués ensemble.
- Chaque partition de table de la région source reproduit ses opérations d'écriture en parallel avec toutes les autres partitions. La séquence des opérations d'écriture dans une région distante peut ne pas correspondre à la séquence des opérations d'écriture effectuées dans la région source. Pour plus d'informations sur les partitions de table, consultez le billet de blog Mise à l'échelle de DynamoDB : comment les partitions, les touches de raccourci et les divisions de chaleur ont un impact sur les performances.
- Un élément nouvellement écrit est généralement propagé à toutes les tables de réplica en une seconde. Les régions voisines ont tendance à se propager plus rapidement.
- Amazon CloudWatch fournit une `ReplicationLatency` métrique pour chaque paire de régions. Il est calculé en examinant les articles qui arrivent, en comparant leur heure d'arrivée avec leur heure d'écriture initiale et en calculant une moyenne. Les horaires sont stockés CloudWatch dans

la région source. L'affichage des délais moyen et maximum peut être utile pour déterminer le délai de réplication moyen et le pire des cas. Il n'existe aucun SLA sur cette latence.

- Si un élément individuel est mis à jour à peu près au même moment (dans cette `ReplicationLatency` fenêtre) dans deux régions différentes et que la deuxième opération d'écriture a lieu avant que la première opération d'écriture ne soit répliquée, il existe un risque de conflit d'écriture. Les tables globales résolvent ces conflits en utilisant un mécanisme de victoire du dernier rédacteur, basé sur l'horodatage des opérations d'écriture. La première opération « perd » au profit de la seconde. Ces conflits ne sont pas enregistrés dans CloudWatch ouAWS CloudTrail.
- Chaque élément possède un horodatage de la dernière écriture conservé en tant que propriété système privée. L'approche du dernier rédacteur gagne est mise en œuvre en utilisant une opération d'écriture conditionnelle qui nécessite que l'horodatage de l'élément entrant soit supérieur à l'horodatage de l'élément existant.
- Un tableau global reproduit tous les éléments dans toutes les régions participantes. Si vous souhaitez disposer de différentes étendues de réplication, vous pouvez créer plusieurs tables globales et attribuer à chaque table des régions participantes différentes.
- La région locale accepte les opérations d'écriture même si la région de réplique est hors ligne ou s'`ReplicationLatency`agrandit. La table locale continue de tenter de répliquer les éléments vers la table distante jusqu'à ce que chaque élément réussisse.
- Dans le cas peu probable où une région se retrouverait complètement hors ligne, toutes les réplications sortantes et entrantes en attente seront réessayées ultérieurement. Aucune action particulière n'est requise pour rétablir la synchronisation des tables. Le mécanisme du dernier auteur gagne garantit que les données finissent par devenir cohérentes.
- Vous pouvez ajouter une nouvelle région à une table DynamoDB à tout moment. DynamoDB gère la synchronisation initiale et la réplication continue. Vous pouvez également supprimer une région (même la région d'origine), ce qui supprimera la table locale de cette région.
- DynamoDB ne possède pas de point de terminaison global. Toutes les demandes sont adressées à un point de terminaison régional qui accède à l'instance de table globale locale à cette région.
- Les appels à DynamoDB ne doivent pas passer d'une région à l'autre. La meilleure pratique consiste à ce qu'une application hébergée dans une région accède directement uniquement au point de terminaison DynamoDB local de sa région. Si des problèmes sont détectés dans une région (dans la couche DynamoDB ou dans la pile environnante), le trafic de l'utilisateur final doit être acheminé vers un point de terminaison d'application différent hébergé dans une autre région. Les tables globales garantissent que l'application hébergée dans chaque région a accès aux mêmes données.

Cas d'utilisation

Les tables globales offrent les avantages communs suivants :

- Opérations de lecture à faible latence. Vous pouvez placer une copie des données plus près de l'utilisateur final afin de réduire la latence du réseau lors des opérations de lecture. Les données sont conservées aussi fraîches que la `ReplicationLatency` valeur.
- Opérations d'écriture à faible latence. Un utilisateur final peut écrire dans une région voisine afin de réduire la latence du réseau et le temps nécessaire pour terminer l'opération d'écriture. Le trafic d'écriture doit être acheminé avec soin pour éviter tout conflit. Les techniques de routage sont abordées dans une section ultérieure.
- Résilience et reprise après sinistre accrues. Si les performances d'une région sont dégradées ou si elle est totalement indisponible, vous pouvez l'évacuer (déplacer une partie ou la totalité des demandes destinées à cette région) et atteindre un objectif de point de reprise (RPO) et un objectif de temps de reprise (RTO) mesurés en secondes. L'utilisation de tables globales augmente également le SLA de DynamoDB pour le pourcentage de disponibilité mensuel de 99,99 % à 99,999 %.
- Migration fluide entre les régions. Vous pouvez ajouter une nouvelle région, puis supprimer l'ancienne région pour migrer un déploiement d'une région vers une autre, sans aucune interruption au niveau de la couche de données.

Par exemple, Fidelity Investments a présenté à re:Invent 2022 comment elle utilise les tables globales DynamoDB pour son système de gestion des commandes. Leur objectif était de parvenir à un traitement fiable à faible latence à une échelle impossible à atteindre avec le traitement sur site, tout en maintenant la résilience face aux défaillances des zones de disponibilité et des régions.

Modes d'écriture pour les tables globales

Les tables globales sont toujours actives-actives au niveau de la table. Toutefois, vous pouvez les traiter comme actives-passives en contrôlant la façon dont vous acheminez les demandes d'écriture. Par exemple, vous pouvez décider d'acheminer les demandes d'écriture vers une seule région afin d'éviter d'éventuels conflits d'écriture.

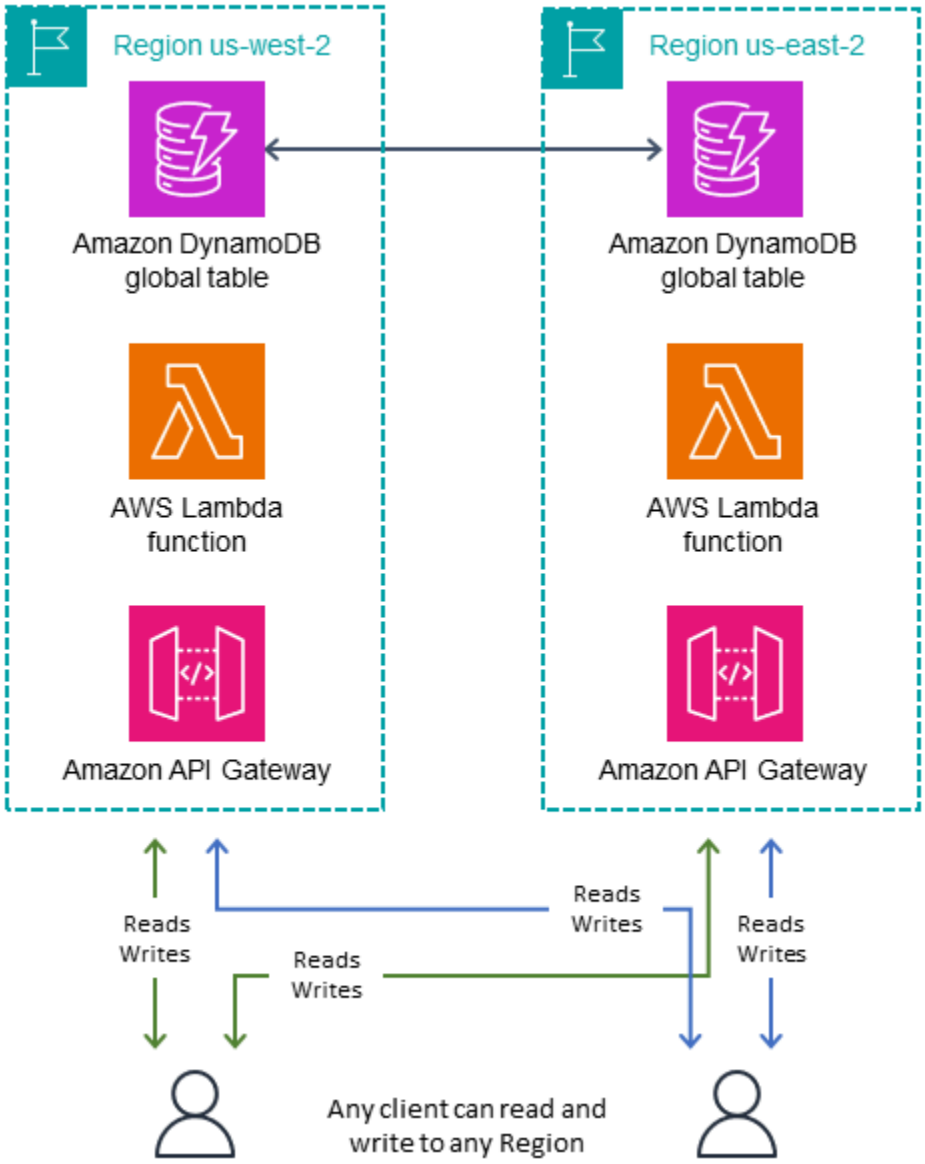
Il existe trois principaux modèles d'écriture gérés, comme expliqué dans les trois sections suivantes. Vous devez déterminer quel modèle d'écriture correspond à votre cas d'utilisation. Ce choix affecte la manière dont vous acheminez les demandes, évacuez une région et gérez la reprise après sinistre. Les instructions fournies dans les sections suivantes dépendent du mode d'écriture de votre application.

Rubriques

- Mode Écrire dans n'importe quelle région (non primaire)
- Mode Écrire dans une région (primaire unique)
- Mode Écrire dans votre région (primaire mixte)

Mode Écrire dans n'importe quelle région (non primaire)

Le mode d'écriture dans n'importe quelle région est entièrement actif-actif et n'impose aucune restriction quant à l'endroit où une opération d'écriture peut avoir lieu. Toute région peut accepter une demande écrite à tout moment. Il s'agit du mode le plus simple, mais il ne peut être utilisé qu'avec certains types d'applications. Il convient lorsque toutes les opérations d'écriture sont idempotentes. Idempotent signifie qu'elles sont reproductibles en toute sécurité afin que les opérations d'écriture simultanées ou répétées entre les régions ne soient pas en conflit, par exemple lorsqu'un utilisateur met à jour ses données de contact. Cela fonctionne également bien pour un ensemble de données à ajout uniquement où toutes les opérations d'écriture sont des insertions uniques sous une clé primaire déterministe, ce qui constitue un cas particulier d'idempotent. Enfin, ce mode convient lorsque le risque d'opérations d'écriture conflictuelles est acceptable.



Le mode Écrire dans n'importe quelle région est l'architecture la plus simple à implémenter. Le routage est plus facile car n'importe quelle région peut être la cible d'écriture à tout moment. Le basculement est plus facile, car toutes les opérations d'écriture récentes peuvent être réexécutées autant de fois que vous le souhaitez dans n'importe quelle région secondaire. Dans la mesure du possible, votre conception doit suivre ce mode d'écriture.

Par exemple, plusieurs services de streaming vidéo utilisent des tableaux globaux pour suivre les favoris, les avis, les indicateurs d'état des visionnages, etc. Ces déploiements peuvent utiliser le mode d'écriture vers n'importe quel mode Region tant qu'ils garantissent que chaque opération d'écriture est idempotente. Ce sera le cas si chaque mise à jour (par exemple, la définition d'un nouveau code horaire, l'attribution d'un nouvel avis ou la définition d'un nouveau statut de suivi)

attribue directement le nouvel état à l'utilisateur, et que la prochaine valeur correcte pour un article ne dépend pas de sa valeur actuelle. Si, par hasard, les demandes d'écriture de l'utilisateur sont acheminées vers différentes régions, la dernière opération d'écriture persistera et l'état global se stabilisera en fonction de la dernière affectation. Les opérations de lecture dans ce mode finiront par devenir cohérentes, retardées par la dernière `ReplicationLatency` valeur.

Autre exemple : une entreprise de services financiers utilise des tables globales dans le cadre d'un système visant à tenir à jour le décompte des achats par carte de débit pour chaque client, afin de calculer les remises en argent de ce client. Les nouvelles transactions arrivent du monde entier et vont vers plusieurs régions. Cette entreprise a pu utiliser le mode d'écriture dans n'importe quelle région grâce à une refonte minutieuse. Le croquis de conception initial ne contenait qu'un seul `RunningBalance` article par client. Les actions du client ont mis à jour la balance avec une `ADD` expression, qui n'est pas idempotente (car la nouvelle valeur correcte dépend de la valeur actuelle), et la balance était désynchronisée si deux opérations d'écriture étaient effectuées sur la même balance à peu près au même moment dans différentes régions. La refonte utilise le streaming d'événements, qui fonctionne comme un registre avec un flux de travail d'ajout uniquement. Chaque action du client ajoute un nouvel élément à la collection d'éléments gérée pour ce client. (Une collection d'articles est un ensemble d'éléments qui partagent une clé primaire mais qui ont des clés de tri différentes.) Chaque opération d'écriture est une insertion idempotente qui utilise l'ID client comme clé de partition et l'ID de transaction comme clé de tri. Cette conception complique le calcul de la balance car elle nécessite d'extraire les éléments puis de `Query` faire quelques calculs côté client, mais elle rend toutes les opérations d'écriture impuissantes et permet de simplifier considérablement le routage et le basculement. (Ce point est présentée plus en détail plus en détail dans ce guide.)

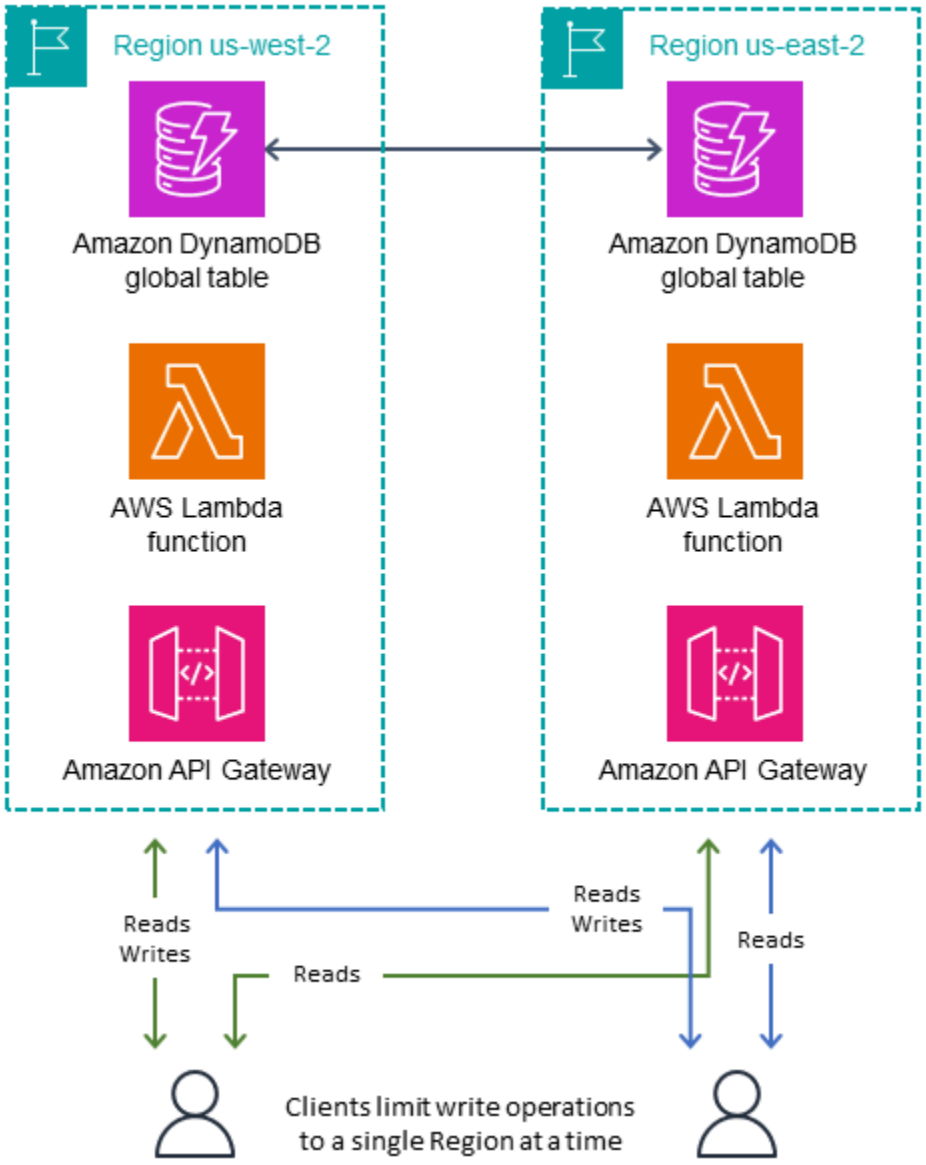
Un troisième exemple concerne une entreprise qui fournit des services de placement d'annonces en ligne. Cette société a décidé qu'un faible risque de perte de données serait acceptable pour simplifier la conception du mode d'écriture vers n'importe quelle région. Lorsqu'ils diffusent des publicités, ils ne disposent que de quelques millisecondes pour récupérer suffisamment de métadonnées afin de déterminer l'annonce à diffuser, puis pour enregistrer l'impression de l'annonce afin de ne pas répéter la même annonce rapidement. Ils utilisent des tables globales pour obtenir à la fois des opérations de lecture à faible latence pour les utilisateurs finaux du monde entier et des opérations d'écriture à faible latence. Ils enregistrent toutes les impressions publicitaires d'un utilisateur au sein d'un seul élément, qui est représenté sous la forme d'une liste croissante. Ils utilisent un seul élément au lieu de l'ajouter à une collection d'articles, ce qui leur permet de supprimer les anciennes impressions publicitaires lors de chaque opération de rédaction sans avoir à payer pour une opération de suppression. Cette opération d'écriture n'est pas impuissante ; si le même utilisateur final voit des

publicités diffusées dans plusieurs régions à peu près au même moment, il est possible qu'une opération d'écriture pour une impression d'annonce en remplace une autre. Le risque est qu'un utilisateur voie une annonce se répéter de temps en temps. Ils ont décidé que c'était acceptable.

Mode Écrire dans une région (primaire unique)

Le mode d'écriture dans une région est actif-passif et achemine toutes les opérations d'écriture de table vers une seule région active. (DynamoDB n'a pas la notion d'une région active unique ; c'est la couche externe à DynamoDB qui gère cela.) Le mode d'écriture dans une région évite les conflits d'écriture en garantissant que les opérations d'écriture ne concernent qu'une région à la fois. Ce mode d'écriture est utile lorsque vous souhaitez utiliser des expressions conditionnelles ou des transactions. Ces expressions ne sont pas possibles à moins que vous ne sachiez que vous agissez à l'encontre des données les plus récentes. Elles nécessitent donc d'envoyer toutes les demandes d'écriture à une seule région disposant des données les plus récentes.

À terme, des opérations de lecture cohérentes peuvent être effectuées vers n'importe laquelle des régions de réplication pour réduire les latences. Les opérations de lecture hautement cohérentes doivent être effectuées vers la seule région principale.



Il est parfois nécessaire de modifier la région active en réponse à une défaillance régionale, comme nous le verrons plus loin. Certains utilisateurs modifient régulièrement la région actuellement active, par exemple en mettant en œuvre un follow-the-sundéploiement. Cela place la région active à proximité de la zone géographique la plus active (généralement là où il fait jour, d'où son nom), ce qui se traduit par la latence la plus faible des opérations de lecture et d'écriture. Cela présente également l'avantage d'appeler quotidiennement le code qui change de région et de s'assurer qu'il est bien testé avant toute reprise après sinistre.

La ou les régions passives peuvent conserver une infrastructure réduite autour de DynamoDB qui ne se développe que si elle devient la région active. Ce guide ne couvre pas les modèles de veilleuse et

de veille chaude. Pour plus d'informations, vous pouvez lire le billet de blog Disaster Recovery (DR) surAWS, Part III : Pilot Light and Warm Standby.

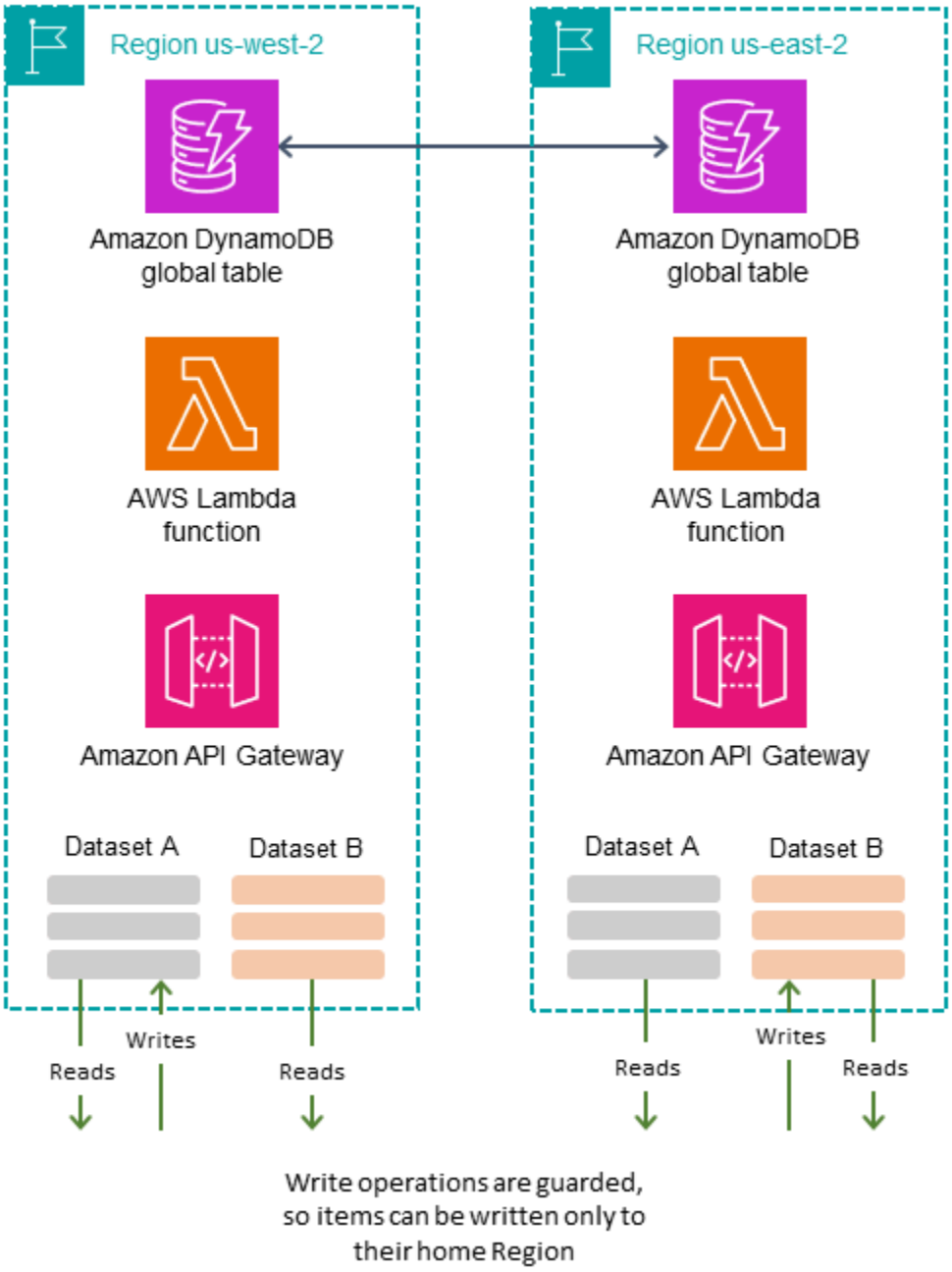
L'utilisation du mode d'écriture dans une région fonctionne bien lorsque vous utilisez des tables globales pour des opérations de lecture à faible latence et distribuées dans le monde entier. Prenons l'exemple d'une grande entreprise de réseaux sociaux qui doit disposer des mêmes données de référence dans toutes les régions du monde. Ils ne mettent pas souvent à jour les données, mais lorsqu'ils le font, ils écrivent dans une seule région afin d'éviter tout conflit d'écriture potentiel. Les opérations de lecture sont toujours autorisées depuis n'importe quelle région.

À titre d'autre exemple, prenons l'exemple de la société de services financiers évoquée précédemment qui a mis en œuvre le calcul des remises en argent quotidiennes. Ils ont utilisé le mode écriture vers n'importe quelle région pour calculer le solde, mais le mode écriture vers une région pour suivre les remboursements. S'ils veulent récompenser un centime pour chaque tranche de 10$ dépensée, ils doivent, `Query` pour toutes les transactions effectuées la veille, calculer le total dépensé, inscrire la décision de remboursement dans un nouveau tableau, supprimer l'ensemble d'articles demandé pour les marquer comme consommés, et les remplacer par un article unique contenant le reste qui devrait être pris en compte dans les calculs du lendemain. Ce travail nécessite des transactions. Il fonctionne donc mieux avec le mode d'écriture dans une seule région. Une application peut mélanger les modes d'écriture, même sur la même table, tant que les charges de travail ne risquent pas de se chevaucher.

Mode Écrire dans votre région (primaire mixte)

Le mode d'écriture dans votre région attribue différents sous-ensembles de données à différentes régions d'origine et autorise les opérations d'écriture sur un élément uniquement via sa région d'origine. Ce mode est actif-passif mais attribue la région active en fonction de l'élément. Chaque région est principale pour son propre jeu de données qui ne se chevauche pas, et les opérations d'écriture doivent être protégées pour garantir une localisation correcte.

Ce mode est similaire à l'écriture dans une région, sauf qu'il permet des opérations d'écriture à faible latence, car les données associées à chaque utilisateur peuvent être placées plus près du réseau par rapport à cet utilisateur. Cela permet également de répartir l'infrastructure environnante de manière plus uniforme entre les régions et de réduire le travail de mise en place de l'infrastructure lors d'un scénario de basculement, car une partie de l'infrastructure de toutes les régions est déjà active.



Vous pouvez déterminer la région d'origine des objets de différentes manières :

- Intrinsèque : certains aspects des données, tels qu'un attribut spécial ou une valeur incorporée dans leur clé de partition, indiquent clairement leur région d'origine. Cette technique est décrite dans le billet de blog Utiliser l'épinglage par région pour définir une région d'accueil pour les articles d'un tableau global Amazon DynamoDB.

- Négocié : la région d'origine de chaque jeu de données est négociée d'une manière externe, par exemple avec un service global distinct qui gère les attributions. La cession peut avoir une durée limitée après laquelle elle peut faire l'objet d'une renégociation.
- Orienté vers les tables : au lieu de créer une seule table globale de réplication, vous créez le même nombre de tables globales que les régions de réplication. Le nom de chaque table indique sa région d'origine. Dans les opérations standard, toutes les données sont écrites dans la région d'origine tandis que les autres régions conservent une copie en lecture seule. Lors d'un basculement, une autre région adopte temporairement des fonctions d'écriture pour cette table.

Par exemple, imaginez que vous travaillez pour une société de jeux vidéo. Vous avez besoin d'opérations de lecture et d'écriture à faible latence pour tous les joueurs du monde entier. Vous attribuez à chaque joueur la région la plus proche de lui. Cette région prend en charge toutes ses opérations de lecture et d'écriture, garantissant ainsi une forte read-after-write cohérence. Toutefois, lorsqu'un joueur voyage ou si sa région d'origine est en panne, une copie complète de ses données est disponible dans d'autres régions, et le joueur peut être affecté à une autre région d'origine.

Autre exemple, imaginez que vous travaillez dans une entreprise de visioconférence. Les métadonnées de chaque conférence téléphonique sont attribuées à une région particulière. Les appelants peuvent utiliser la région la plus proche de chez eux pour minimiser la latence. En cas de panne de région, l'utilisation de tables globales permet une restauration rapide car le système peut déplacer le traitement de l'appel vers une autre région où une copie répliquée des données existe déjà.

Stratégies de routage pour les tables globales

La partie la plus complexe d'un déploiement de tables globales est peut-être la gestion du routage des demandes. Les demandes doivent d'abord aller d'un utilisateur final vers une région choisie et acheminée d'une manière ou d'une autre. La demande rencontre une pile de services dans cette région, notamment une couche de calcul qui consiste peut-être en un équilibreur de charge soutenu par une AWS Lambda fonction, un conteneur ou un nœud Amazon Elastic Compute Cloud (AmazonEC2), et éventuellement d'autres services, y compris peut-être une autre base de données. Cette couche de calcul communique avec DynamoDB. Pour ce faire, il doit utiliser le point de terminaison local de cette région. Les données de la table globale sont répliquées dans toutes les autres régions participantes et chaque région dispose d'une pile de services similaire autour de sa table DynamoDB.

La table globale fournit à chaque pile des différentes régions une copie locale des mêmes données. Vous pouvez envisager de concevoir une pile unique dans une seule région et prévoir de passer des appels distants vers le point de terminaison DynamoDB d'une région secondaire en cas de problème avec la table DynamoDB locale. Il ne s'agit pas d'une bonne pratique. Les latences associées au passage d'une région à une autre peuvent être 100 fois plus élevées que pour l'accès local. Une back-and-forth série de 5 requêtes peut prendre quelques millisecondes lorsqu'elle est exécutée localement, mais quelques secondes lorsqu'elle traverse le globe. Il est préférable d'acheminer le traitement de l'utilisateur final vers une autre région. Pour garantir la résilience, vous avez besoin d'une réplication sur plusieurs régions : réplication de la couche de calcul et de la couche de données.

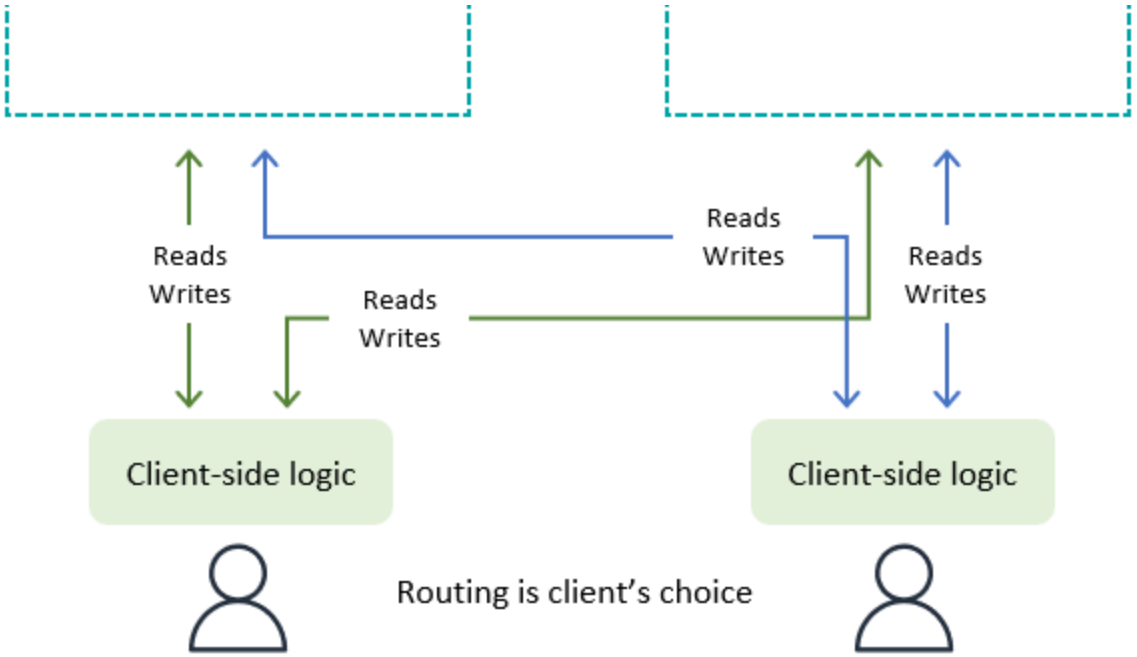
Il existe de nombreuses techniques pour acheminer une demande d'utilisateur final vers une région en vue de son traitement. Le bon choix dépend de votre mode d'écriture et de vos considérations relatives au basculement. Cette section décrit quatre options : axée sur le client, couche informatique, Amazon Route 53 et. AWS Global Accelerator

Rubriques

- Routage des demandes piloté par le client
- Routage des demandes dans la couche de calcul
- Routage des demandes avec Route 53
- Routage des demandes avec Global Accelerator

Routage des demandes piloté par le client

Avec le routage des demandes piloté par le client, le client de l'utilisateur final (une application, une page Web avec ou un autre client) garde une trace des points de terminaison d'application valides (par exemple JavaScript, un point de terminaison Amazon API Gateway plutôt qu'un point de terminaison DynamoDB littéral) et utilise sa propre logique intégrée pour choisir la région avec laquelle communiquer. Il peut choisir en fonction d'une sélection aléatoire, des latences les plus faibles observées, des mesures de bande passante les plus élevées observées ou des contrôles de santé effectués localement.



L'avantage est que le routage des demandes piloté par le client peut s'adapter à des facteurs tels que les conditions réelles du trafic Internet public pour changer de région s'il constate une dégradation des performances. Le client doit connaître tous les points de terminaison potentiels, mais il n'est pas courant de lancer un nouveau point de terminaison régional.

Avec le mode d'écriture dans n'importe quelle région, un client peut sélectionner unilatéralement son point de terminaison préféré. Si son accès à une région est perturbé, le client peut effectuer un routage vers un autre point de terminaison.

Avec le mode écriture dans une région, le client a besoin d'un mécanisme pour acheminer ses demandes d'écriture vers la région actuellement active. Il pourrait s'agir d'un mécanisme de base, comme tester empiriquement quelle région accepte actuellement les demandes écrites (en notant tout rejet d'une demande d'écriture et en revenant à une autre). Il peut également s'agir d'un

mécanisme complexe, tel que l'utilisation d'un coordinateur global pour demander l'état actuel de l'application (peut-être basé sur le contrôle de routage Amazon Application Recovery Controller (ARCARC) (), qui fournit un système à cinq régions piloté par quorum pour maintenir l'état global pour des besoins tels que celui-ci). Le client peut décider si les demandes de lecture peuvent être envoyées à n'importe quelle région pour une cohérence éventuelle ou doivent être acheminées vers la région active pour une meilleure cohérence.

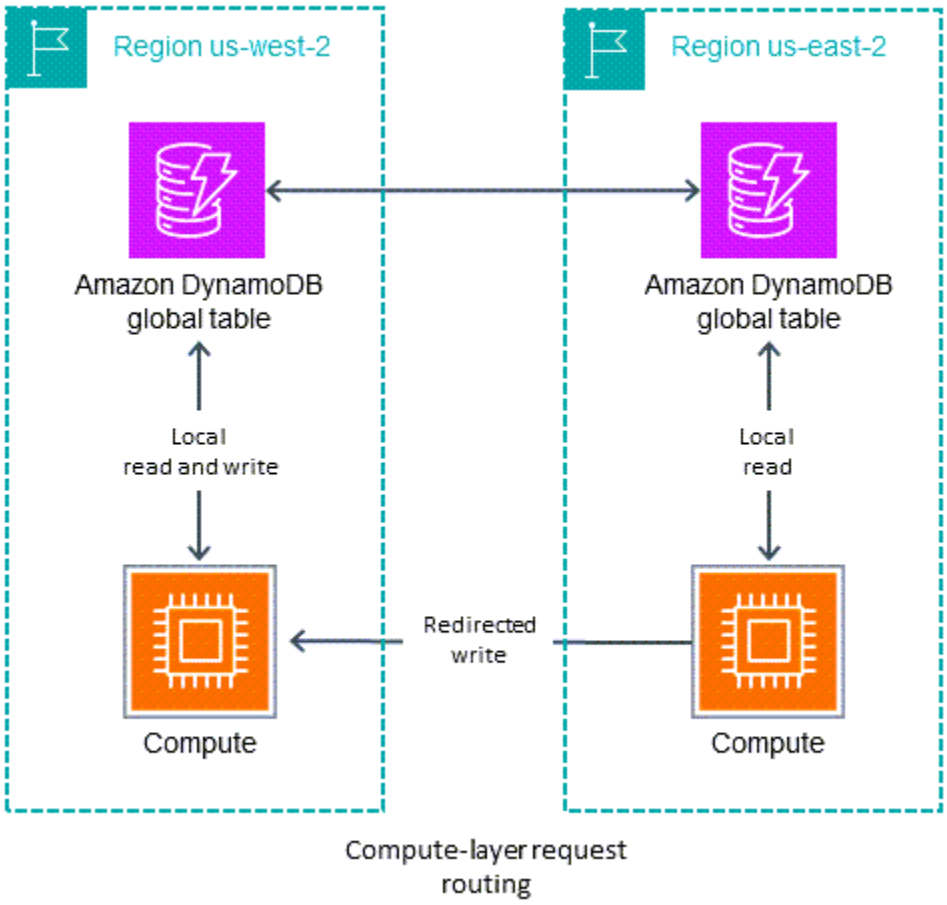
Avec le mode écriture dans votre région, le client doit déterminer la région d'origine du jeu de données avec lequel il travaille. Par exemple, si le client correspond à un compte utilisateur et que chaque compte utilisateur est associé à une région, le client peut demander l'attribution de point de terminaison appropriée à utiliser avec ses informations d'identification auprès d'un système de connexion global.

Par exemple, une société de services financiers qui aide les utilisateurs à gérer les finances de leur entreprise via le Web utilise des tables globales avec un mode écriture dans votre région. Chaque utilisateur doit se connecter à un service central. Ce service renvoie les informations d'identification ainsi que le point de terminaison de la région où ces informations d'identification fonctionneront. La région renvoyée est basée sur l'endroit où se trouve actuellement l'ensemble de données de l'utilisateur. Les informations d'identification sont valides pendant une courte période. Ensuite, la page Web négocie automatiquement une nouvelle connexion, ce qui permet de rediriger potentiellement l'activité de l'utilisateur vers une nouvelle région.

Routage des demandes dans la couche de calcul

Avec le routage des demandes par couche informatique, le code qui s'exécute dans la couche informatique détermine s'il convient de traiter la demande localement ou de la transmettre à une copie de lui-même exécutée dans une autre région. Lorsque vous utilisez le mode écriture dans une région, la couche de calcul peut détecter qu'il ne s'agit pas de la région active et autoriser les opérations de lecture locales tout en transférant toutes les opérations d'écriture vers une autre région. Ce code de couche de calcul doit tenir compte de la topologie des données et des règles de routage, et les appliquer de manière fiable, en fonction des derniers paramètres qui spécifient quelles régions sont actives pour quelles données. La pile logicielle externe au sein de la région n'a pas besoin de savoir comment les demandes de lecture et d'écriture sont acheminées par le microservice. Dans une conception robuste, la région réceptrice vérifie s'il s'agit de la région principale actuelle pour l'opération d'écriture. Si ce n'est pas le cas, cela génère une erreur indiquant que l'état global doit être corrigé. La région réceptrice peut également mettre en mémoire tampon l'opération d'écriture pendant un certain temps si la région principale est en cours de modification. Dans tous les cas, la pile de

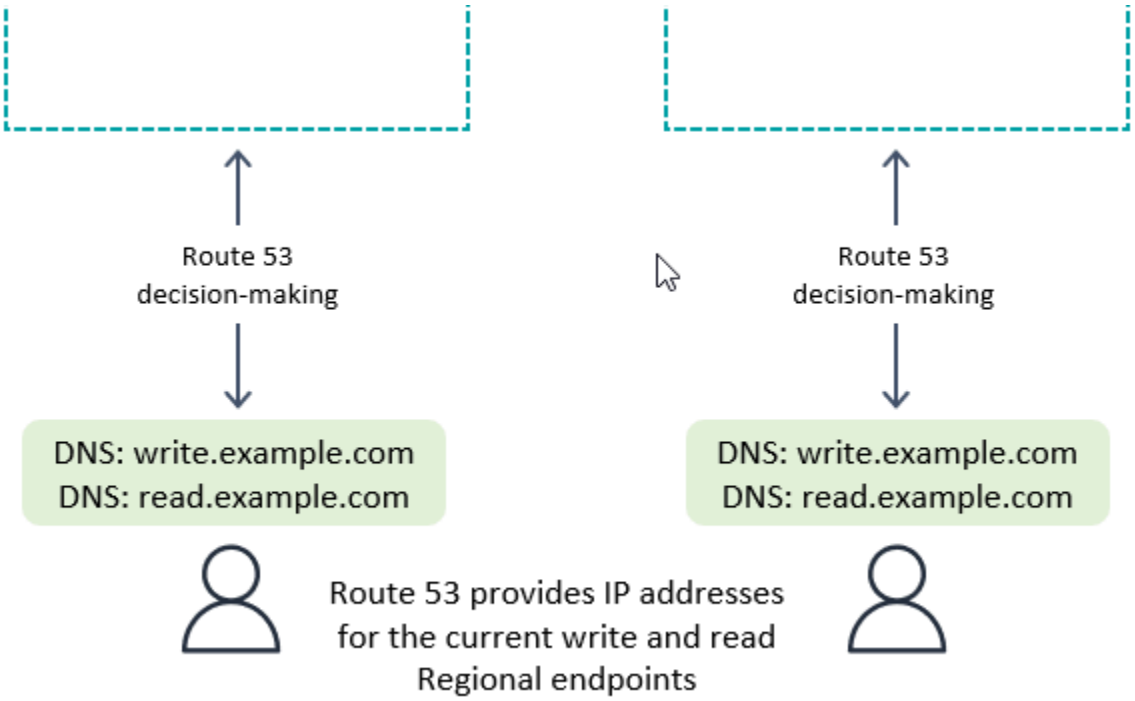
calcul d'une région écrit uniquement sur son point de terminaison DynamoDB local, mais les piles de calcul peuvent communiquer entre elles.



Le Vanguard Group utilise un système appelé Global Orchestration and Status Tool (GOaST) et une bibliothèque appelée Global Multi-Region library (GMRlib) pour ce processus de routage, tel que présenté à re:Invent 2022. Ils utilisent un follow-the-sun seul modèle principal. GOaSTmaintient l'état global, de la même manière que le contrôle de ARC routage décrit dans la section précédente. Il utilise un tableau global pour savoir quelle région est la région principale et quand le prochain changement principal est planifié. Toutes les opérations de lecture et d'écriture sont effectuéesGMRlib, ce qui est coordonné avecGOaST. GMRlibpermet d'effectuer des opérations de lecture localement, avec une faible latence. Pour les opérations d'écriture, GMRlib vérifie si la région locale est la région principale actuelle. Si tel est le cas, l'opération d'écriture se termine directement. Dans le cas contraire, GMRlib transmet la tâche d'écriture GMRlib à la région principale. Cette bibliothèque réceptrice confirme qu'elle se considère également comme la région principale et génère une erreur si ce n'est pas le cas, ce qui indique un délai de propagation par rapport à l'état global. Cette approche offre un avantage en matière de validation car elle ne permet pas d'écrire directement sur un point de terminaison DynamoDB distant.

Routage des demandes avec Route 53

Amazon Route 53 est une technologie de service de noms de domaine (DNS). Avec Route 53, le client demande son point de terminaison en recherchant un nom de DNS domaine connu, et Route 53 renvoie l'adresse IP correspondant au ou aux points de terminaison régionaux qu'il juge les plus appropriés. La Route 53 possède une longue liste de politiques de routage qu'elle utilise pour déterminer la région appropriée. Il peut également effectuer un routage sur incident pour acheminer le trafic hors des régions qui échouent aux tests de santé.



Avec le mode écriture dans n'importe quelle région, ou s'il est combiné au routage des demandes par la couche informatique sur le backend, Route 53 peut être totalement libre de renvoyer la région sur la base de règles internes complexes, telles que le choix de la région dans le réseau ou la proximité géographique le plus proche, ou tout autre choix.

Avec le mode écriture dans une région, vous pouvez configurer Route 53 pour renvoyer la région actuellement active (en utilisantARC). Si le client souhaite se connecter à une région passive (par exemple, pour des opérations de lecture), il peut rechercher un autre DNS nom.

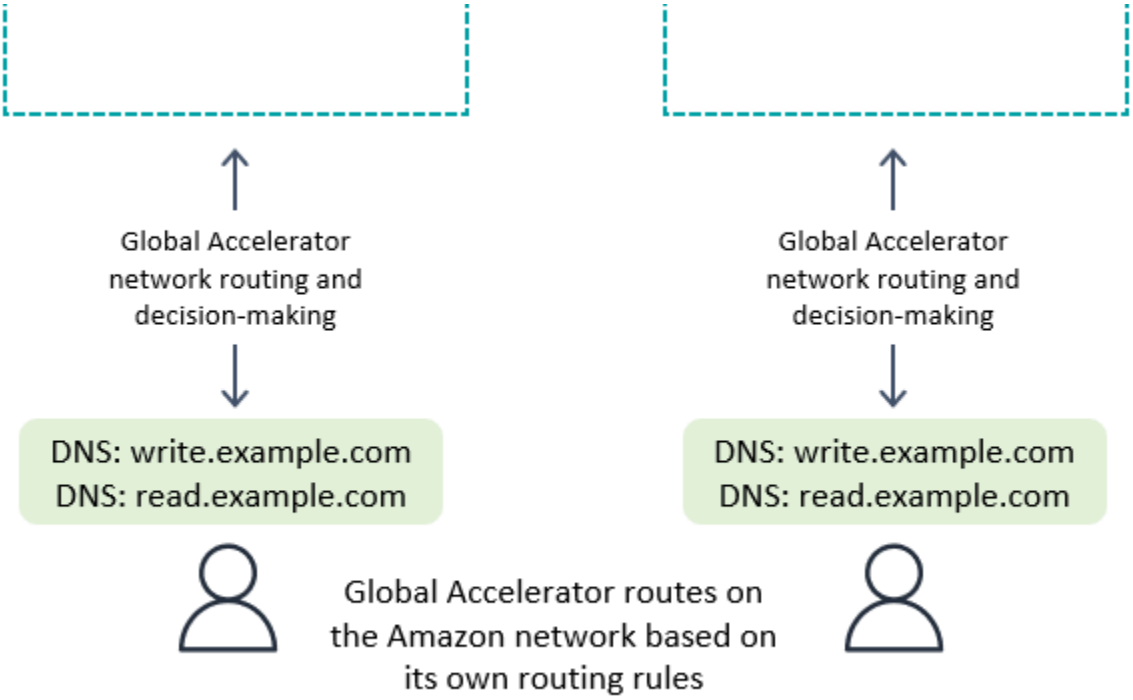
> **Note**
>
> Les clients mettent en cache les adresses IP dans la réponse de Route 53 pendant une durée indiquée par le paramètre time to live (TTL) du nom de domaine. Une durée plus

longue TTL prolonge l'objectif de temps de restauration (RTO) pour que tous les clients reconnaissent le nouveau point de terminaison. Une valeur de 60 secondes est généralement utilisée pour un basculement. Tous les logiciels ne respectent pas parfaitement le DNS TTL délai d'expiration et il peut y avoir plusieurs niveaux de DNS mise en cache, par exemple au niveau du système d'exploitation, de la machine virtuelle et de l'application.

En mode écriture dans votre région, il est préférable d'éviter la Route 53, sauf si vous utilisez également le routage des demandes par couche informatique.

Routage des demandes avec Global Accelerator

Avec AWS Global Acceleratorun client, recherche le nom de domaine bien connu dans Route 53. Toutefois, au lieu de récupérer une adresse IP correspondant à un point de terminaison régional, le client obtient une adresse IP statique anycast qui achemine vers l'emplacement AWS périphérique le plus proche. À partir de cet emplacement périphérique, tout le trafic est acheminé sur le AWS réseau privé vers un point de terminaison (équilibreurs de charge réseau, équilibreurs de charge d'application, EC2 instances ou adresses IP élastiques) dans une région choisie par des règles de routage maintenues dans Global Accelerator. Comparé au routage basé sur les règles Route 53, le routage des demandes avec Global Accelerator présente des latences plus faibles car il réduit le volume de trafic sur l'Internet public. En outre, étant donné que Global Accelerator ne dépend pas de l'DNSTTLexpiration pour modifier les règles de routage, il peut ajuster le routage plus rapidement.



Avec le mode écriture dans n'importe quelle région, ou s'il est combiné au routage des demandes par la couche informatique sur le backend, Global Accelerator fonctionne parfaitement. Le client se connecte à l'emplacement périphérique le plus proche et n'a pas à se soucier de savoir quelle région reçoit la demande.

Avec le mode écriture dans une région, les règles de routage de Global Accelerator doivent envoyer des demandes à la région actuellement active. Vous pouvez utiliser des surveillances de l'état qui signalent artificiellement une défaillance dans une région qui n'est pas considérée par votre système global comme étant la région active. De mêmeDNS, il est possible d'utiliser un autre nom de DNS domaine pour acheminer les demandes de lecture, si celles-ci peuvent provenir de n'importe quelle région.

En mode écriture dans votre région, il est préférable d'éviter Global Accelerator, sauf si vous utilisez également le routage des demandes par couche informatique.

processus d'évacuation pour les tables globales

L'évacuation d'une région est le processus de migration d'une activité (généralement une activité d'écriture, éventuellement une activité de lecture) hors de cette région.

Évacuation d'une région active

Vous pouvez décider d'évacuer une région active pour plusieurs raisons : dans le cadre de vos activités commerciales habituelles (par exemple, si vous utilisez un mode écrire dans une région)follow-the-sun, en raison d'une décision commerciale visant à modifier la région actuellement active, en réponse à des défaillances de la suite logicielle en dehors de DynamoDB ou parce que vous rencontrez des problèmes généraux tels que des latences plus élevées que d'habitude au sein de la région.

Avec le mode Écrire dans n'importe quelle région, il est facile d'évacuer une région active. Vous pouvez acheminer le trafic vers d'autres régions en utilisant n'importe quel système de routage et laisser les opérations d'écriture déjà effectuées dans la région évacuée se répliquer comme d'habitude.

Avec les modes écriture dans une région et écriture dans votre région, vous devez vous assurer que toutes les opérations d'écriture dans la région active ont été entièrement enregistrées, traitées en flux et propagées globalement avant de commencer les opérations d'écriture dans la nouvelle région active, afin de garantir que les opérations d'écriture future seront traitées par rapport à la dernière version des données.

Supposons que la région A soit active et que la région B soit passive (soit pour la table complète, soit pour les éléments qui se trouvent dans la région A). Le mécanisme habituel pour une évacuation consiste à suspendre les opérations d'écriture vers A, à attendre suffisamment longtemps pour que ces opérations se propagent entièrement vers B, à mettre à jour la pile d'architecture pour reconnaître B comme étant active, puis à reprendre les opérations d'écriture vers B. Aucune métrique n'indique avec une certitude absolue que la région A a entièrement répliqué ses données vers la région B. Si la région A est saine, suspendre les opérations d'écriture dans la région A et attendre 10 fois la valeur maximale récente de la métrique `ReplicationLatency` serait généralement suffisant pour déterminer si la réplication est terminée. Si la région A n'est pas saine et indique d'autres zones présentant des latences accrues, vous devez choisir un multiple plus élevé pour le temps d'attente.

Évacuation d'une région déconnectée

Il y a un cas particulier à prendre en compte : et si la région A était complètement déconnectée sans préavis ? C'est très peu probable, mais il faut tout de même y réfléchir. Dans ce cas, toutes les opérations d'écriture dans la région A qui n'ont pas encore été propagées sont conservées et propagées après la remise en ligne de la région A. Les opérations d'écriture ne sont pas perdues, mais leur propagation est retardée indéfiniment.

La procédure à suivre dans ce cas dépend de la décision de l'application. Pour la continuité des activités, les opérations d'écriture peuvent devoir être effectuées vers la nouvelle région principale B. Toutefois, si un élément de la région B reçoit une mise à jour alors qu'une opération d'écriture pour cet élément est en attente de propagation depuis la région A, la propagation est supprimée selon le modèle victoire du dernier auteur. Toute mise à jour dans la région B peut supprimer une demande d'écriture entrante.

En mode écriture vers n'importe quelle région, les opérations de lecture et d'écriture peuvent se poursuivre dans la région B, en sachant que les éléments de la région A finiront par se propager dans la région B et en reconnaissant la possibilité que des éléments soient manquants jusqu'à ce que la région A revienne en ligne. Dans la mesure du possible, par exemple pour les opérations d'écriture idempotentes, vous devez envisager de rejouer le trafic d'écriture récent (par exemple, en utilisant une source d'événements en amont) afin de combler les lacunes liées aux opérations d'écriture potentiellement manquantes et de laisser le dernier auteur gagner la résolution du conflit pour supprimer la propagation éventuelle de l'opération d'écriture entrante.

Avec les autres modes d'écriture, vous devez tenir compte de la mesure dans laquelle le travail peut se poursuivre avec une légère out-of-date vision du monde. Certaines opérations d'écriture de courte durée, telles que suivies par `ReplicationLatency`, sont absentes jusqu'à ce que la région A revienne en ligne. Les entreprises peuvent-elles aller de l'avant ? Dans certains cas d'utilisation, c'est possible, mais dans d'autres, pas sans des mécanismes d'atténuation supplémentaires.

Imaginons par exemple que vous deviez maintenir un solde créditeur disponible sans interruption, même après une panne complète d'une région. Vous pouvez diviser le solde en deux objets différents, l'un réservé à la Région A et l'autre à la Région B, et commencer chacun avec la moitié du solde disponible. Cela utiliserait le mode Écrire dans votre région. Les mises à jour transactionnelles traitées dans chaque région seraient comparées à la copie locale du solde. Si la région A est complètement déconnectée, le traitement des transactions pourrait toujours se poursuivre dans la région B et les opérations d'écriture seraient limitées à la partie du solde détenue dans la région B. Le fractionnement du solde introduit des difficultés lorsque le solde baisse ou que le crédit doit être

rééquilibré, mais cela fournit un exemple de reprise de l'entreprise sûre, même en cas d'opérations d'écriture incertaines en cours.

Autre exemple, imaginez que vous capturez des données de formulaire Web. Vous pouvez utiliser le contrôle de concurrence optimiste (OCC) pour attribuer des versions aux éléments de données et intégrer la dernière version dans le formulaire Web en tant que champ masqué. À chaque envoi, l'opération d'écriture ne réussit que si la version de la base de données correspond toujours à la version avec laquelle le formulaire a été créé. Si les versions ne correspondent pas, le formulaire Web peut être actualisé (ou soigneusement fusionné) avec la version actuelle de la base de données et l'utilisateur peut recommencer. Le modèle OCC offre généralement une protection contre l'écrasement par un autre client et la production d'une nouvelle version des données, mais il peut également être utile en cas de basculement lorsqu'un client peut rencontrer d'anciennes versions des données. Imaginons que vous utilisez l'horodatage comme version. Le formulaire a d'abord été créé pour la région A à midi mais (après le basculement) essaie d'écrire dans la région B et remarque que la dernière version de la base de données est 11:59. Dans ce scénario, le client peut soit attendre que la version de 12 h 00 se propage vers la région B, puis écrire sur cette version, soit utiliser la version de 11 h 59 et créer une nouvelle version à 12 h 01 (qui, après écriture, supprime la version entrante une fois la région A rétablie).

Troisième exemple : une société de services financiers conserve les données relatives aux comptes clients et à leurs transactions financières dans une base de données DynamoDB. En cas de panne complète dans la région A, ils veulent s'assurer que toute activité d'écriture liée à leurs comptes est entièrement disponible dans la région B, ou ils souhaitent mettre leurs comptes en quarantaine comme étant considérés comme partiels jusqu'à ce que la région A soit à nouveau en ligne. Au lieu de suspendre toutes les activités, elle a décidé de suspendre uniquement celles de l'infime fraction des comptes qui, selon elle, contenaient des transactions non propagées. Pour ce faire, elle a utilisé une troisième région, que nous appellerons région C. Avant de traiter les opérations d'écriture dans la région A, elle a placé un résumé succinct des opérations en attente (par exemple, un nouveau nombre de transactions pour un compte) dans la région C. Ce résumé était suffisant pour permettre à la région B de déterminer si sa vue était entièrement à jour. Cette action a effectivement verrouillé le compte entre le moment de l'écriture dans la région C et le moment où la région A a accepté les opérations d'écriture et que la région B les a reçues. Les données de la région C n'ont pas été utilisées, sauf dans le cadre d'un processus de basculement, après quoi la région B a pu recouper ses données avec la région C pour vérifier si l'un de ses comptes était obsolète. Ces comptes seraient marqués comme étant mis en quarantaine jusqu'à ce que la restauration de la région A propage les données partielles vers la région B. En cas de défaillance de la région C, une nouvelle région D pourrait être créée pour être utilisée à la place. Les données de la région C étaient très

transitoires et, au bout de quelques minutes, la région D disposerait d'un up-to-date enregistrement suffisant des opérations d'écriture en vol pour être pleinement utile. En cas d'échec de la région B, la région A pourrait continuer à accepter les demandes d'écriture en coopération avec la région C. Cette entreprise était prête à accepter des écritures à latence plus élevée (vers deux régions : C puis A) et a eu la chance de disposer d'un modèle de données permettant de résumer succinctement l'état d'un compte.

Planification de la capacité de débit pour les tables globales

La migration du trafic d'une région vers une autre nécessite un examen attentif des paramètres de la table DynamoDB concernant la capacité.

Voici quelques considérations relatives à la gestion de la capacité d'écriture :

- Une table globale doit être en mode à la demande ou provisionné avec Auto Scaling activé.
- En cas de provisionnement avec Auto Scaling, les paramètres d'écriture (utilisation minimale, maximale et cible) sont répliqués entre les régions. Bien que les paramètres d'Auto Scaling soient synchronisés, la capacité d'écriture réellement provisionnée peut varier indépendamment entre les régions.
- L'une des raisons pour lesquelles vous pouvez constater une capacité d'écriture allouée différente est due à la fonctionnalité time to live (TTL). Lorsque vous l'activez TTL dans DynamoDB, vous pouvez spécifier un nom d'attribut dont la valeur indique l'heure d'expiration de l'élément, au format Epoch Time Unix en secondes. Passé ce délai, DynamoDB peut supprimer l'élément sans frais d'écriture. Avec les tables globales, vous pouvez configurer TTL dans n'importe quelle région, et le paramètre est automatiquement répliqué dans les autres régions associées à la table globale. Lorsqu'un article est éligible à la suppression par le biais d'une TTL règle, ce travail peut être effectué dans n'importe quelle région. L'opération de suppression est effectuée sans consommer d'unités d'écriture sur la table source, mais les tables répliquées recevront une écriture répliquée de cette opération de suppression et entraîneront des coûts unitaires d'écriture répliqués.
- Si vous utilisez le dimensionnement automatique, assurez-vous que le paramètre de capacité d'écriture maximale allouée est suffisamment élevé pour gérer toutes les opérations d'écriture ainsi que toutes les opérations de TTL suppression potentielles. Auto Scaling ajuste chaque région en fonction de sa consommation d'écriture. Les tables à la demande n'ont pas de paramètre de capacité d'écriture maximale provisionnée, mais la limite de débit d'écriture maximal au niveau de la table indique la capacité d'écriture soutenue maximale autorisée par la table à la demande. La limite par défaut est de 40 000, mais elle peut être ajustée. Nous vous recommandons de le définir suffisamment haut pour gérer toutes les opérations d'écriture (y compris les opérations d'TTLécriture) dont la table à la demande peut avoir besoin. Cette valeur doit être la même dans toutes les régions participantes lorsque vous configurez des tables globales.

Voici quelques considérations relatives à la gestion de la capacité de lecture :

- Les paramètres de gestion de la capacité de lecture peuvent différer entre les régions car il est supposé que les différentes régions peuvent avoir des modèles de lecture indépendants. La première fois que vous ajoutez un réplica global à une table, la capacité de la région source est propagée. Après sa création, vous pouvez ajuster les paramètres de capacité de lecture, qui ne sont pas transférés de l'autre côté.
- Lorsque vous utilisez Auto Scaling de DynamoDB, veillez à ce que les paramètres de capacité de lecture maximale allouée soient suffisamment élevés pour gérer toutes les opérations de lecture dans toutes les régions. Au cours des opérations standard, la capacité de lecture peut être répartie entre les régions, mais lors du basculement, la table doit pouvoir s'adapter automatiquement à l'augmentation de la charge de travail de lecture. Les tables à la demande n'ont pas de paramètre de capacité de lecture maximale provisionnée, mais la limite de débit de lecture maximal au niveau de la table indique la capacité de lecture soutenue maximale autorisée par la table à la demande. La limite par défaut est de 40 000, mais elle peut être ajustée. Nous vous recommandons de la définir à un niveau suffisamment élevé pour gérer toutes les opérations de lecture dont la table pourrait avoir besoin si toutes les opérations de lecture devaient être acheminées vers cette région unique.
- Si une table d'une région ne reçoit généralement pas de trafic de lecture mais qu'elle doit en absorber une grande partie après un basculement, vous pouvez augmenter la capacité de lecture allouée à la table, attendre la fin de la mise à jour de la table, puis la réapprovisionner. Vous pouvez laisser la table en mode provisionné ou la passer en mode à la demande. Cela permet de préchauffer la table pour qu'elle accepte un niveau de trafic de lecture plus élevé.

ARCpropose des contrôles de préparation qui peuvent être utiles pour confirmer que les régions DynamoDB ont des paramètres de table et des quotas de compte similaires, que vous utilisiez Route 53 pour acheminer les demandes ou non. Ces contrôles de préparation vous aident également à ajuster les quotas au niveau du compte pour qu'ils correspondent.

Liste de contrôle pour la préparation des tableaux globaux

Utilisez la liste de contrôle suivante pour les décisions et les tâches lorsque vous déployez des tables globales.

- Déterminez combien et quelles régions doivent participer à la table globale.
- Déterminez le mode d'écriture de votre application.
- Planifiez votre stratégie de routage en fonction de votre mode d'écriture.
- Définissez votre plan d'évacuation en fonction de votre mode d'écriture et de votre stratégie de routage.
- Capturez des indicateurs sur l'état, la latence et les erreurs dans chaque région. Pour obtenir la liste des métriques DynamoDB, consultez AWS le billet de blog Monitoring Amazon DynamoDB pour une prise de conscience opérationnelle. Vous devez également utiliser des canaris synthétiques (requêtes artificielles conçues pour détecter les défaillances) ainsi que l'observation en direct du trafic client. Les problèmes n'apparaissent pas tous dans les métriques DynamoDB.
- Réglez des alarmes en cas d'une augmentation soutenue de la `ReplicationLatency`. Une augmentation peut indiquer une mauvaise configuration accidentelle dans laquelle la table globale possède des paramètres d'écriture différents selon les régions, ce qui entraîne l'échec des demandes répliquées et une augmentation des latences. Cela pourrait également indiquer qu'il y a une perturbation régionale. Un bon exemple est de générer une alerte si la moyenne récente dépasse 180 000 millisecondes. Vous pouvez également surveiller la `ReplicationLatency` chute à 0, ce qui indique un blocage de la réplication.
- Attribuez des paramètres de lecture et d'écriture maximaux suffisants pour chaque table globale.
- Identifiez les conditions dans lesquelles vous évacueriez une région. Si la décision implique un jugement humain, documentez toutes les considérations. Ce travail doit être effectué avec soin à l'avance, sans stress.
- Conservez un runbook pour chaque action qui doit avoir lieu lorsque vous évacuez une région. En général, très peu de travail est nécessaire pour les tables globales, mais le déplacement du reste de la pile peut s'avérer complexe.

> **Note**
>
> En ce qui concerne les procédures de basculement, il est recommandé de s'appuyer uniquement sur les opérations du plan de données et non sur les opérations du plan de contrôle, car certaines opérations du plan de contrôle peuvent être dégradées lors de

défaillances régionales. Pour plus d'informations, consultez le billet de AWS blog Créer des applications résilientes avec les tables globales Amazon DynamoDB : partie 4.

- Testez régulièrement tous les aspects du runbook, y compris les évacuations régionales. Un runbook non testé est un runbook peu fiable.
- Envisagez AWS Resilience Hubde l'utiliser pour évaluer la résilience de l'ensemble de votre application (y compris les tables globales). Ce service fournit une vue complète de l'état de résilience de votre portefeuille d'applications via son tableau de bord.
- Envisagez d'utiliser des contrôles de ARCpréparation pour évaluer la configuration actuelle de votre application et suivre tout écart par rapport aux meilleures pratiques.
- Lorsque vous rédigez des bilans de santé à utiliser avec Route 53 ou Global Accelerator, effectuez une série d'appels couvrant l'ensemble du flux de base de données. Si vous limitez votre vérification pour confirmer uniquement que le point de terminaison DynamoDB est actif, vous ne serez pas en mesure de couvrir de nombreux modes de défaillance AWS Identity and Access Management tels que IAM () les erreurs de configuration, les problèmes de déploiement du code, les défaillances de la pile en dehors de DynamoDB, les latences de lecture ou d'écriture supérieures à la moyenne, etc.

Questions fréquentes (FAQ) sur les tables globales

Cette section contient les réponses aux questions fréquentes sur les tables globales DynamoDB.

Quel est le coût des tables globales ?

- Le coût d'une opération d'écriture dans une table DynamoDB classique est exprimé en unités de capacité d'écriture (WCU) pour les tables provisionnées ou en unités de demande d'écriture (WRU) pour les tables à la demande. Si vous écrivez un élément de 5 Ko, il implique des frais de 5 unités. Le coût d'une écriture dans une table globale est exprimé en unités de capacité d'écriture répliquée (rWCU) pour les tables provisionnées ou en unités de demande d'écriture répliquée (rWRU) pour les tables à la demande.
- Les rWCU et rWRU incluent le coût de l'infrastructure de diffusion en continu nécessaire à la gestion de la réplication. À ce titre, leur coût est 50 pour cent plus élevé que celui des WCU et des WRU. Des frais de transfert de données entre régions s'appliquent.
- Des frais RWcu et RWru sont facturés dans chaque région où l'élément est écrit directement ou écrit par réplication.
- L'écriture dans un index secondaire global (GSI) est considérée comme une opération d'écriture locale et utilise des unités d'écriture normales.
- Aucune capacité réservée n'est disponible pour les rWCU pour le moment. L'achat de capacité réservée pour les WCU peut toujours être avantageux pour les tables dont les GSI consomment des unités d'écriture.
- Lorsque vous ajoutez une nouvelle région à une table globale, DynamoDB démarre automatiquement la nouvelle région et vous facture comme s'il s'agissait d'une restauration de table, en fonction de la taille en Go de la table. Il facture également des frais de transfert de données entre régions.

Quelles sont les régions prises en charge par les tables globales ?

Les tables globales prennent en charge toutes les Régions AWS.

Comment les index secondaires globaux (GSI) sont-ils gérés avec les tables globales ?

Dans les tables globales (version actuelle de 2019), lorsque vous créez un GSI dans une région, il est automatiquement créé dans les autres régions participantes et automatiquement rempli.

Comment arrêter la réplication d'une table globale ?

Vous pouvez supprimer une table de réplica de la même manière qu'une autre table. La suppression de la table globale arrête la réplication vers cette région et supprime la copie de la table conservée dans cette région. Toutefois, vous ne pouvez pas arrêter la réplication tout en conservant des copies de la table en tant qu'entités indépendantes, ni suspendre la réplication.

Comment Amazon DynamoDB Streams interagit-il avec les tables globales ?

Chaque table globale produit un flux indépendant basé sur toutes ses opérations d'écriture, d'où qu'elles proviennent. Vous pouvez choisir de consommer ce flux DynamoDB dans une région ou dans toutes les régions (indépendamment). Si vous souhaitez traiter des opérations d'écritures locales mais pas répliquées, vous pouvez ajouter votre propre attribut de région à chaque élément afin d'identifier la région d'écriture. Vous pouvez ensuite utiliser un filtre d'événements AWS Lambda pour appeler uniquement la fonction Lambda pour les opérations d'écriture dans la région locale. Cela facilite les opérations d'insertion et de mise à jour, mais pas les opérations de suppression.

Comment les tables globales gèrent-elles les transactions ?

Les opérations transactionnelles offrent des garanties d'atomicité, de cohérence, d'isolement et de durabilité (ACID) uniquement dans la région de laquelle provient l'opération d'écriture. Les transactions ne sont pas prises en charge entre les régions dans les tables globales. Par exemple, si vous avez une table globale avec des réplicas dans les régions USA Est (Ohio) et USA Ouest (Oregon), et que vous réalisez une opération `TransactWriteItems` dans la région USA Est (Ohio), vous remarquerez peut-être des transactions partiellement incomplètes dans la région USA Ouest (Oregon) lorsque les changements sont répliqués. Les modifications seront uniquement répliquées sur les autres régions une fois validées dans la région source.

Comment les tables globales interagissent-elles avec le cache de DynamoDB Accelerator (DAX) ?

Les tables globales contournent le DAX en mettant directement à jour DynamoDB. Ainsi, DAX ne sait pas qu'il contient des données obsolètes. Le cache DAX n'est actualisé que lorsque la durée de vie du cache expire.

Les balises présentes sur les tables se propagent-elles ?

Non, les balises ne se propagent pas automatiquement.

Dois-je sauvegarder des tables dans toutes les régions ou dans une seule ?

La réponse dépend de l'objectif de la sauvegarde.

- Si vous souhaitez garantir la durabilité des données, DynamoDB fournit déjà cette garantie. Le service garantit la durabilité.
- Si vous souhaitez conserver un instantané des enregistrements historiques (par exemple, pour répondre à des exigences réglementaires), une sauvegarde dans une région doit suffire. Vous pouvez copier la sauvegarde vers d'autres régions en utilisant AWS Backup.
- Si vous souhaitez récupérer des données supprimées ou modifiées par erreur, utilisez DynamoDB point-in-time recovery (PITR) dans une région.

Comment déployer des tables globales avec AWS CloudFormation ?

- CloudFormation représente une table DynamoDB et une table globale sous la forme de deux ressources distinctes : et. `AWS::DynamoDB::Table` `AWS::DynamoDB::GlobalTable` Une méthode consiste à créer toutes les tables susceptibles d'être globales à l'aide de la construction `GlobalTable`, à les conserver dans un premier temps sous forme de tables autonomes et à ajouter des régions ultérieurement, si nécessaire.
- Dans CloudFormation, chaque table globale est contrôlée par une seule pile, dans une seule région, quel que soit le nombre de répliques. Lorsque vous déployez votre modèle, il

CloudFormation crée et met à jour toutes les répliques dans le cadre d'une opération de pile unique. Vous ne devez pas déployer la même ressource AWS::DynamoDB::GlobalTable dans plusieurs régions. Cette opération n'est pas prise en charge et entraînera des erreurs. Si vous déployez votre modèle d'application dans plusieurs régions, vous pouvez utiliser des conditions pour ne créer la ressource `AWS::DynamoDB::GlobalTable` que dans une seule région. Vous pouvez également choisir de définir vos ressources `AWS::DynamoDB::GlobalTable` dans une pile distincte de votre pile d'applications et vous assurer qu'elle n'est déployée que dans une seule région.

- Si vous avez une table normale et que vous souhaitez la convertir en table globale tout en la gérant en CloudFormation : définissez la politique de suppression sur`Retain`, supprimez la table de la pile, convertissez-la en table globale dans la console, puis importez la table globale en tant que nouvelle ressource dans la pile. Pour plus d'informations, consultez le AWS GitHub référentiel amazon-dynamodb-table-to- global-table-cdk.
- La réplication entre comptes n'est pas prise en charge pour le moment.

Conclusion et ressources

Les tables globales DynamoDB comportent très peu de contrôles, mais elles nécessitent tout de même un examen attentif. Vous devez déterminer votre mode d'écriture, votre modèle de routage et vos processus d'évacuation. Vous devez équiper votre application dans chaque région et être prêt à ajuster votre routage ou à effectuer une évacuation pour préserver l'état global. La récompense est de disposer d'un ensemble de données distribué dans le monde entier avec des opérations de lecture et d'écriture à faible latence, conçu pour une disponibilité de 99,999 %.

Pour plus d'informations sur les tables globales DynamoDB, consultez les ressources suivantes :

- Documentation Amazon DynamoDB
- Contrôleur de restauration d'applications Amazon Route 53
- ARCvérifications de l'état de préparation (AWS documentation)
- Politiques de routage Route 53 (AWS documentation)
- AWS Global Accelerator
- Contrat de niveau de service DynamoDB
- AWS Principes fondamentaux de plusieurs régions (AWS livre blanc)
- Modèles de conception de résilience des données avec AWS(présentationAWS re:Invent 2022)
- Comment Fidelity Investments et Reltio se sont modernisés avec Amazon DynamoDB AWS (présentation re:Invent 2022)
- Modèles de conception multirégionaux et meilleures pratiques (présentationAWS re:Invent 2022)
- Architecture de reprise après sinistre (DR) activée AWS, en partie III : veilleuse et veille chaude (article de AWS blog)
- Utiliser l'épinglage régional pour définir une région d'origine pour les éléments d'une table globale AWS Amazon DynamoDB (article de blog)
- Surveillance d'Amazon DynamoDB à des fins de connaissance AWS opérationnelle (article de blog)
- Mise à l'échelle de DynamoDB : comment les partitions, les raccourcis clavier et le fractionnement ont un impact sur les performances AWS (article de blog)

Historique du document

Le tableau suivant décrit les modifications importantes apportées à ce guide. Pour être averti des mises à jour à venir, abonnez-vous à un fil RSS.

Modification	Description	Date
AWS Global Accelerator Informations mises à jour	Correction des points de terminaison pour le routage des demandes de Global Accelerator.	14 mars 2024
Informations d' Région AWS assistance mises à jour	Mise à jour des Questions fréquentes (FAQ) pour indiquer que les tables globales prennent désormais en charge toutes les Régions AWS.	15 novembre 2023
Publication initiale	—	19 mai 2023

AWS Glossaire des directives prescriptives

Les termes suivants sont couramment utilisés dans les stratégies, les guides et les modèles fournis par les directives AWS prescriptives. Pour suggérer des entrées, veuillez utiliser le lien Faire un commentaire à la fin du glossaire.

Nombres

7 R

Sept politiques de migration courantes pour transférer des applications vers le cloud. Ces politiques s'appuient sur les 5 R identifiés par Gartner en 2011 et sont les suivantes :

- Refactorisation/réarchitecture : transférez une application et modifiez son architecture en tirant pleinement parti des fonctionnalités natives cloud pour améliorer l'agilité, les performances et la capacité de mise à l'échelle. Cela implique généralement le transfert du système d'exploitation et de la base de données. Exemple : migrez votre base de données Oracle sur site vers l'édition compatible avec Amazon Aurora PostgreSQL.
- Replateformer (déplacer et remodeler) : transférez une application vers le cloud et introduisez un certain niveau d'optimisation pour tirer parti des fonctionnalités du cloud. Exemple : migrez votre base de données Oracle sur site vers Amazon Relational Database Service (Amazon RDS) pour Oracle dans le. AWS Cloud
- Racheter (rachat) : optez pour un autre produit, généralement en passant d'une licence traditionnelle à un modèle SaaS. Exemple : migrez votre système de gestion de la relation client (CRM) vers Salesforce.com.
- Réhéberger (lift and shift) : transférez une application vers le cloud sans apporter de modifications pour tirer parti des fonctionnalités du cloud. Exemple : migrez votre base de données Oracle sur site vers Oracle sur une instance EC2 dans le. AWS Cloud
- Relocaliser (lift and shift au niveau de l'hyperviseur) : transférez l'infrastructure vers le cloud sans acheter de nouveau matériel, réécrire des applications ou modifier vos opérations existantes. Vous migrez des serveurs d'une plateforme sur site vers un service cloud pour la même plateforme. Exemple : migrer une Microsoft Hyper-V application vers AWS.
- Retenir : conservez les applications dans votre environnement source. Il peut s'agir d'applications nécessitant une refactorisation majeure, que vous souhaitez retarder, et d'applications existantes que vous souhaitez retenir, car rien ne justifie leur migration sur le plan commercial.

- Retirer : mettez hors service ou supprimez les applications dont vous n'avez plus besoin dans votre environnement source.

A

ABAC

Voir contrôle d'accès basé sur les attributs.

services abstraits

Consultez la section Services gérés.

ACIDE

Voir atomicité, consistance, isolation, durabilité.

migration active-active

Méthode de migration de base de données dans laquelle la synchronisation des bases de données source et cible est maintenue (à l'aide d'un outil de réplication bidirectionnelle ou d'opérations d'écriture double), tandis que les deux bases de données gèrent les transactions provenant de la connexion d'applications pendant la migration. Cette méthode prend en charge la migration par petits lots contrôlés au lieu d'exiger un basculement ponctuel. Elle est plus flexible mais demande plus de travail qu'une migration active-passive.

migration active-passive

Méthode de migration de base de données dans laquelle la synchronisation des bases de données source et cible est maintenue, mais seule la base de données source gère les transactions provenant de la connexion d'applications pendant que les données sont répliquées vers la base de données cible. La base de données cible n'accepte aucune transaction pendant la migration.

fonction d'agrégation

Fonction SQL qui agit sur un groupe de lignes et calcule une valeur de retour unique pour le groupe. Des exemples de fonctions d'agrégation incluent `SUM` et`MAX`.

AI

Voir intelligence artificielle.

AIOps

Voir les opérations d'intelligence artificielle.

anonymisation

Processus de suppression définitive d'informations personnelles dans un ensemble de données. L'anonymisation peut contribuer à protéger la vie privée. Les données anonymisées ne sont plus considérées comme des données personnelles.

anti-motif

Solution fréquemment utilisée pour un problème récurrent lorsque la solution est contre-productive, inefficace ou moins efficace qu'une solution alternative.

contrôle des applications

Une approche de sécurité qui permet d'utiliser uniquement des applications approuvées afin de protéger un système contre les logiciels malveillants.

portefeuille d'applications

Ensemble d'informations détaillées sur chaque application utilisée par une organisation, y compris le coût de génération et de maintenance de l'application, ainsi que sa valeur métier. Ces informations sont essentielles pour le processus de découverte et d'analyse du portefeuille et permettent d'identifier et de prioriser les applications à migrer, à moderniser et à optimiser.

intelligence artificielle (IA)

Domaine de l'informatique consacré à l'utilisation des technologies de calcul pour exécuter des fonctions cognitives généralement associées aux humains, telles que l'apprentissage, la résolution de problèmes et la reconnaissance de modèles. Pour plus d'informations, veuillez consulter Qu'est-ce que l'intelligence artificielle ?

opérations d'intelligence artificielle (AIOps)

Processus consistant à utiliser des techniques de machine learning pour résoudre les problèmes opérationnels, réduire les incidents opérationnels et les interventions humaines, mais aussi améliorer la qualité du service. Pour plus d'informations sur la façon dont les AIOps sont utilisées dans la stratégie de migration AWS , veuillez consulter le guide d'intégration des opérations.

chiffrement asymétrique

Algorithme de chiffrement qui utilise une paire de clés, une clé publique pour le chiffrement et une clé privée pour le déchiffrement. Vous pouvez partager la clé publique, car elle n'est pas utilisée pour le déchiffrement, mais l'accès à la clé privée doit être très restreint.

atomicité, cohérence, isolement, durabilité (ACID)

Ensemble de propriétés logicielles garantissant la validité des données et la fiabilité opérationnelle d'une base de données, même en cas d'erreur, de panne de courant ou d'autres problèmes.

contrôle d'accès par attributs (ABAC)

Pratique qui consiste à créer des autorisations détaillées en fonction des attributs de l'utilisateur, tels que le service, le poste et le nom de l'équipe. Pour plus d'informations, consultez ABAC pour AWS dans la documentation AWS Identity and Access Management (IAM).

source de données faisant autorité

Emplacement où vous stockez la version principale des données, considérée comme la source d'information la plus fiable. Vous pouvez copier les données de la source de données officielle vers d'autres emplacements à des fins de traitement ou de modification des données, par exemple en les anonymisant, en les expurgant ou en les pseudonymisant.

Zone de disponibilité

Un emplacement distinct au sein d'un Région AWS réseau isolé des défaillances dans d'autres zones de disponibilité et fournissant une connectivité réseau peu coûteuse et à faible latence aux autres zones de disponibilité de la même région.

AWS Cadre d'adoption du cloud (AWS CAF)

Un cadre de directives et de meilleures pratiques visant AWS à aider les entreprises à élaborer un plan efficace pour réussir leur migration vers le cloud. AWS La CAF organise ses conseils en six domaines prioritaires appelés perspectives : les affaires, les personnes, la gouvernance, les plateformes, la sécurité et les opérations. Les perspectives d'entreprise, de personnes et de gouvernance mettent l'accent sur les compétences et les processus métier, tandis que les perspectives relatives à la plateforme, à la sécurité et aux opérations se concentrent sur les compétences et les processus techniques. Par exemple, la perspective liée aux personnes cible les parties prenantes qui s'occupent des ressources humaines (RH), des fonctions de dotation en personnel et de la gestion des personnes. Dans cette perspective, la AWS CAF fournit des conseils pour le développement du personnel, la formation et les communications afin de préparer l'organisation à une adoption réussie du cloud. Pour plus d'informations, veuillez consulter le site Web AWS CAF et le livre blanc AWS CAF.

AWS Cadre de qualification de la charge de travail (AWS WQF)

Outil qui évalue les charges de travail liées à la migration des bases de données, recommande des stratégies de migration et fournit des estimations de travail. AWS Le WQF est inclus avec

AWS Schema Conversion Tool (AWS SCT). Il analyse les schémas de base de données et les objets de code, le code d'application, les dépendances et les caractéristiques de performance, et fournit des rapports d'évaluation.

B

mauvais bot

Un bot destiné à perturber ou à nuire à des individus ou à des organisations.

BCP

Consultez la section Planification de la continuité des activités.

graphique de comportement

Vue unifiée et interactive des comportements des ressources et des interactions au fil du temps. Vous pouvez utiliser un graphique de comportement avec Amazon Detective pour examiner les tentatives de connexion infructueuses, les appels d'API suspects et les actions similaires. Pour plus d'informations, veuillez consulter Data in a behavior graph dans la documentation Detective.

système de poids fort

Système qui stocke d'abord l'octet le plus significatif. Voir aussi endianité.

classification binaire

Processus qui prédit un résultat binaire (l'une des deux classes possibles). Par exemple, votre modèle de machine learning peut avoir besoin de prévoir des problèmes tels que « Cet e-mail est-il du spam ou non ? » ou « Ce produit est-il un livre ou une voiture ? ».

filtre de Bloom

Structure de données probabiliste et efficace en termes de mémoire qui est utilisée pour tester si un élément fait partie d'un ensemble.

déploiement bleu/vert

Stratégie de déploiement dans laquelle vous créez deux environnements distincts mais identiques. Vous exécutez la version actuelle de l'application dans un environnement (bleu) et la nouvelle version de l'application dans l'autre environnement (vert). Cette stratégie vous permet de revenir rapidement en arrière avec un impact minimal.

bot

Application logicielle qui exécute des tâches automatisées sur Internet et simule l'activité ou l'interaction humaine. Certains robots sont utiles ou bénéfiques, comme les robots d'exploration Web qui indexent des informations sur Internet. D'autres robots, connus sous le nom de mauvais robots, sont destinés à perturber ou à nuire à des individus ou à des organisations.

botnet

Réseaux de robots infectés par des logiciels malveillants et contrôlés par une seule entité, connue sous le nom d'herder ou d'opérateur de bots. Les botnets sont le mécanisme le plus connu pour faire évoluer les bots et leur impact.

branche

Zone contenue d'un référentiel de code. La première branche créée dans un référentiel est la branche principale. Vous pouvez créer une branche à partir d'une branche existante, puis développer des fonctionnalités ou corriger des bogues dans la nouvelle branche. Une branche que vous créez pour générer une fonctionnalité est communément appelée branche de fonctionnalités. Lorsque la fonctionnalité est prête à être publiée, vous fusionnez à nouveau la branche de fonctionnalités dans la branche principale. Pour plus d'informations, consultez À propos des branches (GitHub documentation).

accès par brise-vitre

Dans des circonstances exceptionnelles et par le biais d'un processus approuvé, c'est un moyen rapide pour un utilisateur d'accéder à un accès auquel Compte AWS il n'est généralement pas autorisé. Pour plus d'informations, consultez l'indicateur Implementation break-glass procedures dans le guide Well-Architected AWS .

stratégie existante (brownfield)

L'infrastructure existante de votre environnement. Lorsque vous adoptez une stratégie existante pour une architecture système, vous concevez l'architecture en fonction des contraintes des systèmes et de l'infrastructure actuels. Si vous étendez l'infrastructure existante, vous pouvez combiner des politiques brownfield (existantes) et greenfield (inédites).

cache de tampon

Zone de mémoire dans laquelle sont stockées les données les plus fréquemment consultées.

capacité métier

Ce que fait une entreprise pour générer de la valeur (par exemple, les ventes, le service client ou le marketing). Les architectures de microservices et les décisions de développement

peuvent être dictées par les capacités métier. Pour plus d'informations, veuillez consulter la section Organisation en fonction des capacités métier du livre blanc Exécution de microservices conteneurisés sur AWS.

planification de la continuité des activités (BCP)

Plan qui tient compte de l'impact potentiel d'un événement perturbateur, tel qu'une migration à grande échelle, sur les opérations, et qui permet à une entreprise de reprendre ses activités rapidement.

C

CAF

Voir le cadre d'adoption du AWS cloud.

déploiement de Canary

Diffusion lente et progressive d'une version pour les utilisateurs finaux. Lorsque vous êtes sûr, vous déployez la nouvelle version et remplacez la version actuelle dans son intégralité.

CCoE

Voir le Centre d'excellence du cloud.

CDC

Consultez la section Capture des données de modification.

capture des données de modification (CDC)

Processus de suivi des modifications apportées à une source de données, telle qu'une table de base de données, et d'enregistrement des métadonnées relatives à ces modifications. Vous pouvez utiliser la CDC à diverses fins, telles que l'audit ou la réplication des modifications dans un système cible afin de maintenir la synchronisation.

ingénierie du chaos

Introduire intentionnellement des défaillances ou des événements perturbateurs pour tester la résilience d'un système. Vous pouvez utiliser AWS Fault Injection Service (AWS FIS) pour effectuer des expériences qui stressent vos AWS charges de travail et évaluer leur réponse.

CI/CD

Découvrez l'intégration continue et la livraison continue.

classification

Processus de catégorisation qui permet de générer des prédictions. Les modèles de ML pour les problèmes de classification prédisent une valeur discrète. Les valeurs discrètes se distinguent toujours les unes des autres. Par exemple, un modèle peut avoir besoin d'évaluer la présence ou non d'une voiture sur une image.

chiffrement côté client

Chiffrement des données localement, avant que la cible ne les Service AWS reçoive.

Centre d'excellence cloud (CCoE)

Une équipe multidisciplinaire qui dirige les efforts d'adoption du cloud au sein d'une organisation, notamment en développant les bonnes pratiques en matière de cloud, en mobilisant des ressources, en établissant des délais de migration et en guidant l'organisation dans le cadre de transformations à grande échelle. Pour plus d'informations, consultez les articles du CCoE sur le blog de stratégie AWS Cloud d'entreprise.

cloud computing

Technologie cloud généralement utilisée pour le stockage de données à distance et la gestion des appareils IoT. Le cloud computing est généralement associé à la technologie informatique de pointe.

modèle d'exploitation du cloud

Dans une organisation informatique, modèle d'exploitation utilisé pour créer, faire évoluer et optimiser un ou plusieurs environnements cloud. Pour plus d'informations, consultez la section Création de votre modèle d'exploitation cloud.

étapes d'adoption du cloud

Les quatre phases que les entreprises traversent généralement lorsqu'elles migrent vers AWS Cloud :

- Projet : exécution de quelques projets liés au cloud à des fins de preuve de concept et d'apprentissage
- Base : réaliser des investissements fondamentaux pour mettre à l'échelle l'adoption du cloud (par exemple, en créant une zone de destination, en définissant un CCoE ou en établissant un modèle opérationnel)
- Migration : migration d'applications individuelles

- Réinvention : optimisation des produits et services et innovation dans le cloud

Ces étapes ont été définies par Stephen Orban dans le billet de blog The Journey Toward Cloud-First & the Stages of Adoption publié sur le blog AWS Cloud Enterprise Strategy. Pour plus d'informations sur leur lien avec la stratégie de AWS migration, consultez le guide de préparation à la migration.

CMDB

Voir base de données de gestion de configuration.

référentiel de code

Emplacement où le code source et d'autres ressources, comme la documentation, les exemples et les scripts, sont stockés et mis à jour par le biais de processus de contrôle de version. Les référentiels cloud courants incluent GitHub ou AWS CodeCommit. Chaque version du code est appelée branche. Dans une structure de microservice, chaque référentiel est consacré à une seule fonctionnalité. Un seul pipeline CI/CD peut utiliser plusieurs référentiels.

cache passif

Cache tampon vide, mal rempli ou contenant des données obsolètes ou non pertinentes. Cela affecte les performances, car l'instance de base de données doit lire à partir de la mémoire principale ou du disque, ce qui est plus lent que la lecture à partir du cache tampon.

données gelées

Données rarement consultées et généralement historiques. Lorsque vous interrogez ce type de données, les requêtes lentes sont généralement acceptables. Le transfert de ces données vers des niveaux ou classes de stockage moins performants et moins coûteux peut réduire les coûts.

vision par ordinateur (CV)

Domaine de l'IA qui utilise l'apprentissage automatique pour analyser et extraire des informations à partir de formats visuels tels que des images numériques et des vidéos. Par exemple, AWS Panorama propose des appareils qui ajoutent des CV aux réseaux de caméras locaux, et Amazon SageMaker fournit des algorithmes de traitement d'image pour les CV.

dérive de configuration

Pour une charge de travail, une modification de configuration par rapport à l'état attendu. Cela peut entraîner une non-conformité de la charge de travail, et cela est généralement progressif et involontaire.

base de données de gestion des configurations (CMDB)

Référentiel qui stocke et gère les informations relatives à une base de données et à son environnement informatique, y compris les composants matériels et logiciels ainsi que leurs configurations. Vous utilisez généralement les données d'une CMDB lors de la phase de découverte et d'analyse du portefeuille de la migration.

pack de conformité

Ensemble de AWS Config règles et d'actions correctives que vous pouvez assembler pour personnaliser vos contrôles de conformité et de sécurité. Vous pouvez déployer un pack de conformité en tant qu'entité unique dans une région Compte AWS et, ou au sein d'une organisation, à l'aide d'un modèle YAML. Pour plus d'informations, consultez la section Packs de conformité dans la AWS Config documentation.

intégration continue et livraison continue (CI/CD)

Processus d'automatisation des étapes source, de génération, de test, intermédiaire et de production du processus de publication du logiciel. CI/CD est communément décrit comme un pipeline. CI/CD peut vous aider à automatiser les processus, à améliorer la productivité, à améliorer la qualité du code et à accélérer les livraisons. Pour plus d'informations, veuillez consulter Avantages de la livraison continue. CD peut également signifier déploiement continu. Pour plus d'informations, veuillez consulter Livraison continue et déploiement continu.

CV

Voir vision par ordinateur.

D

données au repos

Données stationnaires dans votre réseau, telles que les données stockées.

classification des données

Processus permettant d'identifier et de catégoriser les données de votre réseau en fonction de leur sévérité et de leur sensibilité. Il s'agit d'un élément essentiel de toute stratégie de gestion des risques de cybersécurité, car il vous aide à déterminer les contrôles de protection et de conservation appropriés pour les données. La classification des données est une composante du pilier de sécurité du AWS Well-Architected Framework. Pour plus d'informations, veuillez consulter Classification des données.

dérive des données

Une variation significative entre les données de production et les données utilisées pour entraîner un modèle ML, ou une modification significative des données d'entrée au fil du temps. La dérive des données peut réduire la qualité, la précision et l'équité globales des prédictions des modèles ML.

données en transit

Données qui circulent activement sur votre réseau, par exemple entre les ressources du réseau.

maillage de données

Un cadre architectural qui fournit une propriété des données distribuée et décentralisée avec une gestion et une gouvernance centralisées.

minimisation des données

Le principe de collecte et de traitement des seules données strictement nécessaires. La pratique de la minimisation des données AWS Cloud peut réduire les risques liés à la confidentialité, les coûts et l'empreinte carbone de vos analyses.

périmètre de données

Ensemble de garde-fous préventifs dans votre AWS environnement qui permettent de garantir que seules les identités fiables accèdent aux ressources fiables des réseaux attendus. Pour plus d'informations, voir Création d'un périmètre de données sur AWS.

prétraitement des données

Pour transformer les données brutes en un format facile à analyser par votre modèle de ML. Le prétraitement des données peut impliquer la suppression de certaines colonnes ou lignes et le traitement des valeurs manquantes, incohérentes ou en double.

provenance des données

Le processus de suivi de l'origine et de l'historique des données tout au long de leur cycle de vie, par exemple la manière dont les données ont été générées, transmises et stockées.

sujet des données

Personne dont les données sont collectées et traitées.

entrepôt des données

Un système de gestion des données qui prend en charge les informations commerciales, telles que les analyses. Les entrepôts de données contiennent généralement de grandes quantités de données historiques et sont généralement utilisés pour les requêtes et les analyses.

langage de définition de base de données (DDL)

Instructions ou commandes permettant de créer ou de modifier la structure des tables et des objets dans une base de données.

langage de manipulation de base de données (DML)

Instructions ou commandes permettant de modifier (insérer, mettre à jour et supprimer) des informations dans une base de données.

DDL

Voir langage de définition de base de données.

ensemble profond

Sert à combiner plusieurs modèles de deep learning à des fins de prédiction. Vous pouvez utiliser des ensembles profonds pour obtenir une prévision plus précise ou pour estimer l'incertitude des prédictions.

deep learning

Un sous-champ de ML qui utilise plusieurs couches de réseaux neuronaux artificiels pour identifier le mappage entre les données d'entrée et les variables cibles d'intérêt.

defense-in-depth

Approche de la sécurité de l'information dans laquelle une série de mécanismes et de contrôles de sécurité sont judicieusement répartis sur l'ensemble d'un réseau informatique afin de protéger la confidentialité, l'intégrité et la disponibilité du réseau et des données qu'il contient. Lorsque vous adoptez cette stratégie AWS, vous ajoutez plusieurs contrôles à différentes couches de la AWS Organizations structure afin de sécuriser les ressources. Par exemple, une defense-in-depth approche peut combiner l'authentification multifactorielle, la segmentation du réseau et le chiffrement.

administrateur délégué

Dans AWS Organizations, un service compatible peut enregistrer un compte AWS membre pour administrer les comptes de l'organisation et gérer les autorisations pour ce service. Ce compte est

appelé administrateur délégué pour ce service. Pour plus d'informations et une liste des services compatibles, veuillez consulter la rubrique Services qui fonctionnent avec AWS Organizations dans la documentation AWS Organizations .

déploiement

Processus de mise à disposition d'une application, de nouvelles fonctionnalités ou de corrections de code dans l'environnement cible. Le déploiement implique la mise en œuvre de modifications dans une base de code, puis la génération et l'exécution de cette base de code dans les environnements de l'application.

environnement de développement

Voir environnement.

contrôle de détection

Contrôle de sécurité conçu pour détecter, journaliser et alerter après la survenue d'un événement. Ces contrôles constituent une deuxième ligne de défense et vous alertent en cas d'événements de sécurité qui ont contourné les contrôles préventifs en place. Pour plus d'informations, veuillez consulter la rubrique Contrôles de détection dans Implementing security controls on AWS.

cartographie de la chaîne de valeur du développement (DVSM)

Processus utilisé pour identifier et hiérarchiser les contraintes qui nuisent à la rapidité et à la qualité du cycle de vie du développement logiciel. DVSM étend le processus de cartographie de la chaîne de valeur initialement conçu pour les pratiques de production allégée. Il met l'accent sur les étapes et les équipes nécessaires pour créer et transférer de la valeur tout au long du processus de développement logiciel.

jumeau numérique

Représentation virtuelle d'un système réel, tel qu'un bâtiment, une usine, un équipement industriel ou une ligne de production. Les jumeaux numériques prennent en charge la maintenance prédictive, la surveillance à distance et l'optimisation de la production.

tableau des dimensions

Dans un schéma en étoile, table plus petite contenant les attributs de données relatifs aux données quantitatives d'une table de faits. Les attributs des tables de dimensions sont généralement des champs de texte ou des nombres discrets qui se comportent comme du texte. Ces attributs sont couramment utilisés pour la contrainte des requêtes, le filtrage et l'étiquetage des ensembles de résultats.

catastrophe

Un événement qui empêche une charge de travail ou un système d'atteindre ses objectifs commerciaux sur son site de déploiement principal. Ces événements peuvent être des catastrophes naturelles, des défaillances techniques ou le résultat d'actions humaines, telles qu'une mauvaise configuration involontaire ou une attaque de logiciel malveillant.

reprise après sinistre (DR)

La stratégie et le processus que vous utilisez pour minimiser les temps d'arrêt et les pertes de données causés par un sinistre. Pour plus d'informations, consultez Disaster Recovery of Workloads on AWS : Recovery in the Cloud in the AWS Well-Architected Framework.

DML

Voir langage de manipulation de base de données.

conception axée sur le domaine

Approche visant à développer un système logiciel complexe en connectant ses composants à des domaines évolutifs, ou objectifs métier essentiels, que sert chaque composant. Ce concept a été introduit par Eric Evans dans son ouvrage Domain-Driven Design: Tackling Complexity in the Heart of Software (Boston : Addison-Wesley Professional, 2003). Pour plus d'informations sur l'utilisation du design piloté par domaine avec le modèle de figuier étrangleur, veuillez consulter Modernizing legacy Microsoft ASP.NET (ASMX) web services incrementally by using containers and Amazon API Gateway.

DR

Consultez la section Reprise après sinistre.

détection de dérive

Suivi des écarts par rapport à une configuration de référence. Par exemple, vous pouvez l'utiliser AWS CloudFormation pour détecter la dérive des ressources du système ou AWS Control Tower pour détecter les modifications de votre zone d'atterrissage susceptibles d'affecter le respect des exigences de gouvernance.

DVSM

Voir la cartographie de la chaîne de valeur du développement.

E

EDA

Voir analyse exploratoire des données.

informatique de périphérie

Technologie qui augmente la puissance de calcul des appareils intelligents en périphérie d'un réseau IoT. Comparé au cloud computing, l'informatique de pointe peut réduire la latence des communications et améliorer le temps de réponse.

chiffrement

Processus informatique qui transforme des données en texte clair, lisibles par l'homme, en texte chiffré.

clé de chiffrement

Chaîne cryptographique de bits aléatoires générée par un algorithme cryptographique. La longueur des clés peut varier, et chaque clé est conçue pour être imprévisible et unique.

endianisme

Ordre selon lequel les octets sont stockés dans la mémoire de l'ordinateur. Les systèmes de poids fort stockent d'abord l'octet le plus significatif. Les systèmes de poids faible stockent d'abord l'octet le moins significatif.

point de terminaison

Voir point de terminaison de service.

service de point de terminaison

Service que vous pouvez héberger sur un cloud privé virtuel (VPC) pour le partager avec d'autres utilisateurs. Vous pouvez créer un service de point de terminaison avec AWS PrivateLink et accorder des autorisations à d'autres principaux Comptes AWS ou à AWS Identity and Access Management (IAM) principaux. Ces comptes ou principaux peuvent se connecter à votre service de point de terminaison de manière privée en créant des points de terminaison d'un VPC d'interface. Pour plus d'informations, veuillez consulter Création d'un service de point de terminaison dans la documentation Amazon Virtual Private Cloud (Amazon VPC).

planification des ressources d'entreprise (ERP)

Système qui automatise et gère les principaux processus métier (tels que la comptabilité, le MES et la gestion de projet) pour une entreprise.

chiffrement d'enveloppe

Processus de chiffrement d'une clé de chiffrement à l'aide d'une autre clé de chiffrement. Pour plus d'informations, consultez la section Chiffrement des enveloppes dans la documentation AWS Key Management Service (AWS KMS).

environment

Instance d'une application en cours d'exécution. Les types d'environnement les plus courants dans le cloud computing sont les suivants :

- Environnement de développement : instance d'une application en cours d'exécution à laquelle seule l'équipe principale chargée de la maintenance de l'application peut accéder. Les environnements de développement sont utilisés pour tester les modifications avant de les promouvoir dans les environnements supérieurs. Ce type d'environnement est parfois appelé environnement de test.
- Environnements inférieurs : tous les environnements de développement d'une application, tels que ceux utilisés pour les générations et les tests initiaux.
- Environnement de production : instance d'une application en cours d'exécution à laquelle les utilisateurs finaux peuvent accéder. Dans un pipeline CI/CD, l'environnement de production est le dernier environnement de déploiement.
- Environnements supérieurs : tous les environnements accessibles aux utilisateurs autres que l'équipe de développement principale. Ils peuvent inclure un environnement de production, des environnements de préproduction et des environnements pour les tests d'acceptation par les utilisateurs.

épopée

Dans les méthodologies agiles, catégories fonctionnelles qui aident à organiser et à prioriser votre travail. Les épopées fournissent une description détaillée des exigences et des tâches d'implémentation. Par exemple, les points forts de la AWS CAF en matière de sécurité incluent la gestion des identités et des accès, les contrôles de détection, la sécurité des infrastructures, la protection des données et la réponse aux incidents. Pour plus d'informations sur les épopées dans la stratégie de migration AWS , veuillez consulter le guide d'implémentation du programme.

ERP

Voir Planification des ressources d'entreprise.

analyse exploratoire des données (EDA)

Processus d'analyse d'un jeu de données pour comprendre ses principales caractéristiques. Vous collectez ou agrégez des données, puis vous effectuez des enquêtes initiales pour trouver des modèles, détecter des anomalies et vérifier les hypothèses. L'EDA est réalisée en calculant des statistiques récapitulatives et en créant des visualisations de données.

F

tableau des faits

La table centrale dans un schéma en étoile. Il stocke des données quantitatives sur les opérations commerciales. Généralement, une table de faits contient deux types de colonnes : celles qui contiennent des mesures et celles qui contiennent une clé étrangère pour une table de dimensions.

échouer rapidement

Une philosophie qui utilise des tests fréquents et progressifs pour réduire le cycle de vie du développement. C'est un élément essentiel d'une approche agile.

limite d'isolation des défauts

Dans le AWS Cloud, une limite telle qu'une zone de disponibilité Région AWS, un plan de contrôle ou un plan de données qui limite l'effet d'une panne et contribue à améliorer la résilience des charges de travail. Pour plus d'informations, consultez la section Limites d'isolation des AWS pannes.

branche de fonctionnalités

Voir la succursale.

fonctionnalités

Les données d'entrée que vous utilisez pour faire une prédiction. Par exemple, dans un contexte de fabrication, les fonctionnalités peuvent être des images capturées périodiquement à partir de la ligne de fabrication.

importance des fonctionnalités

Le niveau d'importance d'une fonctionnalité pour les prédictions d'un modèle. Il s'exprime généralement sous la forme d'un score numérique qui peut être calculé à l'aide de différentes

techniques, telles que la méthode Shapley Additive Explanations (SHAP) et les gradients intégrés. Pour plus d'informations, voir Interprétabilité du modèle d'apprentissage automatique avec :AWS.

transformation de fonctionnalité

Optimiser les données pour le processus de ML, notamment en enrichissant les données avec des sources supplémentaires, en mettant à l'échelle les valeurs ou en extrayant plusieurs ensembles d'informations à partir d'un seul champ de données. Cela permet au modèle de ML de tirer parti des données. Par exemple, si vous décomposez la date « 2021-05-27 00:15:37 » en « 2021 », « mai », « jeudi » et « 15 », vous pouvez aider l'algorithme d'apprentissage à apprendre des modèles nuancés associés à différents composants de données.

FGAC

Découvrez le contrôle d'accès détaillé.

contrôle d'accès détaillé (FGAC)

Utilisation de plusieurs conditions pour autoriser ou refuser une demande d'accès.

migration instantanée (flash-cut)

Méthode de migration de base de données qui utilise la réplication continue des données via la capture des données de modification afin de migrer les données dans les plus brefs délais, au lieu d'utiliser une approche progressive. L'objectif est de réduire au maximum les temps d'arrêt.

G

blocage géographique

Voir les restrictions géographiques.

restrictions géographiques (blocage géographique)

Sur Amazon CloudFront, option permettant d'empêcher les utilisateurs de certains pays d'accéder aux distributions de contenu. Vous pouvez utiliser une liste d'autorisation ou une liste de blocage pour spécifier les pays approuvés et interdits. Pour plus d'informations, consultez la section Restreindre la distribution géographique de votre contenu dans la CloudFront documentation.

Flux de travail Gitflow

Approche dans laquelle les environnements inférieurs et supérieurs utilisent différentes branches dans un référentiel de code source. Le flux de travail Gitflow est considéré comme existant, et le flux de travail basé sur les troncs est l'approche moderne préférée.

stratégie inédite

L'absence d'infrastructures existantes dans un nouvel environnement. Lorsque vous adoptez une stratégie inédite pour une architecture système, vous pouvez sélectionner toutes les nouvelles technologies sans restriction de compatibilité avec l'infrastructure existante, également appelée brownfield. Si vous étendez l'infrastructure existante, vous pouvez combiner des politiques brownfield (existantes) et greenfield (inédites).

barrière de protection

Règle de haut niveau qui permet de régir les ressources, les politiques et la conformité au sein des unités d'organisation (UO). Les barrières de protection préventives appliquent des politiques pour garantir l'alignement sur les normes de conformité. Elles sont mises en œuvre à l'aide de politiques de contrôle des services et de limites des autorisations IAM. Les barrières de protection de détection détectent les violations des politiques et les problèmes de conformité, et génèrent des alertes pour y remédier. Ils sont implémentés à l'aide d'Amazon AWS Config AWS Security Hub GuardDuty AWS Trusted Advisor, d'Amazon Inspector et de AWS Lambda contrôles personnalisés.

H

HA

Découvrez la haute disponibilité.

migration de base de données hétérogène

Migration de votre base de données source vers une base de données cible qui utilise un moteur de base de données différent (par exemple, Oracle vers Amazon Aurora). La migration hétérogène fait généralement partie d'un effort de réarchitecture, et la conversion du schéma peut s'avérer une tâche complexe. AWS propose AWS SCT qui facilite les conversions de schémas.

haute disponibilité (HA)

Capacité d'une charge de travail à fonctionner en continu, sans intervention, en cas de difficultés ou de catastrophes. Les systèmes HA sont conçus pour basculer automatiquement, fournir constamment des performances de haute qualité et gérer différentes charges et défaillances avec un impact minimal sur les performances.

modernisation de l'historien

Approche utilisée pour moderniser et mettre à niveau les systèmes de technologie opérationnelle (OT) afin de mieux répondre aux besoins de l'industrie manufacturière. Un historien est un type de base de données utilisé pour collecter et stocker des données provenant de diverses sources dans une usine.

migration de base de données homogène

Migration de votre base de données source vers une base de données cible qui partage le même moteur de base de données (par exemple, Microsoft SQL Server vers Amazon RDS for SQL Server). La migration homogène s'inscrit généralement dans le cadre d'un effort de réhébergement ou de replateforme. Vous pouvez utiliser les utilitaires de base de données natifs pour migrer le schéma.

données chaudes

Données fréquemment consultées, telles que les données en temps réel ou les données translationnelles récentes. Ces données nécessitent généralement un niveau ou une classe de stockage à hautes performances pour fournir des réponses rapides aux requêtes.

correctif

Solution d'urgence à un problème critique dans un environnement de production. En raison de son urgence, un correctif est généralement créé en dehors du flux de travail de DevOps publication habituel.

période de soins intensifs

Immédiatement après le basculement, période pendant laquelle une équipe de migration gère et surveille les applications migrées dans le cloud afin de résoudre les problèmes éventuels. En règle générale, cette période dure de 1 à 4 jours. À la fin de la période de soins intensifs, l'équipe de migration transfère généralement la responsabilité des applications à l'équipe des opérations cloud.

I

IaC

Considérez l'infrastructure comme un code.

politique basée sur l’identité

Politique attachée à un ou plusieurs principaux IAM qui définit leurs autorisations au sein de l' AWS Cloud environnement.

application inactive

Application dont l'utilisation moyenne du processeur et de la mémoire se situe entre 5 et 20 % sur une période de 90 jours. Dans un projet de migration, il est courant de retirer ces applications ou de les retenir sur site.

IIoT

Voir Internet industriel des objets.

infrastructure immuable

Modèle qui déploie une nouvelle infrastructure pour les charges de travail de production au lieu de mettre à jour, d'appliquer des correctifs ou de modifier l'infrastructure existante. Les infrastructures immuables sont intrinsèquement plus cohérentes, fiables et prévisibles que les infrastructures mutables. Pour plus d'informations, consultez les meilleures pratiques de déploiement à l'aide d'une infrastructure immuable dans le AWS Well-Architected Framework.

VPC entrant (d'entrée)

Dans une architecture AWS multi-comptes, un VPC qui accepte, inspecte et achemine les connexions réseau depuis l'extérieur d'une application. L'architecture de référence de sécuritéAWS recommande de configurer votre compte réseau avec des VPC entrants, sortants et d'inspection afin de protéger l'interface bidirectionnelle entre votre application et Internet en général.

migration incrémentielle

Stratégie de basculement dans le cadre de laquelle vous migrez votre application par petites parties au lieu d'effectuer un basculement complet unique. Par exemple, il se peut que vous ne transfériez que quelques microservices ou utilisateurs vers le nouveau système dans un premier temps. Après avoir vérifié que tout fonctionne correctement, vous pouvez transférer progressivement des microservices ou des utilisateurs supplémentaires jusqu'à ce que vous puissiez mettre hors service votre système hérité. Cette stratégie réduit les risques associés aux migrations de grande ampleur.

Industry 4.0

Terme introduit par Klaus Schwab en 2016 pour désigner la modernisation des processus de fabrication grâce aux avancées en matière de connectivité, de données en temps réel, d'automatisation, d'analyse et d'IA/ML.

infrastructure

Ensemble des ressources et des actifs contenus dans l'environnement d'une application.

infrastructure en tant que code (IaC)

Processus de mise en service et de gestion de l'infrastructure d'une application via un ensemble de fichiers de configuration. IaC est conçue pour vous aider à centraliser la gestion de l'infrastructure, à normaliser les ressources et à mettre à l'échelle rapidement afin que les nouveaux environnements soient reproductibles, fiables et cohérents.

internet industriel des objets (IIoT)

L'utilisation de capteurs et d'appareils connectés à Internet dans les secteurs industriels tels que la fabrication, l'énergie, l'automobile, les soins de santé, les sciences de la vie et l'agriculture. Pour plus d'informations, veuillez consulter Building an industrial Internet of Things (IIoT) digital transformation strategy.

VPC d'inspection

Dans une architecture AWS multi-comptes, un VPC centralisé qui gère les inspections du trafic réseau entre les VPC (identiques ou Régions AWS différents), Internet et les réseaux sur site. L'architecture de référence de sécuritéAWS recommande de configurer votre compte réseau avec des VPC entrants, sortants et d'inspection afin de protéger l'interface bidirectionnelle entre votre application et Internet en général.

Internet des objets (IoT)

Réseau d'objets physiques connectés dotés de capteurs ou de processeurs intégrés qui communiquent avec d'autres appareils et systèmes via Internet ou via un réseau de communication local. Pour plus d'informations, veuillez consulter la section Qu'est-ce que l'IoT ?.

interprétabilité

Caractéristique d'un modèle de machine learning qui décrit dans quelle mesure un être humain peut comprendre comment les prédictions du modèle dépendent de ses entrées. Pour plus d'informations, veuillez consulter Machine learning model interpretability with AWS.

IoT

Voir Internet des objets.

Bibliothèque d'informations informatiques (ITIL)

Ensemble de bonnes pratiques pour proposer des services informatiques et les aligner sur les exigences métier. L'ITIL constitue la base de l'ITSM.

gestion des services informatiques (ITSM)

Activités associées à la conception, à la mise en œuvre, à la gestion et à la prise en charge de services informatiques d'une organisation. Pour plus d'informations sur l'intégration des opérations cloud aux outils ITSM, veuillez consulter le guide d'intégration des opérations.

ITIL

Consultez la bibliothèque d'informations informatiques.

ITSM

Consultez la section Gestion des services informatiques.

L

contrôle d'accès basé sur des étiquettes (LBAC)

Une implémentation du contrôle d'accès obligatoire (MAC) dans laquelle une valeur d'étiquette de sécurité est explicitement attribuée aux utilisateurs et aux données elles-mêmes. L'intersection entre l'étiquette de sécurité utilisateur et l'étiquette de sécurité des données détermine les lignes et les colonnes visibles par l'utilisateur.

zone de destination

Une zone d'atterrissage est un AWS environnement multi-comptes bien conçu, évolutif et sécurisé. Il s'agit d'un point de départ à partir duquel vos entreprises peuvent rapidement lancer et déployer des charges de travail et des applications en toute confiance dans leur environnement de sécurité et d'infrastructure. Pour plus d'informations sur les zones de destination, veuillez consulter Setting up a secure and scalable multi-account AWS environment.

migration de grande envergure

Migration de 300 serveurs ou plus.

LBAC

Voir contrôle d'accès basé sur des étiquettes.

principe de moindre privilège

Bonne pratique de sécurité qui consiste à accorder les autorisations minimales nécessaires à l'exécution d'une tâche. Pour plus d'informations, veuillez consulter la rubrique Accorder les autorisations de moindre privilège dans la documentation IAM.

lift and shift

Voir 7 Rs.

système de poids faible

Système qui stocke d'abord l'octet le moins significatif. Voir aussi endianité.

environnements inférieurs

Voir environnement.

M

machine learning (ML)

Type d'intelligence artificielle qui utilise des algorithmes et des techniques pour la reconnaissance et l'apprentissage de modèles. Le ML analyse et apprend à partir de données enregistrées, telles que les données de l'Internet des objets (IoT), pour générer un modèle statistique basé sur des modèles. Pour plus d'informations, veuillez consulter Machine Learning.

branche principale

Voir la succursale.

malware

Logiciel conçu pour compromettre la sécurité ou la confidentialité de l'ordinateur. Les logiciels malveillants peuvent perturber les systèmes informatiques, divulguer des informations sensibles ou obtenir un accès non autorisé. Parmi les malwares, on peut citer les virus, les vers, les rançongiciels, les chevaux de Troie, les logiciels espions et les enregistreurs de frappe.

services gérés

Services AWS qui AWS gère la couche d'infrastructure, le système d'exploitation et les plateformes, et vous accédez aux points de terminaison pour stocker et récupérer des données.

Amazon Simple Storage Service (Amazon S3) et Amazon DynamoDB sont des exemples de services gérés. Ils sont également connus sous le nom de services abstraits.

système d'exécution de la fabrication (MES)

Un système logiciel pour le suivi, la surveillance, la documentation et le contrôle des processus de production qui convertissent les matières premières en produits finis dans l'atelier.

MAP

Voir Migration Acceleration Program.

mécanisme

Processus complet au cours duquel vous créez un outil, favorisez son adoption, puis inspectez les résultats afin de procéder aux ajustements nécessaires. Un mécanisme est un cycle qui se renforce et s'améliore lorsqu'il fonctionne. Pour plus d'informations, voir Création de mécanismes dans le cadre AWS Well-Architected.

compte membre

Tous, à l' Comptes AWS exception du compte de gestion, qui font partie d'une organisation dans AWS Organizations. Un compte ne peut être membre que d'une seule organisation à la fois.

MAILLES

Voir le système d'exécution de la fabrication.

Transport télémétrique en file d'attente de messages (MQTT)

Protocole de communication léger machine-to-machine (M2M), basé sur le modèle de publication/d'abonnement, pour les appareils IoT aux ressources limitées.

microservice

Petit service indépendant qui communique via des API bien définies et qui est généralement détenu par de petites équipes autonomes. Par exemple, un système d'assurance peut inclure des microservices qui mappent à des capacités métier, telles que les ventes ou le marketing, ou à des sous-domaines, tels que les achats, les réclamations ou l'analytique. Les avantages des microservices incluent l'agilité, la flexibilité de la mise à l'échelle, la facilité de déploiement, la réutilisation du code et la résilience. Pour plus d'informations, consultez la section Intégration de microservices à l'aide de services AWS sans serveur.

architecture de microservices

Approche de création d'une application avec des composants indépendants qui exécutent chaque processus d'application en tant que microservice. Ces microservices communiquent via une

interface bien définie à l'aide d'API légères. Chaque microservice de cette architecture peut être mis à jour, déployé et mis à l'échelle pour répondre à la demande de fonctions spécifiques d'une application. Pour plus d'informations, consultez la section Implémentation de microservices sur AWS.

Programme d'accélération des migrations (MAP)

Un AWS programme qui fournit un support de conseil, des formations et des services pour aider les entreprises à établir une base opérationnelle solide pour passer au cloud, et pour aider à compenser le coût initial des migrations. MAP inclut une méthodologie de migration pour exécuter les migrations héritées de manière méthodique, ainsi qu'un ensemble d'outils pour automatiser et accélérer les scénarios de migration courants.

migration à grande échelle

Processus consistant à transférer la majeure partie du portefeuille d'applications vers le cloud par vagues, un plus grand nombre d'applications étant déplacées plus rapidement à chaque vague. Cette phase utilise les bonnes pratiques et les enseignements tirés des phases précédentes pour implémenter une usine de migration d'équipes, d'outils et de processus en vue de rationaliser la migration des charges de travail grâce à l'automatisation et à la livraison agile. Il s'agit de la troisième phase de la stratégie de migration AWS.

usine de migration

Équipes interfonctionnelles qui rationalisent la migration des charges de travail grâce à des approches automatisées et agiles. Les équipes de Migration Factory comprennent généralement les opérations, les analystes commerciaux et les propriétaires, les ingénieurs de migration, les développeurs et les DevOps professionnels travaillant dans le cadre de sprints. Entre 20 et 50 % du portefeuille d'applications d'entreprise est constitué de modèles répétés qui peuvent être optimisés par une approche d'usine. Pour plus d'informations, veuillez consulter la rubrique discussion of migration factories et le guide Cloud Migration Factory dans cet ensemble de contenus.

métadonnées de migration

Informations relatives à l'application et au serveur nécessaires pour finaliser la migration. Chaque modèle de migration nécessite un ensemble de métadonnées de migration différent. Les exemples de métadonnées de migration incluent le sous-réseau cible, le groupe de sécurité et le AWS compte.

modèle de migration

Tâche de migration reproductible qui détaille la stratégie de migration, la destination de la migration et l'application ou le service de migration utilisé. Exemple : réorganisez la migration vers Amazon EC2 AWS avec le service de migration d'applications.

Évaluation du portefeuille de migration (MPA)

Outil en ligne qui fournit des informations pour valider l'analyse de rentabilisation en faveur de la migration vers le. AWS Cloud La MPA propose une évaluation détaillée du portefeuille (dimensionnement approprié des serveurs, tarification, comparaison du coût total de possession, analyse des coûts de migration), ainsi que la planification de la migration (analyse et collecte des données d'applications, regroupement des applications, priorisation des migrations et planification des vagues). L'outil MPA (connexion requise) est disponible gratuitement pour tous les AWS consultants et consultants APN Partner.

Évaluation de la préparation à la migration (MRA)

Processus qui consiste à obtenir des informations sur l'état de préparation d'une organisation au cloud, à identifier les forces et les faiblesses et à élaborer un plan d'action pour combler les lacunes identifiées, à l'aide du AWS CAF. Pour plus d'informations, veuillez consulter le guide de préparation à la migration. La MRA est la première phase de la stratégie de migration AWS.

stratégie de migration

L'approche utilisée pour migrer une charge de travail vers le AWS Cloud. Pour plus d'informations, reportez-vous aux 7 R de ce glossaire et à Mobiliser votre organisation pour accélérer les migrations à grande échelle.

ML

Voir apprentissage automatique.

modernisation

Transformation d'une application obsolète (héritée ou monolithique) et de son infrastructure en un système agile, élastique et hautement disponible dans le cloud afin de réduire les coûts, de gagner en efficacité et de tirer parti des innovations. Pour plus d'informations, consultez la section Stratégie de modernisation des applications dans le AWS Cloud.

évaluation de la préparation à la modernisation

Évaluation qui permet de déterminer si les applications d'une organisation sont prêtes à être modernisées, d'identifier les avantages, les risques et les dépendances, et qui détermine dans quelle mesure l'organisation peut prendre en charge l'état futur de ces applications. Le résultat

de l'évaluation est un plan de l'architecture cible, une feuille de route détaillant les phases de développement et les étapes du processus de modernisation, ainsi qu'un plan d'action pour combler les lacunes identifiées. Pour plus d'informations, consultez la section Évaluation de l'état de préparation à la modernisation des applications dans le AWS Cloud.

applications monolithiques (monolithes)

Applications qui s'exécutent en tant que service unique avec des processus étroitement couplés. Les applications monolithiques ont plusieurs inconvénients. Si une fonctionnalité de l'application connaît un pic de demande, l'architecture entière doit être mise à l'échelle. L'ajout ou l'amélioration des fonctionnalités d'une application monolithique devient également plus complexe lorsque la base de code s'élargit. Pour résoudre ces problèmes, vous pouvez utiliser une architecture de microservices. Pour plus d'informations, veuillez consulter Decomposing monoliths into microservices.

MPA

Voir Évaluation du portefeuille de migration.

MQTT

Voir Message Queuing Telemetry Transport.

classification multi-classes

Processus qui permet de générer des prédictions pour plusieurs classes (prédiction d'un résultat parmi plus de deux). Par exemple, un modèle de ML peut demander « Ce produit est-il un livre, une voiture ou un téléphone ? » ou « Quelle catégorie de produits intéresse le plus ce client ? ».

infrastructure mutable

Modèle qui met à jour et modifie l'infrastructure existante pour les charges de travail de production. Pour améliorer la cohérence, la fiabilité et la prévisibilité, le AWS Well-Architected Framework recommande l'utilisation d'une infrastructure immuable comme meilleure pratique.

O

OAC

Voir Contrôle d'accès à l'origine.

OAI

Voir l'identité d'accès à l'origine.

OCM

Voir gestion du changement organisationnel.

migration hors ligne

Méthode de migration dans laquelle la charge de travail source est supprimée au cours du processus de migration. Cette méthode implique un temps d'arrêt prolongé et est généralement utilisée pour de petites charges de travail non critiques.

OI

Consultez la section Intégration des opérations.

OLA

Voir l'accord au niveau opérationnel.

migration en ligne

Méthode de migration dans laquelle la charge de travail source est copiée sur le système cible sans être mise hors ligne. Les applications connectées à la charge de travail peuvent continuer à fonctionner pendant la migration. Cette méthode implique un temps d'arrêt nul ou minimal et est généralement utilisée pour les charges de travail de production critiques.

OPC-UA

Voir Open Process Communications - Architecture unifiée.

Communications par processus ouvert - Architecture unifiée (OPC-UA)

Un protocole de communication machine-to-machine (M2M) pour l'automatisation industrielle. L'OPC-UA fournit une norme d'interopérabilité avec des schémas de cryptage, d'authentification et d'autorisation des données.

accord au niveau opérationnel (OLA)

Accord qui précise ce que les groupes informatiques fonctionnels s'engagent à fournir les uns aux autres, afin de prendre en charge un contrat de niveau de service (SLA).

examen de l'état de préparation opérationnelle (ORR)

Une liste de questions et de bonnes pratiques associées qui vous aident à comprendre, évaluer, prévenir ou réduire l'ampleur des incidents et des défaillances possibles. Pour plus d'informations, voir Operational Readiness Reviews (ORR) dans le AWS Well-Architected Framework.

technologie opérationnelle (OT)

Systèmes matériels et logiciels qui fonctionnent avec l'environnement physique pour contrôler les opérations, les équipements et les infrastructures industriels. Dans le secteur manufacturier, l'intégration des systèmes OT et des technologies de l'information (IT) est au cœur des transformations de l'industrie 4.0.

intégration des opérations (OI)

Processus de modernisation des opérations dans le cloud, qui implique la planification de la préparation, l'automatisation et l'intégration. Pour en savoir plus, veuillez consulter le guide d'intégration des opérations.

journal de suivi d'organisation

Un parcours créé par AWS CloudTrail qui enregistre tous les événements pour tous les membres Comptes AWS d'une organisation dans AWS Organizations. Ce journal de suivi est créé dans chaque Compte AWS qui fait partie de l'organisation et suit l'activité de chaque compte. Pour plus d'informations, consultez la section Création d'un suivi pour une organisation dans la CloudTrail documentation.

gestion du changement organisationnel (OCM)

Cadre pour gérer les transformations métier majeures et perturbatrices du point de vue des personnes, de la culture et du leadership. L'OCM aide les organisations à se préparer et à effectuer la transition vers de nouveaux systèmes et de nouvelles politiques en accélérant l'adoption des changements, en abordant les problèmes de transition et en favorisant des changements culturels et organisationnels. Dans la stratégie de AWS migration, ce cadre est appelé accélération du personnel, en raison de la rapidité du changement requise dans les projets d'adoption du cloud. Pour plus d'informations, veuillez consulter le guide OCM.

contrôle d'accès d'origine (OAC)

Dans CloudFront, une option améliorée pour restreindre l'accès afin de sécuriser votre contenu Amazon Simple Storage Service (Amazon S3). L'OAC prend en charge tous les compartiments S3 dans leur ensemble Régions AWS, le chiffrement côté serveur avec AWS KMS (SSE-KMS) et les requêtes dynamiques PUT adressées au compartiment S3. DELETE

identité d'accès d'origine (OAI)

Dans CloudFront, une option permettant de restreindre l'accès afin de sécuriser votre contenu Amazon S3. Lorsque vous utilisez OAI, il CloudFront crée un principal auprès duquel Amazon S3 peut s'authentifier. Les principaux authentifiés ne peuvent accéder au contenu d'un compartiment

S3 que par le biais d'une distribution spécifique CloudFront . Voir également OAC, qui fournit un contrôle d'accès plus précis et amélioré.

OU

Voir l'examen de l'état de préparation opérationnelle.

DE

Voir technologie opérationnelle.

VPC sortant (de sortie)

Dans une architecture AWS multi-comptes, un VPC qui gère les connexions réseau initiées depuis une application. L'architecture de référence de sécuritéAWS recommande de configurer votre compte réseau avec des VPC entrants, sortants et d'inspection afin de protéger l'interface bidirectionnelle entre votre application et Internet en général.

P

limite des autorisations

Politique de gestion IAM attachée aux principaux IAM pour définir les autorisations maximales que peut avoir l'utilisateur ou le rôle. Pour plus d'informations, veuillez consulter la rubrique Limites des autorisations dans la documentation IAM.

informations personnelles identifiables (PII)

Informations qui, lorsqu'elles sont consultées directement ou associées à d'autres données connexes, peuvent être utilisées pour déduire raisonnablement l'identité d'une personne. Les exemples d'informations personnelles incluent les noms, les adresses et les informations de contact.

PII

Voir les informations personnelles identifiables.

manuel stratégique

Ensemble d'étapes prédéfinies qui capturent le travail associé aux migrations, comme la fourniture de fonctions d'opérations de base dans le cloud. Un manuel stratégique peut revêtir la forme de scripts, de runbooks automatisés ou d'un résumé des processus ou des étapes nécessaires au fonctionnement de votre environnement modernisé.

PLC

Voir contrôleur logique programmable.

PLM

Consultez la section Gestion du cycle de vie des produits.

politique

Objet capable de définir les autorisations (voir la politique basée sur l'identité), de spécifier les conditions d'accès (voir la politique basée sur les ressources) ou de définir les autorisations maximales pour tous les comptes d'une organisation dans AWS Organizations (voir la politique de contrôle des services).

persistance polyglotte

Choix indépendant de la technologie de stockage de données d'un microservice en fonction des modèles d'accès aux données et d'autres exigences. Si vos microservices utilisent la même technologie de stockage de données, ils peuvent rencontrer des difficultés d'implémentation ou présenter des performances médiocres. Les microservices sont plus faciles à mettre en œuvre, atteignent de meilleures performances, ainsi qu'une meilleure capacité de mise à l'échelle s'ils utilisent l'entrepôt de données le mieux adapté à leurs besoins. Pour plus d'informations, veuillez consulter Enabling data persistence in microservices.

évaluation du portefeuille

Processus de découverte, d'analyse et de priorisation du portefeuille d'applications afin de planifier la migration. Pour plus d'informations, veuillez consulter Evaluating migration readiness.

predicate

Une condition de requête qui renvoie `true` ou`false`, généralement située dans une `WHERE` clause.

prédicat pushdown

Technique d'optimisation des requêtes de base de données qui filtre les données de la requête avant le transfert. Cela réduit la quantité de données qui doivent être extraites et traitées à partir de la base de données relationnelle et améliore les performances des requêtes.

contrôle préventif

Contrôle de sécurité conçu pour empêcher qu'un événement ne se produise. Ces contrôles constituent une première ligne de défense pour empêcher tout accès non autorisé ou toute

modification indésirable de votre réseau. Pour plus d'informations, veuillez consulter Preventative controls dans Implementing security controls on AWS.

principal

Entité AWS capable d'effectuer des actions et d'accéder aux ressources. Cette entité est généralement un utilisateur root pour un Compte AWS rôle IAM ou un utilisateur. Pour plus d'informations, veuillez consulter la rubrique Principal dans Termes et concepts relatifs aux rôles, dans la documentation IAM.

Confidentialité dès la conception

Une approche de l'ingénierie des systèmes qui prend en compte la confidentialité tout au long du processus d'ingénierie.

zones hébergées privées

Conteneur qui contient des informations concernant la façon dont vous souhaitez qu'Amazon Route 53 réponde aux requêtes DNS pour un domaine et ses sous-domaines dans un ou plusieurs VPC. Pour plus d'informations, veuillez consulter Working with private hosted zones dans la documentation Route 53.

contrôle proactif

Contrôle de sécurité conçu pour empêcher le déploiement de ressources non conformes. Ces contrôles analysent les ressources avant qu'elles ne soient provisionnées. Si la ressource n'est pas conforme au contrôle, elle n'est pas provisionnée. Pour plus d'informations, consultez le guide de référence sur les contrôles dans la AWS Control Tower documentation et consultez la section Contrôles proactifs dans Implémentation des contrôles de sécurité sur AWS.

gestion du cycle de vie des produits (PLM)

Gestion des données et des processus d'un produit tout au long de son cycle de vie, depuis la conception, le développement et le lancement, en passant par la croissance et la maturité, jusqu'au déclin et au retrait.

environnement de production

Voir environnement.

contrôleur logique programmable (PLC)

Dans le secteur manufacturier, un ordinateur hautement fiable et adaptable qui surveille les machines et automatise les processus de fabrication.

pseudonymisation

Processus de remplacement des identifiants personnels dans un ensemble de données par des valeurs fictives. La pseudonymisation peut contribuer à protéger la vie privée. Les données pseudonymisées sont toujours considérées comme des données personnelles.

publier/souscrire (pub/sub)

Modèle qui permet des communications asynchrones entre les microservices afin d'améliorer l'évolutivité et la réactivité. Par exemple, dans un MES basé sur des microservices, un microservice peut publier des messages d'événements sur un canal auquel d'autres microservices peuvent s'abonner. Le système peut ajouter de nouveaux microservices sans modifier le service de publication.

Q

plan de requête

Série d'étapes, telles que des instructions, utilisées pour accéder aux données d'un système de base de données relationnelle SQL.

régression du plan de requêtes

Le cas où un optimiseur de service de base de données choisit un plan moins optimal qu'avant une modification donnée de l'environnement de base de données. Cela peut être dû à des changements en termes de statistiques, de contraintes, de paramètres d'environnement, de liaisons de paramètres de requêtes et de mises à jour du moteur de base de données.

R

Matrice RACI

Voir responsable, responsable, consulté, informé (RACI).

rançongiciel

Logiciel malveillant conçu pour bloquer l'accès à un système informatique ou à des données jusqu'à ce qu'un paiement soit effectué.

Matrice RASCI

Voir responsable, responsable, consulté, informé (RACI).

RCAC

Voir contrôle d'accès aux lignes et aux colonnes.

réplica en lecture

Copie d'une base de données utilisée en lecture seule. Vous pouvez acheminer les requêtes vers le réplica de lecture pour réduire la charge sur votre base de données principale.

réarchitecte

Voir 7 Rs.

objectif de point de récupération (RPO)

Durée maximale acceptable depuis le dernier point de récupération des données. Cela permet de déterminer ce qui est considéré comme une perte de données acceptable entre le dernier point de restauration et l'interruption du service.

objectif de temps de récupération (RTO)

Le délai maximum acceptable entre l'interruption du service et le rétablissement du service.

refactoriser

Voir 7 Rs.

Région

Un ensemble de AWS ressources dans une zone géographique. Chacune Région AWS est isolée et indépendante des autres pour garantir tolérance aux pannes, stabilité et résilience. Pour plus d'informations, voir Spécifier ce que Régions AWS votre compte peut utiliser.

régression

Technique de ML qui prédit une valeur numérique. Par exemple, pour résoudre le problème « Quel sera le prix de vente de cette maison ? », un modèle de ML pourrait utiliser un modèle de régression linéaire pour prédire le prix de vente d'une maison sur la base de faits connus à son sujet (par exemple, la superficie en mètres carrés).

réhéberger

Voir 7 Rs.

version

Dans un processus de déploiement, action visant à promouvoir les modifications apportées à un environnement de production.

déplacer

Voir 7 Rs.

replateforme

Voir 7 Rs.

rachat

Voir 7 Rs.

résilience

La capacité d'une application à résister aux perturbations ou à s'en remettre. La haute disponibilité et la reprise après sinistre sont des considérations courantes lors de la planification de la résilience dans le AWS Cloud. Pour plus d'informations, consultez AWS Cloud Résilience.

politique basée sur les ressources

Politique attachée à une ressource, comme un compartiment Amazon S3, un point de terminaison ou une clé de chiffrement. Ce type de politique précise les principaux auxquels l'accès est autorisé, les actions prises en charge et toutes les autres conditions qui doivent être remplies.

matrice responsable, redevable, consulté et informé (RACI)

Une matrice qui définit les rôles et les responsabilités de toutes les parties impliquées dans les activités de migration et les opérations cloud. Le nom de la matrice est dérivé des types de responsabilité définis dans la matrice : responsable (R), responsable (A), consulté (C) et informé (I). Le type de support (S) est facultatif. Si vous incluez le support, la matrice est appelée matrice RASCI, et si vous l'excluez, elle est appelée matrice RACI.

contrôle réactif

Contrôle de sécurité conçu pour permettre de remédier aux événements indésirables ou aux écarts par rapport à votre référence de sécurité. Pour plus d'informations, veuillez consulter la rubrique Responsive controls dans Implementing security controls on AWS.

retain

Voir 7 Rs.

se retirer

Voir 7 Rs.

rotation

Processus de mise à jour périodique d'un secret pour empêcher un attaquant d'accéder aux informations d'identification.

contrôle d'accès aux lignes et aux colonnes (RCAC)

Utilisation d'expressions SQL simples et flexibles dotées de règles d'accès définies. Le RCAC comprend des autorisations de ligne et des masques de colonnes.

RPO

Voir l'objectif du point de récupération.

RTO

Voir l'objectif en matière de temps de rétablissement.

runbook

Ensemble de procédures manuelles ou automatisées nécessaires à l'exécution d'une tâche spécifique. Elles visent généralement à rationaliser les opérations ou les procédures répétitives présentant des taux d'erreur élevés.

S

SAML 2.0

Un standard ouvert utilisé par de nombreux fournisseurs d'identité (IdPs). Cette fonctionnalité permet l'authentification unique fédérée (SSO), afin que les utilisateurs puissent se connecter AWS Management Console ou appeler les opérations d' AWS API sans que vous ayez à créer un utilisateur dans IAM pour tous les membres de votre organisation. Pour plus d'informations sur la fédération SAML 2.0, veuillez consulter À propos de la fédération SAML 2.0 dans la documentation IAM.

SCADA

Voir Contrôle de supervision et acquisition de données.

SCP

Voir la politique de contrôle des services.

secret

Dans AWS Secrets Manager des informations confidentielles ou restreintes, telles qu'un mot de passe ou des informations d'identification utilisateur, que vous stockez sous forme cryptée. Il comprend la valeur secrète et ses métadonnées. La valeur secrète peut être binaire, une chaîne unique ou plusieurs chaînes. Pour plus d'informations, voir Que contient le secret d'un Secrets Manager ? dans la documentation de Secrets Manager.

contrôle de sécurité

Barrière de protection technique ou administrative qui empêche, détecte ou réduit la capacité d'un assaillant d'exploiter une vulnérabilité de sécurité. Il existe quatre principaux types de contrôles de sécurité : préventifs, détectifs, réactifs et proactifs.

renforcement de la sécurité

Processus qui consiste à réduire la surface d'attaque pour la rendre plus résistante aux attaques. Cela peut inclure des actions telles que la suppression de ressources qui ne sont plus requises, la mise en œuvre des bonnes pratiques de sécurité consistant à accorder le moindre privilège ou la désactivation de fonctionnalités inutiles dans les fichiers de configuration.

système de gestion des informations et des événements de sécurité (SIEM)

Outils et services qui associent les systèmes de gestion des informations de sécurité (SIM) et de gestion des événements de sécurité (SEM). Un système SIEM collecte, surveille et analyse les données provenant de serveurs, de réseaux, d'appareils et d'autres sources afin de détecter les menaces et les failles de sécurité, mais aussi de générer des alertes.

automatisation des réponses de sécurité

Action prédéfinie et programmée conçue pour répondre automatiquement à un événement de sécurité ou y remédier. Ces automatisations servent de contrôles de sécurité détectifs ou réactifs qui vous aident à mettre en œuvre les meilleures pratiques AWS de sécurité. Parmi les actions de réponse automatique, citons la modification d'un groupe de sécurité VPC, l'application de correctifs à une instance Amazon EC2 ou la rotation des informations d'identification.

chiffrement côté serveur

Chiffrement des données à destination, par celui Service AWS qui les reçoit.

Politique de contrôle des services (SCP)

Politique qui propose un contrôle centralisé des autorisations pour tous les comptes d'une organisation dans AWS Organizations. Les SCP définissent des barrières de protection ou des

limites aux actions qu'un administrateur peut déléguer à des utilisateurs ou à des rôles. Vous pouvez utiliser les SCP comme listes d'autorisation ou de refus, pour indiquer les services ou les actions autorisés ou interdits. Pour plus d'informations, consultez la section Politiques de contrôle des services dans la AWS Organizations documentation.

point de terminaison du service

URL du point d'entrée pour un Service AWS. Pour vous connecter par programmation au service cible, vous pouvez utiliser un point de terminaison. Pour plus d'informations, veuillez consulter la rubrique Service AWS endpoints dans Références générales AWS.

contrat de niveau de service (SLA)

Accord qui précise ce qu'une équipe informatique promet de fournir à ses clients, comme le temps de disponibilité et les performances des services.

indicateur de niveau de service (SLI)

Mesure d'un aspect des performances d'un service, tel que son taux d'erreur, sa disponibilité ou son débit.

objectif de niveau de service (SLO)

Mesure cible qui représente l'état d'un service, tel que mesuré par un indicateur de niveau de service.

modèle de responsabilité partagée

Un modèle décrivant la responsabilité que vous partagez en matière AWS de sécurité et de conformité dans le cloud. AWS est responsable de la sécurité du cloud, alors que vous êtes responsable de la sécurité dans le cloud. Pour de plus amples informations, veuillez consulter Modèle de responsabilité partagée.

SIEM

Consultez les informations de sécurité et le système de gestion des événements.

point de défaillance unique (SPOF)

Défaillance d'un seul composant critique d'une application susceptible de perturber le système.

SLA

Voir le contrat de niveau de service.

SLI

Voir l'indicateur de niveau de service.

SLO

Voir l'objectif de niveau de service.

split-and-seed modèle

Modèle permettant de mettre à l'échelle et d'accélérer les projets de modernisation. Au fur et à mesure que les nouvelles fonctionnalités et les nouvelles versions de produits sont définies, l'équipe principale se divise pour créer des équipes de produit. Cela permet de mettre à l'échelle les capacités et les services de votre organisation, d'améliorer la productivité des développeurs et de favoriser une innovation rapide. Pour plus d'informations, consultez la section Approche progressive de la modernisation des applications dans le. AWS Cloud

SPOF

Voir point de défaillance unique.

schéma en étoile

Structure organisationnelle de base de données qui utilise une grande table de faits pour stocker les données transactionnelles ou mesurées et utilise une ou plusieurs tables dimensionnelles plus petites pour stocker les attributs des données. Cette structure est conçue pour être utilisée dans un entrepôt de données ou à des fins de business intelligence.

modèle de figuier étrangleur

Approche de modernisation des systèmes monolithiques en réécrivant et en remplaçant progressivement les fonctionnalités du système jusqu'à ce que le système hérité puisse être mis hors service. Ce modèle utilise l'analogie d'un figuier de vigne qui se développe dans un arbre existant et qui finit par supplanter son hôte. Le schéma a été présenté par Martin Fowler comme un moyen de gérer les risques lors de la réécriture de systèmes monolithiques. Pour obtenir un exemple d'application de ce modèle, veuillez consulter Modernizing legacy Microsoft ASP.NET (ASMX) web services incrementally by using containers and Amazon API Gateway.

sous-réseau

Plage d'adresses IP dans votre VPC. Un sous-réseau doit se trouver dans une seule zone de disponibilité.

contrôle de supervision et acquisition de données (SCADA)

Dans le secteur manufacturier, un système qui utilise du matériel et des logiciels pour surveiller les actifs physiques et les opérations de production.

chiffrement symétrique

Algorithme de chiffrement qui utilise la même clé pour chiffrer et déchiffrer les données.

tests synthétiques

Tester un système de manière à simuler les interactions des utilisateurs afin de détecter les problèmes potentiels ou de surveiller les performances. Vous pouvez utiliser Amazon CloudWatch Synthetics pour créer ces tests.

T

balises

Des paires clé-valeur qui agissent comme des métadonnées pour organiser vos AWS ressources. Les balises peuvent vous aider à gérer, identifier, organiser, rechercher et filtrer des ressources. Pour plus d'informations, veuillez consulter la rubrique Balisage de vos AWS ressources.

variable cible

La valeur que vous essayez de prédire dans le cadre du ML supervisé. Elle est également qualifiée de variable de résultat. Par exemple, dans un environnement de fabrication, la variable cible peut être un défaut du produit.

liste de tâches

Outil utilisé pour suivre les progrès dans un runbook. Liste de tâches qui contient une vue d'ensemble du runbook et une liste des tâches générales à effectuer. Pour chaque tâche générale, elle inclut le temps estimé nécessaire, le propriétaire et l'avancement.

environnement de test

Voir environnement.

entraînement

Pour fournir des données à partir desquelles votre modèle de ML peut apprendre. Les données d'entraînement doivent contenir la bonne réponse. L'algorithme d'apprentissage identifie des modèles dans les données d'entraînement, qui mettent en correspondance les attributs des données d'entrée avec la cible (la réponse que vous souhaitez prédire). Il fournit un modèle de ML qui capture ces modèles. Vous pouvez alors utiliser le modèle de ML pour obtenir des prédictions sur de nouvelles données pour lesquelles vous ne connaissez pas la cible.

passerelle de transit

Hub de transit de réseau que vous pouvez utiliser pour relier vos VPC et vos réseaux sur site. Pour plus d'informations, voir Qu'est-ce qu'une passerelle de transit dans la AWS Transit Gateway documentation.

flux de travail basé sur jonction

Approche selon laquelle les développeurs génèrent et testent des fonctionnalités localement dans une branche de fonctionnalités, puis fusionnent ces modifications dans la branche principale. La branche principale est ensuite intégrée aux environnements de développement, de préproduction et de production, de manière séquentielle.

accès sécurisé

Accorder des autorisations à un service que vous spécifiez pour effectuer des tâches au sein de votre organisation AWS Organizations et dans ses comptes en votre nom. Le service de confiance crée un rôle lié au service dans chaque compte, lorsque ce rôle est nécessaire, pour effectuer des tâches de gestion à votre place. Pour plus d'informations, consultez la section Utilisation AWS Organizations avec d'autres AWS services dans la AWS Organizations documentation.

réglage

Pour modifier certains aspects de votre processus d'entraînement afin d'améliorer la précision du modèle de ML. Par exemple, vous pouvez entraîner le modèle de ML en générant un ensemble d'étiquetage, en ajoutant des étiquettes, puis en répétant ces étapes plusieurs fois avec différents paramètres pour optimiser le modèle.

équipe de deux pizzas

Une petite DevOps équipe que vous pouvez nourrir avec deux pizzas. Une équipe de deux pizzas garantit les meilleures opportunités de collaboration possible dans le développement de logiciels.

U

incertitude

Un concept qui fait référence à des informations imprécises, incomplètes ou inconnues susceptibles de compromettre la fiabilité des modèles de ML prédictifs. Il existe deux types d'incertitude : l'incertitude épistémique est causée par des données limitées et incomplètes, alors que l'incertitude aléatoire est causée par le bruit et le caractère aléatoire inhérents aux données.

Pour plus d'informations, veuillez consulter le guide Quantifying uncertainty in deep learning systems.

tâches indifférenciées

Également connu sous le nom de « levage de charges lourdes », ce travail est nécessaire pour créer et exploiter une application, mais qui n'apporte pas de valeur directe à l'utilisateur final ni d'avantage concurrentiel. Les exemples de tâches indifférenciées incluent l'approvisionnement, la maintenance et la planification des capacités.

environnements supérieurs

Voir environnement.

V

mise à vide

Opération de maintenance de base de données qui implique un nettoyage après des mises à jour incrémentielles afin de récupérer de l'espace de stockage et d'améliorer les performances.

contrôle de version

Processus et outils permettant de suivre les modifications, telles que les modifications apportées au code source dans un référentiel.

Appairage de VPC

Connexion entre deux VPC qui vous permet d'acheminer le trafic à l'aide d'adresses IP privées. Pour plus d'informations, veuillez consulter la rubrique Qu'est-ce que l'appairage de VPC ? dans la documentation Amazon VPC.

vulnérabilités

Défaut logiciel ou matériel qui compromet la sécurité du système.

W

cache actif

Cache tampon qui contient les données actuelles et pertinentes fréquemment consultées. L'instance de base de données peut lire à partir du cache tampon, ce qui est plus rapide que la lecture à partir de la mémoire principale ou du disque.

données chaudes

Données rarement consultées. Lorsque vous interrogez ce type de données, des requêtes modérément lentes sont généralement acceptables.

fonction de fenêtre

Fonction SQL qui effectue un calcul sur un groupe de lignes liées d'une manière ou d'une autre à l'enregistrement en cours. Les fonctions de fenêtre sont utiles pour traiter des tâches, telles que le calcul d'une moyenne mobile ou l'accès à la valeur des lignes en fonction de la position relative de la ligne en cours.

charge de travail

Ensemble de ressources et de code qui fournit une valeur métier, par exemple une application destinée au client ou un processus de backend.

flux de travail

Groupes fonctionnels d'un projet de migration chargés d'un ensemble de tâches spécifique. Chaque flux de travail est indépendant, mais prend en charge les autres flux de travail du projet. Par exemple, le flux de travail du portefeuille est chargé de prioriser les applications, de planifier les vagues et de collecter les métadonnées de migration. Le flux de travail du portefeuille fournit ces actifs au flux de travail de migration, qui migre ensuite les serveurs et les applications.

VER

Voir écrire une fois, lire plusieurs.

WQF

Consultez le cadre de qualification des charges de travail AWS.

écrire une fois, lire plusieurs (WORM)

Modèle de stockage qui écrit les données une seule fois et empêche leur suppression ou leur modification. Les utilisateurs autorisés peuvent lire les données autant de fois que nécessaire, mais ils ne peuvent pas les modifier. Cette infrastructure de stockage de données est considérée comme immuable.

Z

exploit Zero-Day

Une attaque, généralement un logiciel malveillant, qui tire parti d'une vulnérabilité de type « jour zéro ».

vulnérabilité de type « jour zéro »

Une faille ou une vulnérabilité non atténuée dans un système de production. Les acteurs malveillants peuvent utiliser ce type de vulnérabilité pour attaquer le système. Les développeurs prennent souvent conscience de la vulnérabilité à la suite de l'attaque.

application zombie

Application dont l'utilisation moyenne du processeur et de la mémoire est inférieure à 5 %. Dans un projet de migration, il est courant de retirer ces applications.

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.