AWS Principes fondamentaux de plusieurs régions

AWS Directives prescriptives

AWS Directives prescriptives: AWS Principes fondamentaux de plusieurs régions

Table of Contents

AWS Principes fondamentaux de plusieurs régions

John Formento, Amazon Web Services ()AWS

Décembre 2024 (historique du document)

Ce guide avancé de 300 niveaux est destiné aux architectes du cloud et aux cadres supérieurs qui élaborent des charges de travail sur une architecture multirégionale AWS et souhaitent l'utiliser pour améliorer la résilience de leurs charges de travail. Ce guide suppose une connaissance de base de l' AWS infrastructure et des services. Il décrit les cas d'utilisation multirégionaux courants, partage les concepts multirégionaux fondamentaux et les implications en matière de conception, de développement et de déploiement, et fournit des conseils prescriptifs pour vous aider à mieux déterminer si une architecture multirégionale convient à vos charges de travail.

Êtes-vous Well-Architected ?

Le AWS Well-Architected Framework vous aide à comprendre les avantages et les inconvénients des décisions que vous prenez lorsque vous créez des systèmes dans le cloud. Les six piliers du cadre fournissent les meilleures pratiques architecturales pour concevoir et exploiter des systèmes fiables, sécurisés, efficaces, rentables et durables. Vous pouvez utiliser le AWS Well-Architected Tool, qui est disponible gratuitement sur le AWS Management Console, pour évaluer votre charge de travail par rapport à ces meilleures pratiques en répondant à une série de questions pour chaque pilier.

Pour obtenir des conseils d'experts supplémentaires et les meilleures pratiques relatives à votre architecture cloud, notamment des déploiements d'architecture de référence, des diagrammes et des guides techniques, consultez le Centre AWS d'architecture.

Introduction

Chacune Région AWSse compose de plusieurs zones de disponibilité indépendantes et physiquement séparées au sein d'une zone géographique. Une séparation logique stricte entre les services logiciels de chaque région est maintenue. Cette conception ciblée garantit qu'une défaillance d'infrastructure ou de service dans une région n'entraîne pas une défaillance corrélée dans une autre région.

La plupart des AWS utilisateurs peuvent atteindre leurs objectifs de résilience pour une charge de travail dans une seule région en utilisant plusieurs zones de disponibilité ou régionales Services

AWS. Cependant, un sous-ensemble d'utilisateurs opte pour des architectures multirégionales pour trois raisons :

- Ils ont des exigences élevées en matière de disponibilité et de continuité des opérations pour leurs charges de travail les plus élevées et souhaitent établir un temps de rétablissement limité en cas de défaillance ayant un impact sur les ressources d'une seule région.
- Ils doivent satisfaire aux exigences de souveraineté des données (telles que le respect des lois, réglementations et conformité locales) qui obligent les charges de travail à opérer dans une juridiction donnée.
- Ils doivent améliorer les performances et l'expérience client en fonction de la charge de travail en exécutant les charges de travail sur les sites les plus proches de leurs utilisateurs finaux.

Ce guide met l'accent sur les exigences de haute disponibilité et de continuité des opérations, et vous aide à comprendre les considérations relatives à l'adoption d'une architecture multirégionale pour une charge de travail. Il décrit les concepts fondamentaux qui s'appliquent à la conception, au développement et au déploiement d'une charge de travail multirégionale, et fournit un cadre normatif pour vous aider à déterminer si une architecture multirégionale est le bon choix pour une charge de travail particulière. Vous devez vous assurer qu'une architecture multirégionale est le bon choix pour votre charge de travail, car ces architectures sont difficiles, et si l'architecture multirégionale n'est pas construite correctement, il est possible que la disponibilité globale de la charge de travail diminue.

Ingénierie et exploitation pour la résilience dans une seule région

Avant de vous plonger dans les concepts multirégionaux, commencez par vérifier que votre charge de travail est déjà aussi résiliente que possible dans une seule région. Pour y parvenir, évaluez votre charge de travail par rapport au pilier de fiabilité et au pilier d'excellence opérationnelle du AWS Well-Architected Framework, et apportez les modifications nécessaires en fonction des compromis et de l'évaluation des risques. Les concepts suivants sont abordés dans le AWS Well-Architected Framework :

- Segmentation de la charge de travail en fonction des limites du domaine
- Contrats de service bien définis
- Gestion des dépendances et couplage
- Gestion des échecs, des nouvelles tentatives et des stratégies de réduction
- Opérations idempotentes et transactions avec état par opposition aux transactions apatrides
- Préparation opérationnelle et gestion du changement
- Comprendre l'état des charges de travail
- Réagir aux événements

Pour approfondir la résilience d'une seule région, passez en revue et appliquez les concepts abordés dans le document Advanced Multi-AZ Resilience Patterns : Detecting and Mitigating Gray Failures. Ce paper présente les meilleures pratiques pour utiliser des répliques dans chaque zone de disponibilité afin de contenir les défaillances et développe les concepts multi-AZ introduits dans le AWS Well Architected Framework. Bien qu'une architecture multirégionale puisse atténuer les modes de défaillance liés aux zones de disponibilité, l'approche multirégionale comporte des inconvénients que vous devez prendre en compte. C'est pourquoi nous vous recommandons de commencer par une approche multi-AZ, puis d'évaluer une charge de travail spécifique par rapport aux principes fondamentaux des architectures multirégionales afin de déterminer si une approche multirégionale peut augmenter la résilience de la charge de travail.

Fondement multirégional 1 : Comprendre les exigences

Comme indiqué précédemment, la haute disponibilité et la continuité des opérations sont des raisons courantes pour lesquelles on opte pour des architectures multirégionales. Les indicateurs de disponibilité mesurent le pourcentage de temps pendant lequel une charge de travail est disponible pour être utilisée sur une période définie, tandis que les indicateurs de continuité des opérations mesurent le temps de reprise pour des événements à grande échelle, généralement de plus longue durée.

La mesure de la disponibilité est un processus quasi continu. Les mesures spécifiques peuvent varier, mais elles s'articulent généralement autour d'un indicateur de disponibilité cible, le plus souvent appelé neuf (tel que 99,99 % de disponibilité). En ce qui concerne les objectifs de disponibilité, il n'y a pas de solution universelle. Vous devez définir des objectifs de disponibilité au niveau de la charge de travail et séparer les composants non critiques des composants critiques, au lieu d'appliquer un seul objectif à toutes les charges de travail.

Pour assurer la continuité des opérations, les point-in-time mesures suivantes sont généralement utilisées :

- Objectif de temps de rétablissement (RTO) — Le RTO est le délai maximum acceptable entre l'interruption du service et le rétablissement du service. Cette valeur détermine une durée acceptable pendant laquelle le service est perturbé.
- Objectif de point de restauration (RPO) — Le RPO est le délai maximum 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 une interruption de service.

À l'instar de la définition d'objectifs de disponibilité, le RTO et le RPO doivent également être définis au niveau de la charge de travail. Une continuité des opérations plus agressive ou une haute disponibilité nécessitent des investissements accrus. Cela dit, toutes les applications ne peuvent pas exiger ou ne nécessitent pas le même niveau de résilience. Aligner les responsables commerciaux et informatiques pour évaluer la criticité des applications en fonction de l'impact commercial, puis les hiérarchiser en conséquence peut aider à fournir un point de départ. Les tableaux suivants fournissent des exemples de hiérarchisation.

Ce tableau présente un exemple de hiérarchisation de la résilience pour les accords de niveau de service (). SLAs

Niveau de résilience	Disponibilité SLA	Temps d'arrêt acceptable/an
Platine	99,99 %	52,60 minutes
Doré	99,90 %	8,77 heures
Argent	99,5 %	1,83 jours

Le tableau suivant présente un exemple de hiérarchisation de la résilience pour le RTO et le RPO.

Niveau de résilience	RTO maximal	RPO maximal	Critères	Coût
Platine	15 minutes	5 minutes	Charges de travail critiques	$$$
Doré	15 minutes — 6 heures	2 heures	Charges de travail importantes mais non critiques	$$
Argent	6 heures — quelques jours	24 heures	Charges de travail non critiques	$

Lorsque vous concevez des charges de travail dans un souci de résilience, tenez compte de la relation entre haute disponibilité et continuité des opérations. Par exemple, si une charge de travail nécessite une disponibilité de 99,99 %, un temps d'arrêt maximal de 53 minutes par an est tolérable. La détection d'une panne peut prendre au moins 5 minutes et un opérateur peut prendre 10 minutes supplémentaires pour intervenir, prendre des décisions sur les étapes de restauration et exécuter ces étapes. Il n'est pas rare que 30 à 45 minutes soient nécessaires pour remédier à un seul problème. Dans ce cas, il est avantageux d'avoir une stratégie multirégionale pour fournir une instance isolée qui supprime l'impact corrélé. Cela permet de poursuivre les opérations en basculant dans un délai limité pendant que vous triez indépendamment la déficience initiale. C'est là qu'il est nécessaire de définir le temps de rétablissement limité approprié et de garantir l'alignement.

Une approche multirégionale peut être appropriée pour les charges de travail critiques présentant des besoins de disponibilité extrêmes (par exemple, une disponibilité de 99,99 % ou plus) ou des exigences strictes en matière de continuité des opérations qui ne peuvent être satisfaites qu'en basculant vers une autre région. Toutefois, ces exigences ne s'appliquent généralement qu'à un petit sous-ensemble du portefeuille de charges de travail d'une entreprise dont le temps de restauration est limité, mesuré en minutes ou en heures. À moins qu'une application n'ait besoin d'un temps de restauration de quelques minutes ou de quelques heures, il peut être préférable d'attendre qu'une interruption régionale de l'application soit corrigée dans la région affectée. Cette approche est généralement adaptée aux charges de travail de niveau inférieur.

Avant de mettre en œuvre une architecture multirégionale, les décideurs commerciaux et les équipes techniques doivent s'entendre sur les implications financières, y compris les facteurs de coûts opérationnels et d'infrastructure. Une architecture multirégionale typique peut entraîner un coût deux fois plus élevé qu'une approche à région unique. Bien qu'il existe plusieurs modèles multirégionaux de continuité des activités, tels que le fonctionnement en veille chaude, en veille chaude ou en veilleuse, le modèle présentant le moins de risques d'atteindre les objectifs de reprise impliquera le fonctionnement en mode veille chaud, ce qui doublera le coût de votre charge de travail.

Principaux conseils

- Les objectifs de disponibilité et de continuité des opérations tels que le RTO et le RPO doivent être établis par charge de travail et alignés sur les parties prenantes commerciales et informatiques.
- La plupart des objectifs de disponibilité et de continuité des opérations peuvent être atteints au sein d'une seule région. Pour les objectifs qui ne peuvent être atteints au sein d'une seule région, envisagez une approche multirégionale avec une vision claire des compromis entre les coûts, la complexité et les avantages.

Deuxième élément fondamental pour plusieurs régions : comprendre les données

La gestion des données n'est pas un problème trivial lorsque vous adoptez des architectures multirégionales. La distance géographique entre les régions impose une latence inévitable qui se traduit par le temps nécessaire pour répliquer les données entre les régions. Des compromis entre la disponibilité, la cohérence des données et l'introduction d'une latence plus élevée dans une charge de travail utilisant une architecture multirégionale seront nécessaires. Que vous utilisiez la réplication asynchrone ou synchrone, vous devez modifier votre application pour gérer les changements de comportement imposés par la technologie de réplication. Les défis liés à la cohérence des données et à la latence font qu'il est très difficile de transformer une application existante conçue pour une seule région en une application multirégionale. Il est essentiel de comprendre les exigences de cohérence des données et les modèles d'accès aux données pour des charges de travail particulières pour évaluer les compromis.

2.a : Comprendre les exigences de cohérence des données

Le théorème CAP fournit une référence pour raisonner sur les compromis entre la cohérence des données, la disponibilité et les partitions du réseau. Seules deux de ces exigences peuvent être satisfaites en même temps pour une charge de travail. Par définition, une architecture multirégionale inclut des partitions réseau entre les régions. Vous devez donc choisir entre disponibilité et cohérence.

Si vous sélectionnez la disponibilité des données entre les régions, vous ne subirez pas de latence importante lors des opérations d'écriture transactionnelles, car le recours à la réplication asynchrone des données validées entre les régions réduit la cohérence entre les régions jusqu'à la fin de la réplication. Avec la réplication asynchrone, en cas de défaillance dans la région principale, il est fort probable que les opérations d'écriture soient en attente de réplication depuis la région principale. Cela conduit à un scénario dans lequel les données les plus récentes ne sont pas disponibles tant que la réplication ne reprend pas, et un processus de rapprochement est nécessaire pour gérer les transactions en cours qui n'ont pas été répliquées depuis la région ayant subi l'interruption. Ce scénario nécessite de comprendre votre logique métier et de créer un processus spécifique pour rejouer la transaction ou comparer les banques de données entre les régions.

Pour les charges de travail où la réplication asynchrone est privilégiée, vous pouvez utiliser des services tels qu'Amazon Aurora et Amazon DynamoDB pour la réplication asynchrone entre régions.

Les bases de données mondiales Amazon Aurora et les tables globales Amazon DynamoDB disposent de métriques CloudWatchAmazon par défaut pour faciliter le suivi du délai de réplication. Une base de données globale Aurora se compose d'une région principale dans laquelle vos données sont écrites et d'un maximum de cinq régions secondaires en lecture seule. Les tables globales DynamoDB se composent de tables de répliques multiactives couvrant un nombre illimité de régions dans lesquelles vos données sont écrites et lues.

L'ingénierie de la charge de travail pour tirer parti des architectures axées sur les événements constitue un avantage pour une stratégie multirégionale, car cela signifie que la charge de travail peut inclure la réplication asynchrone des données et permet la reconstruction de l'état en rejouant les événements. Étant donné que les services de streaming et de messagerie mettent en mémoire tampon les données utiles des messages dans une seule région, un processus de basculement ou de retour en arrière régional doit inclure un mécanisme permettant de rediriger les flux de données d'entrée des clients. Le processus doit également concilier les charges utiles en vol ou non livrées stockées dans la région qui a subi la perturbation.

Si vous optez pour l'exigence de cohérence CAP et que vous utilisez une base de données répliquée de manière synchrone entre les régions pour prendre en charge vos applications exécutées simultanément à partir de plusieurs régions, vous éliminez le risque de perte de données et vous maintenez les données synchronisées entre les régions. Cependant, cela introduit des caractéristiques de latence plus élevées, car les écritures doivent être validées dans plusieurs régions, et les régions peuvent se trouver à des centaines ou des milliers de kilomètres les unes des autres. Vous devez tenir compte de cette caractéristique de latence dans la conception de votre application. En outre, la réplication synchrone peut entraîner des défaillances corrélées, car les écritures devront être validées dans plusieurs régions pour réussir. S'il y a un handicap dans une région, vous devrez réunir un quorum pour que les écritures soient couronnées de succès. Cela implique généralement de configurer votre base de données dans trois régions et d'établir un quorum de deux régions sur trois. Les technologies telles que Paxos peuvent aider à répliquer et à valider les données de manière synchrone, mais nécessitent un investissement important pour les développeurs.

Lorsque les écritures impliquent une réplication synchrone entre plusieurs régions pour répondre à de fortes exigences de cohérence, la latence d'écriture augmente d'un ordre de grandeur. Une latence d'écriture plus élevée n'est généralement pas quelque chose que vous pouvez intégrer à une application sans modifications importantes, telles que la révision du délai d'expiration et de la stratégie de nouvelle tentative de votre application. Idéalement, il doit être pris en compte lors de la conception initiale de l'application. Pour les charges de travail multirégionales où la réplication synchrone est une priorité, les AWS Partner solutions peuvent être utiles.

2.b : Comprendre les modèles d'accès aux données

Les modèles d'accès aux données de charge de travail nécessitent beaucoup de lecture ou d'écriture. Comprendre cette caractéristique pour une charge de travail particulière vous aidera à sélectionner une architecture multirégionale appropriée.

Pour les charges de travail intensives en lecture, telles que le contenu statique entièrement en lecture seule, vous pouvez obtenir une architecture multirégion active-active qui présente une complexité d'ingénierie moindre par rapport à une charge de travail intensive en écriture. La diffusion de contenu statique à la périphérie à l'aide d'un réseau de diffusion de contenu (CDN) garantit la disponibilité en mettant en cache le contenu le plus proche de l'utilisateur final ; l'utilisation d'ensembles de fonctionnalités tels que le basculement d'origine au sein d'Amazon CloudFront peut y contribuer. Une autre option consiste à déployer le calcul sans état dans plusieurs régions et à utiliser le DNS pour acheminer les utilisateurs vers la région la plus proche afin de lire le contenu. Pour ce faire, vous pouvez utiliser Amazon Route 53 avec une politique de routage par géolocalisation.

Pour les charges de travail intensives en lecture dont le pourcentage de trafic en lecture est supérieur à celui du trafic en écriture, vous pouvez utiliser une stratégie globale de lecture locale et d'écriture globale. Cela signifie que toutes les demandes d'écriture sont envoyées à une base de données d'une région spécifique, que les données sont répliquées de manière asynchrone dans toutes les autres régions et que les lectures peuvent être effectuées dans n'importe quelle région. Cette approche nécessite une charge de travail pour garantir la cohérence finale, car les lectures locales peuvent devenir obsolètes en raison de l'augmentation de la latence lors de la réplication interrégionale des écritures.

Les bases de données globales Aurora peuvent aider à fournir des répliques de lecture dans une région de secours qui peut uniquement gérer l'ensemble du trafic de lecture localement, et à fournir un seul magasin de données principal dans une région spécifique pour gérer le trafic d'écriture. Les données sont répliquées de manière asynchrone de la base de données principale vers les bases de données de secours (répliques en lecture), et les bases de données de secours peuvent être promues au statut principal si vous devez transférer des opérations vers la région de secours. Vous pouvez également utiliser DynamoDB dans le cadre de cette approche. Les tables globales DynamoDB peuvent fournir des répliques de tables réparties dans différentes régions, chacune pouvant être dimensionnée pour prendre en charge n'importe quel volume de trafic local de lecture ou d'écriture. Quand une application écrit des données dans une table de réplique dans une région, DynamoDB propage automatiquement l'écriture aux autres tables de réplique dans les autres régions . Avec cette configuration, les données sont répliquées de manière asynchrone depuis une région principale définie vers des tables de réplication dans des régions de secours. Les tables de

réplication de n'importe quelle région peuvent toujours accepter des écritures. La promotion d'une région de secours au statut de région principale est donc gérée au niveau de l'application. Encore une fois, la charge de travail doit être cohérente, ce qui peut nécessiter une réécriture si elle n'a pas été conçue pour cela dès le départ.

Pour les charges de travail intensives en écriture, une région principale doit être sélectionnée et la capacité de basculer vers une région de secours doit être intégrée à la charge de travail. Par rapport à une approche active-active, une approche de veille primaire comporte des compromis supplémentaires. En effet, pour une architecture active-active, la charge de travail doit être réécrite pour gérer le routage intelligent vers les régions, établir une affinité de session, garantir des transactions idempotentes et gérer les conflits potentiels.

La plupart des charges de travail qui utilisent une approche multirégionale pour la résilience ne nécessiteront pas une approche active-active. Vous pouvez utiliser une stratégie de sharding pour renforcer la résilience en limitant l'impact d'une dépréciation sur l'ensemble de la clientèle. Si vous pouvez partager efficacement une base de clients, vous pouvez sélectionner différentes régions principales pour chaque partition. Par exemple, vous pouvez partager des clients de manière à ce que la moitié des clients soient alignés sur la région 1 et l'autre moitié sur la région deux. En traitant les régions comme des cellules, vous pouvez créer une approche cellulaire multirégionale, ce qui permet de réduire l'impact sur votre charge de travail. Pour plus d'informations, consultez la présentation AWS re:Invent sur cette approche.

Vous pouvez combiner l'approche de partitionnement avec une approche de veille principale afin de fournir des fonctionnalités de basculement pour les partitions. Vous devrez concevoir un processus de basculement testé adapté à la charge de travail ainsi qu'un processus de réconciliation des données, afin de garantir la cohérence transactionnelle des banques de données après le basculement. Elles sont abordées plus en détail plus loin dans ce guide.

Principaux conseils

- Il est fort probable que les écritures en attente de réplication ne soient pas validées dans la région de secours en cas d'échec. Les données ne seront pas disponibles jusqu'à ce que la réplication reprenne (dans l'hypothèse d'une réplication asynchrone).
- Dans le cadre du basculement, un processus de réconciliation des données sera nécessaire pour garantir le maintien d'un état de cohérence transactionnelle pour les magasins de données utilisant la réplication asynchrone. Cela nécessite une logique métier spécifique et n'est pas géré par le magasin de données lui-même.

- Lorsqu'une forte cohérence est requise, les charges de travail doivent être modifiées pour tolérer la latence requise d'un magasin de données qui se réplique de manière synchrone.

Principe fondamental 3 pour plusieurs régions : comprendre les dépendances de votre charge de travail

Une charge de travail spécifique peut avoir plusieurs dépendances dans une région, telles que l' Services AWS utilisation, les dépendances internes, les dépendances tierces, les dépendances réseau, les certificats, les clés, les secrets et les paramètres. Pour garantir le fonctionnement de la charge de travail en cas de panne, il ne doit y avoir aucune dépendance entre la région principale et la région de secours ; chacune doit pouvoir fonctionner indépendamment de l'autre. Pour ce faire, examinez toutes les dépendances de la charge de travail pour vous assurer qu'elles sont disponibles dans chaque région. Cela est nécessaire car une défaillance dans la région principale ne devrait pas affecter la région de secours. En outre, vous devez comprendre le fonctionnement de la charge de travail lorsqu'une dépendance est dégradée ou totalement indisponible, afin de pouvoir concevoir des solutions pour gérer cette situation de manière appropriée.

3.a : Services AWS

Lorsque vous concevez une architecture multirégionale, il est important de comprendre Services AWS ce qui sera utilisé, les fonctionnalités multirégionales de ces services et les solutions que vous devrez concevoir pour atteindre les objectifs multirégionaux. Par exemple, Amazon Aurora et Amazon DynamoDB peuvent répliquer des données de manière asynchrone vers une région de secours. Toutes les Service AWS dépendances devront être disponibles dans toutes les régions à partir desquelles une charge de travail sera exécutée. Pour vérifier que les services que vous utilisez sont disponibles dans les régions souhaitées, consultez la liste Services AWS par région.

3.b : Dépendances internes et tierces

Assurez-vous que les dépendances internes de chaque charge de travail sont disponibles dans les régions à partir desquelles elles opèrent. Par exemple, si la charge de travail est composée de nombreux microservices, identifiez tous les microservices qui constituent une capacité commerciale et vérifiez que tous ces microservices sont déployés dans chaque région à partir de laquelle la charge de travail fonctionne. Vous pouvez également définir une stratégie pour gérer avec élégance les microservices devenus indisponibles.

Les appels interrégionaux entre microservices au sein d'une même charge de travail ne sont pas conseillés, et l'isolement régional doit être maintenu. En effet, la création de dépendances entre régions augmente le risque de défaillance corrélée, ce qui compense les avantages des

implémentations régionales isolées de la charge de travail. Les dépendances sur site peuvent également faire partie de la charge de travail. Il est donc important de comprendre comment les caractéristiques de ces intégrations pourraient changer si la région principale devait changer. Par exemple, si la région de secours est située plus loin de l'environnement sur site, l'augmentation de la latence peut avoir un impact négatif.

Comprendre les solutions logicielles en tant que service (SaaS), les kits de développement logiciel (SDKs) et les autres dépendances de produits tiers, et être en mesure de mettre en œuvre des scénarios dans lesquels ces dépendances sont dégradées ou indisponibles, permettra de mieux comprendre le fonctionnement et le comportement de la chaîne de systèmes en fonction des différents modes de défaillance. Ces dépendances peuvent se trouver dans le code de votre application, comme la gestion des secrets en externe à l'aide AWS Secrets Manager, ou elles peuvent impliquer une solution de coffre-fort tierce (telle que HashiCorp), ou des systèmes d'authentification dépendants AWS IAM Identity Centerdes connexions fédérées.

La redondance en matière de dépendances peut accroître la résilience. Si une solution SaaS ou une dépendance à un tiers utilise le même primaire Région AWS que la charge de travail, contactez le fournisseur pour déterminer si sa posture de résilience correspond à vos exigences en matière de charge de travail.

En outre, soyez conscient du destin partagé entre la charge de travail et ses dépendances, telles que les applications tierces. Si les dépendances ne sont pas disponibles dans (ou depuis) une région secondaire après un basculement, la charge de travail risque de ne pas être complètement rétablie.

3.c : Mécanisme de basculement

Le DNS est couramment utilisé comme mécanisme de basculement pour transférer le trafic de la région principale vers une région de secours. Passez en revue et examinez de manière critique toutes les dépendances prises par le mécanisme de basculement. Par exemple, si votre charge de travail utilise Amazon Route 53, comprendre que le plan de contrôle est hébergé dans cette région `us-east-1` signifie que vous devenez dépendant du plan de contrôle de cette région spécifique. Cela n'est pas recommandé dans le cadre d'un mécanisme de basculement si la région principale l'est également, `us-east-1` car cela crée un point de défaillance unique. Si vous utilisez un autre mécanisme de basculement, vous devez avoir une connaissance approfondie des scénarios dans lesquels le basculement ne fonctionnerait pas comme prévu, puis planifier les mesures d'urgence ou développer un nouveau mécanisme si nécessaire. Consultez le billet de blog Creating Disaster Recovery Mechanisms Using Amazon Route 53 pour découvrir les approches que vous pouvez utiliser pour réussir le basculement.

Comme indiqué dans la section précédente, tous les microservices faisant partie d'une capacité commerciale doivent être disponibles dans chaque région dans laquelle la charge de travail est déployée. Dans le cadre de la stratégie de basculement, tous les microservices faisant partie de la capacité commerciale doivent basculer ensemble afin d'éliminer le risque d'appels interrégionaux. Par ailleurs, si les microservices basculent indépendamment, il existe un risque de comportement indésirable, tel que des appels interrégionaux susceptibles de passer des microservices. Cela introduit de la latence et peut entraîner l'indisponibilité de la charge de travail pendant les délais d'expiration du client.

3.d : Dépendances de configuration

Les certificats, les clés, les secrets, les images Amazon Machine (AMIs), les images de conteneur et les paramètres font partie de l'analyse de dépendance nécessaire lors de la conception d'une architecture multirégionale. Dans la mesure du possible, il est préférable de localiser ces composants au sein de chaque région afin qu'ils ne partagent pas le même sort entre les régions en ce qui concerne ces dépendances. Par exemple, vous devez modifier les dates d'expiration des certificats afin d'éviter qu'un certificat expirant (avec des alarmes réglées pour « avertir à l'avance ») n'ait un impact sur plusieurs régions.

Les clés et les secrets de chiffrement doivent également être spécifiques à la région. Ainsi, en cas d'erreur lors de la rotation d'une clé ou d'un secret, l'impact est limité à une région spécifique.

Enfin, tous les paramètres de charge de travail doivent être stockés localement pour que la charge de travail puisse être récupérée dans la région spécifique.

Principaux conseils

- Une architecture multirégionale bénéficie de la séparation physique et logique entre les régions. L'introduction de dépendances entre régions au niveau de la couche d'application annule cet avantage. Évitez de telles dépendances.
- Les contrôles de basculement devraient fonctionner sans aucune dépendance vis-à-vis de la région principale.
- Le basculement doit être coordonné tout au long du parcours de l'utilisateur afin d'éliminer le risque d'augmentation de la latence et de la dépendance des appels interrégionaux.

Élément fondamental multirégional 4 : Préparation opérationnelle

L'exploitation d'une charge de travail multirégionale est une tâche complexe qui comporte des défis opérationnels spécifiques à une architecture multirégionale. Il s'agit notamment Compte AWS de la gestion, de la refonte des processus de déploiement, de la création d'une stratégie d'observabilité multirégionale, de la création et du test de processus de reprise, puis de la gestion des coûts. Un examen du niveau de préparation opérationnelle (ORR) peut aider les équipes à préparer une charge de travail pour la production, qu'elle soit exécutée dans une seule région ou dans plusieurs régions.

4.a : gestion Compte AWS

Pour déployer une charge de travail entre plusieurs régions Régions AWS, assurez-vous que tous les Service AWS quotas d'un compte sont égaux. Tout d'abord, identifiez tout Services AWS ce qui fait partie de l'architecture, examinez l'utilisation prévue dans les régions de secours, puis comparez l'utilisation prévue à l'utilisation actuelle. Dans certains cas, si la région de secours n'a jamais été utilisée auparavant, vous pouvez vous référer aux quotas de service par défaut pour comprendre le point de départ. Ensuite, pour tous les services qui seront utilisés, demandez une augmentation de quota à l'aide de la console Service Quotas (connexion requise) ou APIs.

Configurez les rôles AWS Identity and Access Management (IAM) dans chaque région pour attribuer aux opérateurs, les outils d'automatisation et Services AWS les autorisations appropriées aux ressources de la région de secours. Pour obtenir une isolation régionale pour les architectures multirégionales, isolez les rôles par région. Assurez-vous que les autorisations sont en place avant de passer en ligne avec une région en veille.

4.b : Pratiques de déploiement

Les fonctionnalités multirégionales peuvent compliquer le déploiement d'une charge de travail dans plusieurs régions. Vous devez vous assurer que vous déployez dans une région à la fois. Par exemple, si vous utilisez une approche active-passive, vous devez d'abord effectuer le déploiement dans la région principale, puis dans la région de secours. AWS CloudFormationvous aide à déployer l'infrastructure dans une ou plusieurs régions et peut être adaptée en fonction de vos besoins. AWS CodePipelinevous permet de créer un pipeline integration/continuous delivery (CI/CD (continu) comportant des actions interrégionales qui permettent le déploiement dans des régions différentes de

la région dans laquelle se trouve le pipeline. Ceci, combiné à des stratégies de déploiement robustes telles que le bleu/vert, permet un déploiement de temps d'arrêt minimal, voire nul.

Toutefois, le déploiement de fonctionnalités dynamiques peut devenir plus complexe lorsque l'état de l'application ou des données n'est pas externalisé vers un magasin persistant. Dans ces situations, adaptez soigneusement le processus de déploiement à vos besoins. Concevez le pipeline et le processus de déploiement pour déployer dans une région à la fois au lieu de déployer simultanément dans plusieurs régions. Cela réduit le risque de défaillances corrélées entre les régions. Pour en savoir plus sur les techniques utilisées par Amazon pour automatiser les déploiements de logiciels, consultez l'article Automating safe and handoff deployments de la AWS Builders' Library.

4.c : Observabilité

Lorsque vous concevez pour plusieurs régions, réfléchissez à la manière dont vous allez surveiller l'état de tous les composants de chaque région afin d'obtenir une vision globale de la santé régionale. Cela peut inclure des mesures de surveillance du délai de réplication, ce qui n'est pas pris en compte pour une charge de travail à région unique.

Lorsque vous créez une architecture multirégionale, pensez également à observer les performances de la charge de travail depuis les régions de secours. Cela inclut un bilan de santé et des canaris (tests synthétiques) effectués depuis la région de réserve afin de fournir un aperçu extérieur de l'état de santé de la région principale. En outre, vous pouvez utiliser Amazon CloudWatch Internet Monitor pour comprendre l'état du réseau externe et les performances de vos charges de travail du point de vue de l'utilisateur final. La région principale doit disposer de la même observabilité pour surveiller la région de réserve.

Les canaris de la région d'attente doivent surveiller les indicateurs de l'expérience client afin de déterminer l'état général de la charge de travail. Cela est nécessaire car en cas de problème dans la région principale, l'observabilité dans la région principale pourrait être altérée et cela aurait un impact sur votre capacité à évaluer l'état de la charge de travail.

Dans ce cas, l'observation en dehors de cette région peut fournir des informations. Ces mesures doivent être regroupées dans des tableaux de bord disponibles dans chaque région et dans des alarmes créées dans chaque région. Comme il CloudWatchs'agit d'un service régional, il est obligatoire d'avoir des alarmes dans les deux régions. Ces données de surveillance seront utilisées pour faire basculer l'appel d'une région principale vers une région de secours.

4.d : Processus et procédures

C'est le meilleur moment pour répondre à la question : « Quand dois-je échouer ? » est bien avant que vous n'en ayez besoin. Définissez des plans de reprise qui incluent les personnes, les processus et les technologies bien avant la survenue d'un problème, et testez-les régulièrement. Décidez d'un cadre décisionnel en matière de recouvrement. S'il existe un processus de restauration bien rodé et que le délai de restauration est bien connu, vous pouvez choisir de démarrer le processus de restauration en utilisant un basculement qui répond à l'objectif du RTO. Ce moment peut être immédiatement après l'identification d'un problème avec l'application dans la région principale, ou il peut s'agir d'un événement ultérieur lorsque les options de restauration de l'application dans la région ont été épuisées.

L'action de basculement elle-même doit être automatisée à 100 %, mais la décision d'activer le basculement doit être prise par des humains, généralement un petit nombre de personnes prédéterminées au sein de l'organisation. Ces personnes devraient tenir compte de la perte de données et d'informations sur l'événement. En outre, les critères d'un basculement doivent être clairement définis et compris globalement au sein de l'organisation. Pour définir et exécuter ces processus, vous pouvez utiliser des AWS Systems Manager runbooks, qui permettent une end-to-end automatisation complète et garantissent la cohérence des processus exécutés pendant les tests et le basculement.

Ces runbooks doivent être disponibles dans les régions principale et de secours pour démarrer les processus de basculement ou de retour en arrière. Lorsque cette automatisation est en place, définissez et suivez une cadence de test régulière. Cela garantit que lorsqu'un événement se produit, la réponse suit un processus bien défini et mis en pratique dans lequel l'organisation a confiance. Il est également important de prendre en compte les tolérances établies pour les processus de rapprochement des données. Confirmez que le processus proposé répond aux exigences RPO/RTO établies.

4.e : Tests

Avoir une approche de rétablissement non testée équivaut à ne pas avoir d'approche de rétablissement. Un niveau de test de base consiste à exécuter une procédure de restauration pour changer de région d'exploitation pour votre application. C'est ce que l'on appelle parfois une approche de rotation des applications. Nous vous recommandons de développer la capacité de changer de région pour adopter votre position opérationnelle normale ; toutefois, ce test à lui seul ne suffit pas.

Les tests de résilience sont également essentiels pour valider l'approche de restauration d'une application. Cela implique d'injecter des scénarios de défaillance particuliers, de comprendre comment votre application et votre processus de restauration réagissent, puis de mettre en œuvre les mesures d'atténuation nécessaires si le test ne se déroule pas comme prévu. Le fait de tester votre procédure de restauration en l'absence d'erreur ne vous indiquera pas comment votre application se comporte dans son ensemble en cas de panne. Vous devez élaborer un plan pour tester votre reprise par rapport aux scénarios de défaillance attendus. AWS Fault Injection Servicefournit une liste croissante de scénarios pour vous aider à démarrer.

Cela est particulièrement important pour les applications à haute disponibilité, où des tests rigoureux sont nécessaires pour garantir que les objectifs de continuité des activités sont atteints. Le test proactif des capacités de restauration réduit le risque de défaillances en production, ce qui permet de s'assurer que l'application peut atteindre le temps de restauration limité souhaité. Des tests réguliers renforcent également l'expertise opérationnelle, ce qui permet à l'équipe de se remettre rapidement et de manière fiable en cas de panne lorsqu'elle survient. L'exercice de l'élément humain, ou du processus, de votre approche de rétablissement est tout aussi essentiel que les aspects techniques.

4.f : Coût et complexité

Les implications financières d'une architecture multirégionale dépendent de l'augmentation de l'utilisation de l'infrastructure, des frais opérationnels et du temps consacré aux ressources. Comme indiqué précédemment, le coût d'infrastructure dans une région de secours est similaire à celui d'une région principale lors du préprovisionnement, ce qui double votre coût total. Fournir une capacité suffisante pour les opérations quotidiennes tout en préservant une capacité tampon suffisante pour tolérer les pics de demande. Configurez ensuite les mêmes limites dans chaque région.

En outre, si vous adoptez une architecture active-active, vous devrez peut-être apporter des modifications au niveau de l'application pour exécuter correctement votre application dans une architecture multirégionale. La conception et l'exploitation de ces modifications peuvent nécessiter beaucoup de temps et de ressources. Au minimum, les entreprises doivent consacrer du temps à comprendre les dépendances techniques et commerciales de chaque région et à concevoir des processus de basculement et de reprise.

Les équipes devraient également effectuer des exercices de basculement et de retour en arrière normaux pour se sentir à l'aise avec les runbooks qui seraient utilisés lors d'un événement. Bien que ces exercices soient essentiels pour obtenir les résultats escomptés d'un investissement multirégional, ils représentent un coût d'opportunité et privent du temps et des ressources d'autres activités.

4.g : Stratégie organisationnelle de basculement multirégionale

Régions AWS fournir des limites d'isolation des défaillances qui empêchent les défaillances corrélées et limitent l'impact Service AWS des déficiences, lorsqu'elles se produisent, sur une seule région. Vous pouvez utiliser ces limites de défauts pour créer des applications multirégionales composées de répliques indépendantes isolées des défauts dans chaque région afin de limiter les scénarios de destin partagé. Cela vous permet de créer des applications multirégionales et d'utiliser une gamme d'approches (sauvegarde et restauration, pilote, active-active) pour implémenter votre architecture multirégionale. Cependant, les applications ne fonctionnent généralement pas de manière isolée. Par conséquent, considérez à la fois les composants que vous utiliserez et leurs dépendances dans le cadre de votre stratégie de basculement. En général, plusieurs applications fonctionnent ensemble pour soutenir une histoire utilisateur, une fonctionnalité spécifique proposée à un utilisateur final, telle que la publication d'une photo et d'une légende sur une application de réseau social ou le paiement sur un site de commerce électronique. Pour cette raison, vous devez développer une stratégie organisationnelle de basculement multirégionale qui assure la coordination et la cohérence nécessaires au succès de votre approche.

Il existe quatre stratégies de haut niveau parmi lesquelles les organisations peuvent choisir pour orienter une approche multirégionale. Elles sont répertoriées de l'approche la plus détaillée à la plus large :

- Basculement au niveau des composants
- Basculement d'une application individuelle
- Basculement du graphe de dépendances
- Basculement de l'ensemble du portefeuille d'applications

Chaque stratégie comporte des compromis et répond à différents défis, notamment la flexibilité de la prise de décision en cas de basculement, la capacité de tester les combinaisons de basculement, la présence d'un comportement modal et l'investissement organisationnel dans la planification et la mise en œuvre. Pour en savoir plus sur chaque stratégie, consultez le billet de AWS blog Création d'une stratégie de basculement multirégionale organisationnelle.

Principaux conseils

- Passez en revue tous les Service AWS quotas pour vous assurer qu'ils sont égaux dans toutes les régions dans lesquelles la charge de travail sera appliquée.

- Le processus de déploiement doit cibler une région à la fois au lieu d'impliquer plusieurs régions simultanément.
- Des mesures supplémentaires telles que le délai de réplication sont spécifiques aux scénarios multirégionaux et doivent être surveillées.
- Étendre la surveillance de la charge de travail au-delà de la région principale. Surveillez les indicateurs de l'expérience client pour chaque région et mesurez ces données en dehors de chaque région dans laquelle une charge de travail est exécutée.
- Testez régulièrement le failover et le failback. Implémentez un seul runbook pour les processus de basculement et de retour en arrière et utilisez-le à la fois pour les tests et les événements en direct. Les runbooks destinés aux tests et aux événements en direct ne devraient pas être différents.
- Comprenez les compromis liés aux stratégies de basculement. Implémentez un graphe de dépendances ou une stratégie de portefeuille d'applications complète.

Conclusion et ressources

Ce guide couvre les cas d'utilisation courants des architectures multirégionales, les principes fondamentaux de la mise en œuvre de ces architectures et les implications de cette approche. Vous pouvez appliquer ces principes fondamentaux à n'importe quelle charge de travail et utiliser les informations comme cadre pour vous aider à décider si une architecture multirégionale est la bonne approche pour votre entreprise.

Pour plus d’informations, consultez les ressources suivantes :

- AWS Centre d'architecture
- AWS Framework Well-Architected
- AWS Well-Architected Tool
- Création d'une stratégie de basculement multirégionale organisationnelle (AWS article de blog)
- AWS Capacités multirégionales (AWS Re:Post article)

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
Mises à jour	Mises à jour tout au long du guide.	27 décembre 2024
Publication initiale	—	20 décembre 2022

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 locale vers Oracle sur une EC2 instance du 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 les bases de données source et cible sont synchronisées, mais seule la base de données source gère les transactions liées à la connexion des 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 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 son AIOps utilisation dans la stratégie de AWS migration, consultez 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'indexation qui indexent des informations sur Internet. D'autres robots, appelés « bots malveillants », 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.

CCo E

Voir le Centre d'excellence du cloud.

CDC

Voir 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 du 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 CCoarticles électroniques du blog sur la 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 accélérer votre adoption du cloud (par exemple, créer une zone de landing zone, définir un CCo E, établir 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

Consultez la base de données de gestion des configurations.

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 ouBitbucket Cloud. 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, Amazon SageMaker AI 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 de source, de construction, de test, de préparation 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

Voir 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.

EDI

Voir échange de données informatisé.

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.

échange de données informatisé (EDI)

L'échange automatique de documents commerciaux entre les organisations. Pour plus d'informations, voir Qu'est-ce que l'échange de données informatisé ?

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 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 CI/CD pipeline, 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 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.

invitation en quelques coups

Fournir à un LLM un petit nombre d'exemples illustrant la tâche et le résultat souhaité avant de lui demander d'effectuer une tâche similaire. Cette technique est une application de l'apprentissage contextuel, dans le cadre de laquelle les modèles apprennent à partir d'exemples (prises de vue) intégrés dans des instructions. Les instructions en quelques étapes peuvent être efficaces pour les tâches qui nécessitent un formatage, un raisonnement ou des connaissances de domaine spécifiques. Voir également l'invite Zero-Shot.

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 par le biais de 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.

FM

Voir le modèle de fondation.

modèle de fondation (FM)

Un vaste réseau neuronal d'apprentissage profond qui s'est entraîné sur d'énormes ensembles de données généralisées et non étiquetées. FMs sont capables d'effectuer une grande variété de tâches générales, telles que comprendre le langage, générer du texte et des images et converser en langage naturel. Pour plus d'informations, voir Que sont les modèles de base ?

G

IA générative

Sous-ensemble de modèles d'IA qui ont été entraînés sur de grandes quantités de données et qui peuvent utiliser une simple invite textuelle pour créer de nouveaux contenus et artefacts, tels que des images, des vidéos, du texte et du son. Pour plus d'informations, consultez Qu'est-ce que l'IA générative.

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.

image dorée

Un instantané d'un système ou d'un logiciel utilisé comme modèle pour déployer de nouvelles instances de ce système ou logiciel. Par exemple, dans le secteur de la fabrication, une image dorée peut être utilisée pour fournir des logiciels sur plusieurs appareils et contribue à améliorer la vitesse, l'évolutivité et la productivité des opérations de fabrication des appareils.

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 organisationnelles (OUs). 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 des historiens

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.

données de rétention

Partie de données historiques étiquetées qui n'est pas divulguée dans un ensemble de données utilisé pour entraîner un modèle d'apprentissage automatique. Vous pouvez utiliser les données de blocage pour évaluer les performances du modèle en comparant les prévisions du modèle aux données de blocage.

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.

IIo T

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 AWS de référence de sécurité recommande de configurer votre compte réseau avec les fonctions entrantes, sortantes et d'inspection VPCs afin de protéger l'interface bidirectionnelle entre votre application et l'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, voir Élaboration d'une stratégie de transformation numérique de l'Internet des objets (IIoT) industriel.

VPC d'inspection

Dans une architecture AWS multi-comptes, un VPC centralisé qui gère les inspections du trafic réseau VPCs entre (identique ou Régions AWS différent), Internet et les réseaux locaux. L'architecture AWS de référence de sécurité recommande de configurer votre compte réseau avec les fonctions entrantes, sortantes et d'inspection VPCs afin de protéger l'interface bidirectionnelle entre votre application et l'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, voir Interprétabilité du modèle d'apprentissage automatique avec 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

Voir 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.

grand modèle de langage (LLM)

Un modèle d'intelligence artificielle basé sur le deep learning qui est préentraîné sur une grande quantité de données. Un LLM peut effectuer plusieurs tâches, telles que répondre à des questions, résumer des documents, traduire du texte dans d'autres langues et compléter des phrases. Pour plus d'informations, voir Que sont LLMs.

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.

LLM

Voir le grand modèle de langage.

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 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 pour lequel AWS fonctionnent 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 AWS Well-Architected Framework.

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

Un petit service indépendant qui communique via un réseau bien défini APIs 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 en utilisant Lightweight. APIs 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 : migration réhébergée vers Amazon EC2 avec le service de migration AWS 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, à évaluer, à prévenir ou à 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.

ORR

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 AWS de référence de sécurité recommande de configurer votre compte réseau avec les fonctions entrantes, sortantes et d'inspection VPCs afin de protéger l'interface bidirectionnelle entre votre application et l'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 d'ingénierie système qui prend en compte la confidentialité tout au long du processus de développement.

zones hébergées privées

Conteneur contenant des informations sur la manière dont vous souhaitez qu'Amazon Route 53 réponde aux requêtes DNS pour un domaine et ses sous-domaines au sein d'un ou de plusieurs VPCs domaines. 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.

chaînage rapide

Utiliser le résultat d'une invite LLM comme entrée pour l'invite suivante afin de générer de meilleures réponses. Cette technique est utilisée pour décomposer une tâche complexe en sous-tâches ou pour affiner ou développer de manière itérative une réponse préliminaire. Cela permet d'améliorer la précision et la pertinence des réponses d'un modèle et permet d'obtenir des résultats plus précis et personnalisés.

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.

publish/subscribe (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).

CHIFFON

Voir Retrieval Augmented Generation.

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. Il détermine ce qui est considéré comme étant une perte de données acceptable entre le dernier point de reprise 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. Chacun Région AWS est isolé et indépendant 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.

Génération augmentée de récupération (RAG)

Technologie d'IA générative dans laquelle un LLM fait référence à une source de données faisant autorité qui se trouve en dehors de ses sources de données de formation avant de générer une réponse. Par exemple, un modèle RAG peut effectuer une recherche sémantique dans la base de connaissances ou dans les données personnalisées d'une organisation. Pour plus d'informations, voir Qu'est-ce que RAG ?

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 de l' 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.

sécurité dès la conception

Une approche d'ingénierie système qui prend en compte la sécurité tout au long du processus de développement.

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 EC2 instance Amazon 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 fournit un contrôle centralisé des autorisations pour tous les comptes d'une organisation dans AWS Organizations. SCPs définissez des garde-fous ou des limites aux actions qu'un administrateur peut déléguer à des utilisateurs ou à des rôles. Vous pouvez les utiliser SCPs comme listes d'autorisation ou de refus pour spécifier 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.

invite du système

Technique permettant de fournir un contexte, des instructions ou des directives à un LLM afin d'orienter son comportement. Les instructions du système aident à définir le contexte et à établir des règles pour les interactions avec les utilisateurs.

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

Un hub de transit réseau que vous pouvez utiliser pour interconnecter vos réseaux VPCs et ceux 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

Une connexion entre deux VPCs qui vous permet d'acheminer le trafic en utilisant des 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

Voir le cadre AWS de qualification de la charge de travail.

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 « 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.

invite Zero-Shot

Fournir à un LLM des instructions pour effectuer une tâche, mais aucun exemple (plans) pouvant aider à la guider. Le LLM doit utiliser ses connaissances pré-entraînées pour gérer la tâche. L'efficacité de l'invite zéro dépend de la complexité de la tâche et de la qualité de l'invite. Voir également les instructions en quelques clics.

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.